#SHAREorg

SHARE
Technology · Connections · Results

GDPS/Active-Active and Load Balancing via Server/Application State Protocol (SASP)

Dr. Steve Guendert
Brocade Communications

Michael Fitzpatrick
IBM Corporation

February 7, 2013
Session 12875

SHARE in San Francisco 2013

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:

- Advanced Peer-to-Peer Networking®
- AIX®
- alphaWorks®
- AnyNet®
- AS/400®
- BladeCenter®
- Candle®
- CICS®
- DataPower®
- DB2 Connect
- DB2®
- DRDA®
- e-business on demand®
- e-business (logo)
- e business(logo)®
- ESCON®
- FICON®
- GDDM®
- GDPS®
- Geographically Dispersed Parallel Sysplex
- HiperSockets
- HPR Channel Connectivity
- HyperSwap
- i5/OS (logo)
- i5/OS®
- IBM eServer
- IBM (logo)®
- IBM®
- IBM zEnterprise™ System
- IMS
- InfiniBand ®
- IP PrintWay
- IPDS
- iSeries
- LANDP®
- Language Environment®
- MQSeries®
- MVS
- NetView®
- OMEGAMON®
- Open Power
- OpenPower
- Operating System/2®
- Operating System/400®
- OS/2®
- OS/390®
- OS/400®
- Parallel Sysplex®
- POWER®
- POWER7®
- PowerVM
- PR/SM
- pSeries®
- RACF®
- Rational Suite®
- Rational®
- Redbooks
- Redbooks (logo)
- Sysplex Timer®
- System i5
- System p5
- System x®
- System z®
- System z9®
- System z10
- Tivoli (logo)®
- Tivoli®
- VTAM®
- WebSphere®
- xSeries®
- z9®
- z10 BC
- z10 EC
- zEnterprise
- zSeries®
- z/Architecture
- z/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.

Abstract

The GDPS/Active-Active sites concept is a fundamental paradigm shift in disaster recovery from a failover model to a continuous availability model. GDPS/Active-Active consists of two sites, separated by virtually unlimited distances, running the same applications and having the same data to provide cross site workload balancing and continuous availability. One of the key components to a GDPS/Active-Active solution are external load balancing IP routers that balance workloads through the Server/Application State Protocol (SASP). This session will discuss the GDPS/Active-Active workload balancing function, with a focus on SASP, the router functionality, and how it functions in conjunction with the IBM Multi-site Workload Lifeline for z/OS.

Agenda

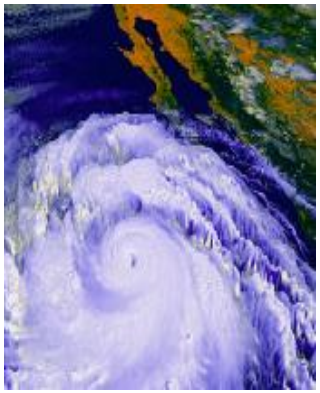
- Business Continuity vs. IT Resiliency
- Introduction to GDPS Active-Active
- Requirements-hardware and software
- Server Application State Protocol (SASP) Overview
  - Motivation and high level overview of the protocol
- Overview of IBM solutions based on SASP
  - z/OS Sysplex Clusters (z/OS Load Balancing Advisor)
  - Active/Active - Next generation of IBM's disaster recovery technology (Multi-site Workload Lifeline product)
- Conclusion and questions.

Definitions

BUSINESS CONTINUITY
IT RESILIENCY

Why Worry?

Lightning Strikes



Hurricanes / Cyclones

Earthquake

Tornadoes and other storms

Tsunami

Data theft and security breaches

Overloaded lines and infrastructure

Cut cables and power

Terrorism

IT resilience

- The ability to rapidly adapt and respond to any internal or external disruption, demand, or threat and continue business operations without significant impact.
  - Continuous/near continuous application availability (CA)
  - Planned and unplanned outages
- Broader in scope than disaster recovery (DR)
  - DR concentrates solely on recovering from unplanned events
- **Bottom Line**
  - **Business continuance is no longer simply IT DR**

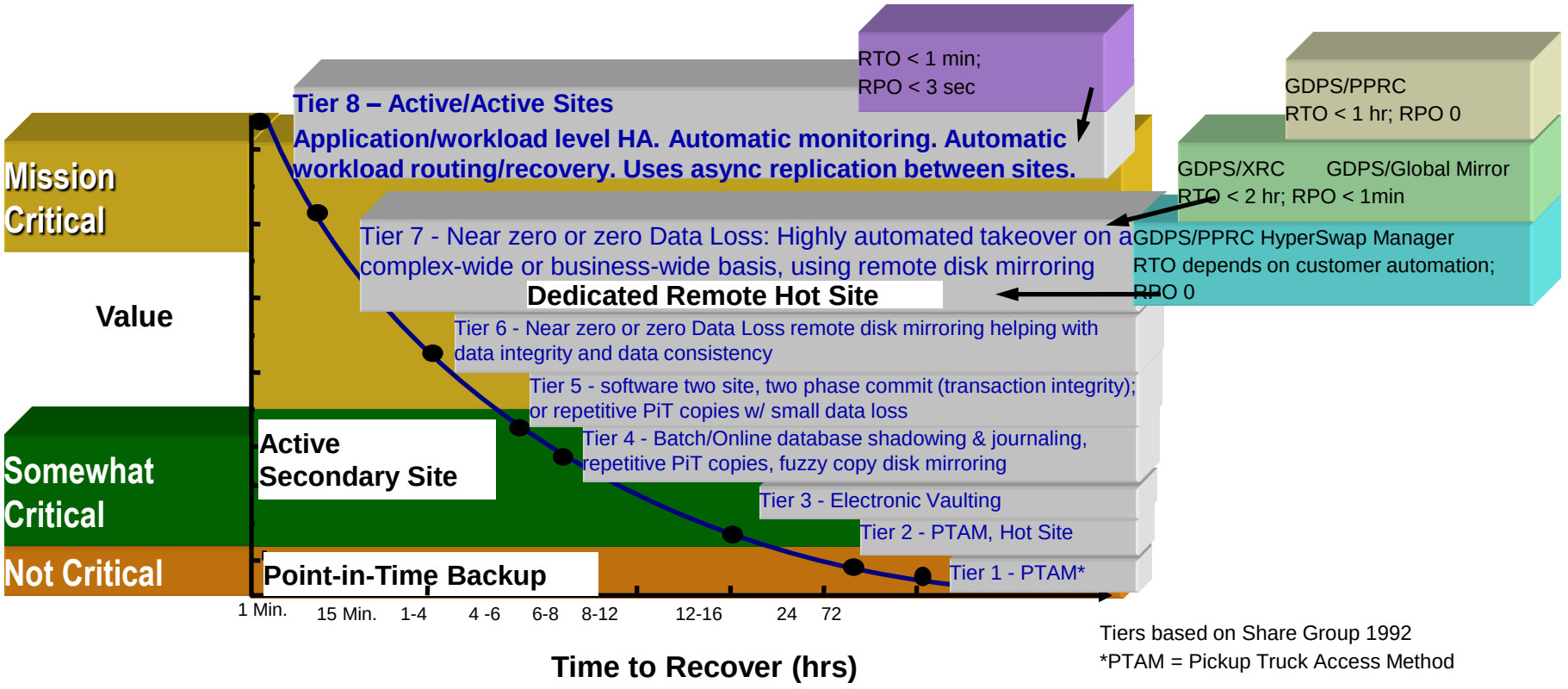
RTO

- Recovery Time Objective (RTO)
  - A metric for how long it takes to recover the application and resume operations after a planned or unplanned outage
  - How long your business can afford to wait for IT services to be resumed.
  - How much pain can you take?
  - Days, hours, or minutes?

RPO

- Recovery Point Objective (RPO)
  - A metric for how much data is lost
  - The actual recovery point to which all data is current and consistent.
  - How much data your company is willing to recreate following an outage.
    - What is the acceptable time difference between the data in your production system and the data at the recovery site?

Tiers of Disaster Recovery

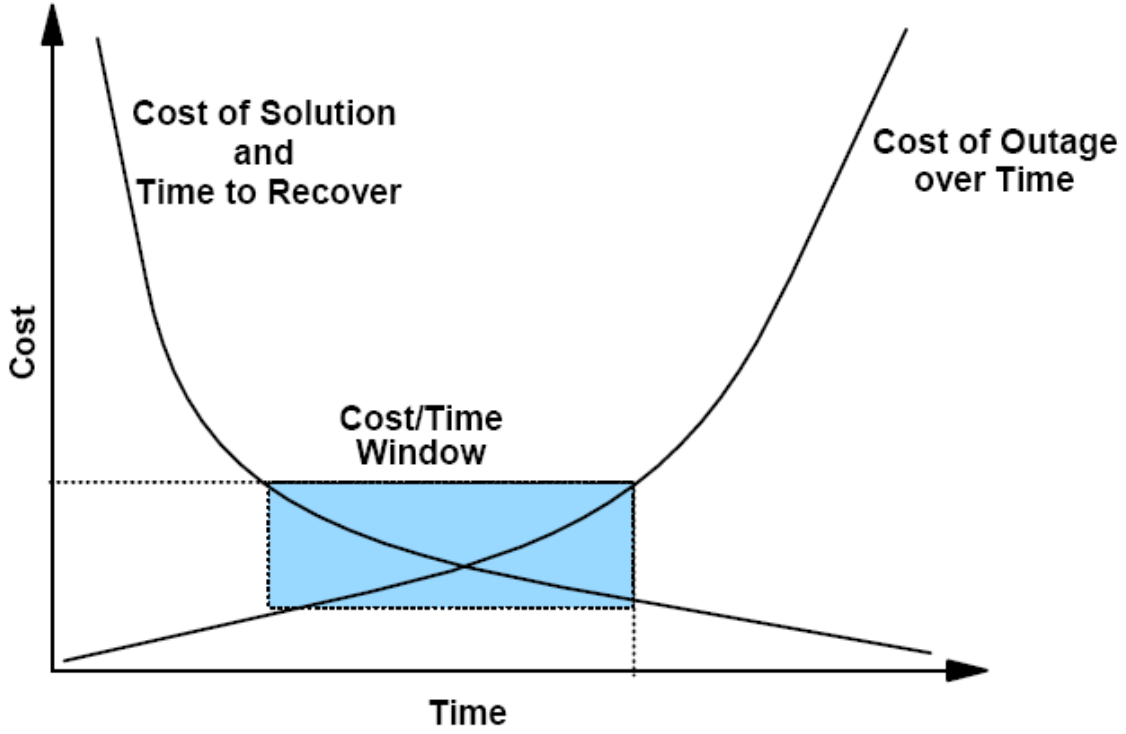


Tiers based on Share Group 1992
*PTAM = Pickup Truck Access Method

Failover models can only achieve so much in improving RTO

RTO and RPO

- Cost tradeoffs- balancing need vs. afford

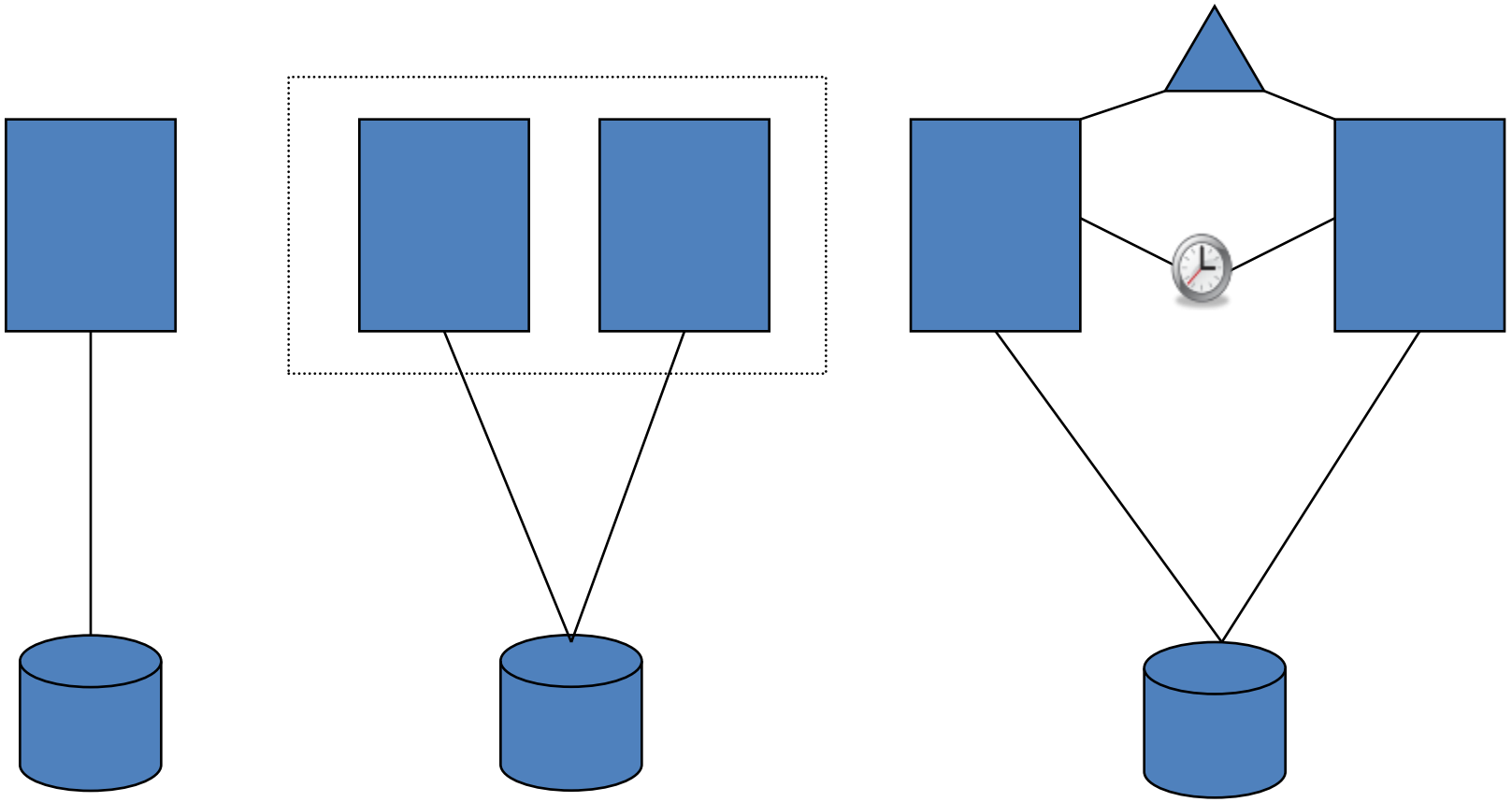


Cost of business continuity solution versus cost of outage

Guendert: Revisiting Business Continuity and Disaster Recovery Planning and Performance For 21st Century Regional Disasters: The case for GDPS. *Journal of Computer Resource Management*. Summer 2007

INTRODUCTION TO GDPS ACTIVE-ACTIVE

Availability and the IBM Mainframe



Single system

Mutliple partitions/images

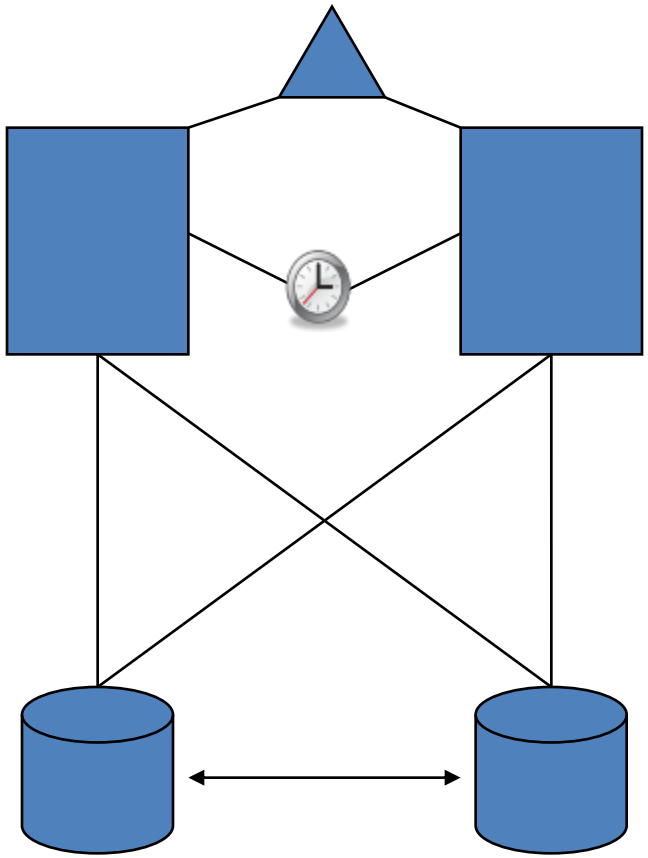
Parallel Sysplex

1964

1982-1990

1995

Availability and the IBM Mainframe



GDPS/PPRC

Parallel Sysplex
Synchronous Mirroring
HyperSwap for primary disk failures
Near-continuous Availability

Limitation: Performance at distance

1998

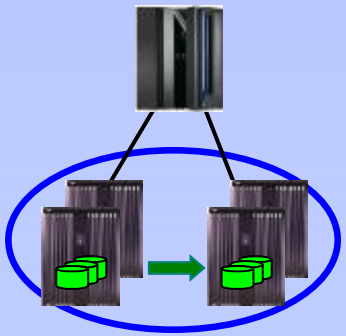
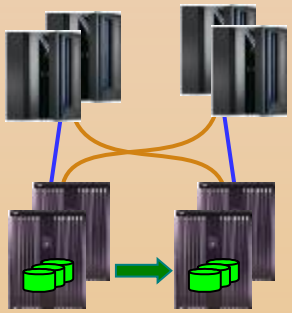
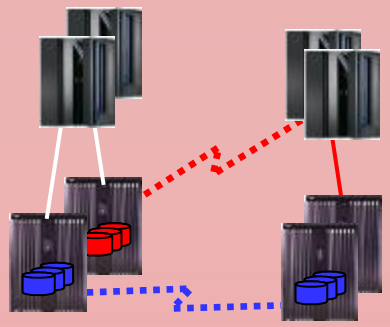
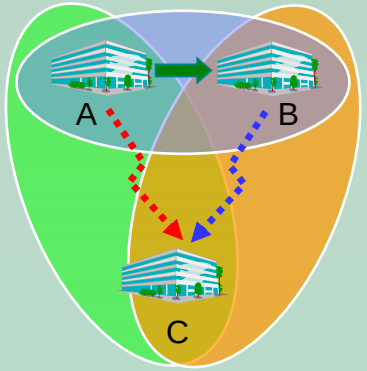
What are GDPS/PPRC customers doing today?

- GDPS/PPRC, based upon a multi-site Parallel Sysplex and synchronous disk replication, is a metro area Continuous Availability (CA) and Disaster Recovery solution (DR)
- GDPS/PPRC supports two configurations:
  - Active/standby or single site workload
  - Active/active or multi-site workload
- Some customers have deployed GDPS/PPRC active/active configurations
  - All critical data must be PPRCed and HyperSwap enabled
  - All critical CF structures must be duplexed
  - Applications must be parallel sysplex enabled
  - Signal latency will impact OLTP thru-put and batch duration resulting in the sites being separated by no more than a couple tens of KM (fiber)
- Issue: the GDPS/PPRC active/active configuration does not provide enough site separation for some enterprises

What are GDPS/XRC & GDPS/GM customers doing today?

- GDPS/XRC and GDPS/GM, based upon asynchronous disk replication, are unlimited distance DR solutions
- The current GDPS async replication products require the failed site's workload to be restarted in the recovery site and this typically will take 30-60 min
  - Power fail consistency
  - Transaction consistency
- There are no identified extensions to the existing GDPS async replication products that will allow the RTO to be substantially reduced.
- Issue: GDPS/XRC and GDPS/GM will not achieve an RTO of seconds being requested by some enterprises

What are GDPS customers doing today ?

Continuous Availability of Data within a Data Center	Continuous Availability / Disaster Recovery within a Metropolitan Region	Disaster Recovery at Extended Distance	Continuous Availability Regionally and Disaster Recovery Extended Distance
Single Data Center Applications remain active Continuous access to data in the event of a storage subsystem outage	Two Data Centers Systems remain active Multi-site workloads can withstand site and/or storage failures	Two Data Centers Rapid Systems Disaster Recovery with "seconds" of Data Loss Disaster recovery for out of region interruptions	Three Data Centers High availability for site disasters Disaster recovery for regional disasters
			
GDPS/HyperSwap Mgr RPO=0 & RTO=0	GDPS/PPRC RPO=0 & RTO<1 hr	GDPS/GM & GDPS/XRC RPO secs & RTO <1 hr	GDPS/MGM & GDPS/MzGM

Drivers for improvements in HA and DR

- Interagency Paper on Sound Practices to Strengthen the Resilience of the U.S. Financial System [Docket No. R-1128] (April 7, 2003)
  - Focus on mission critical workloads, their recovery and resumption of normal processing
- Cost of an outage
  - Financial
  - Reputation
- Global Business Model
  - 24x7 processing
  - Planned outage avoidance

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."

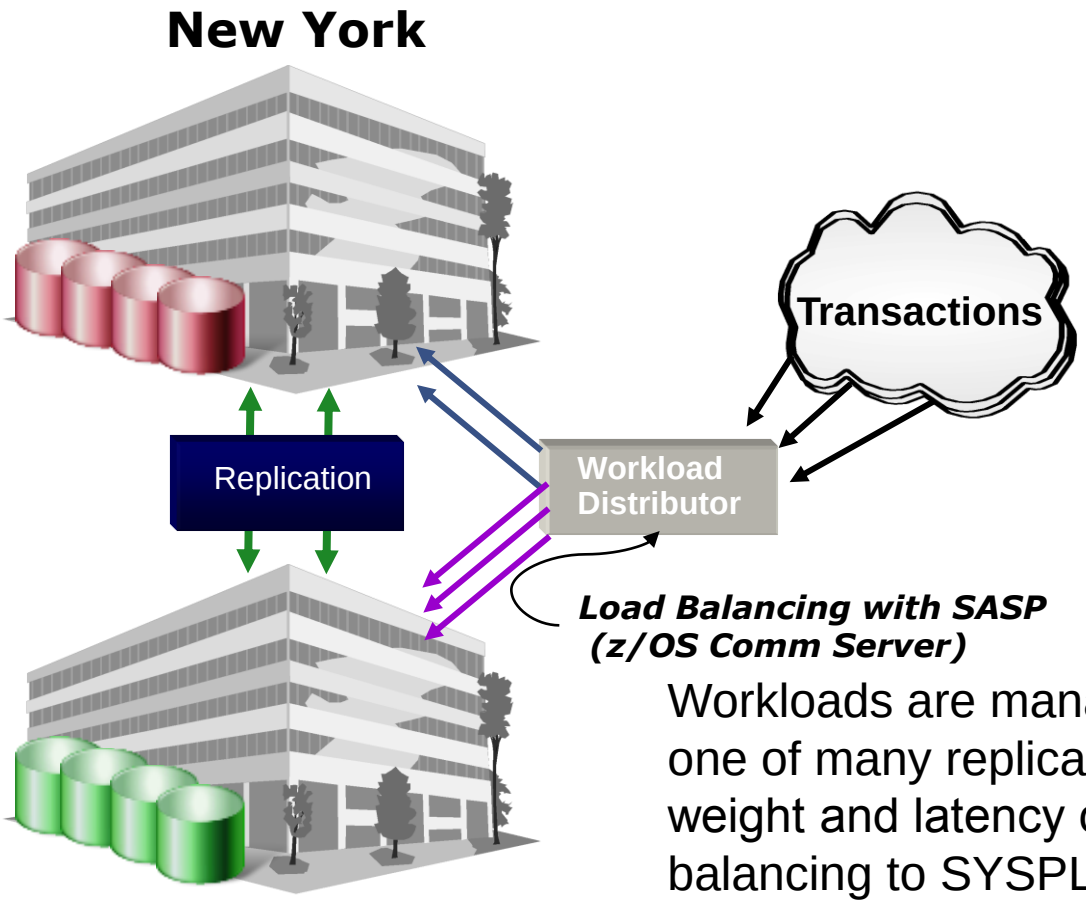
Customer requirements for HA/DR/BC in 2013

- 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).

IBM GDPS active/active

- **Long distance disaster recovery with only seconds of impact**
- Continuous availability
- Fundamental paradigm shift from a failover model to a near continuous availability model.
- Allows for unlimited distance replication with only seconds of user impact if there is a site disaster.
- Uses software based replication and techniques for copying the data between sites
- Provides control over which workloads are being protected.
- GDPS automation provides an end to end automated solution
  - Helps manage the availability of the workload
  - Coordination point/ controller for activities including being a focal point for operating and monitoring the solution and readiness for recovery.

Active/Active concepts



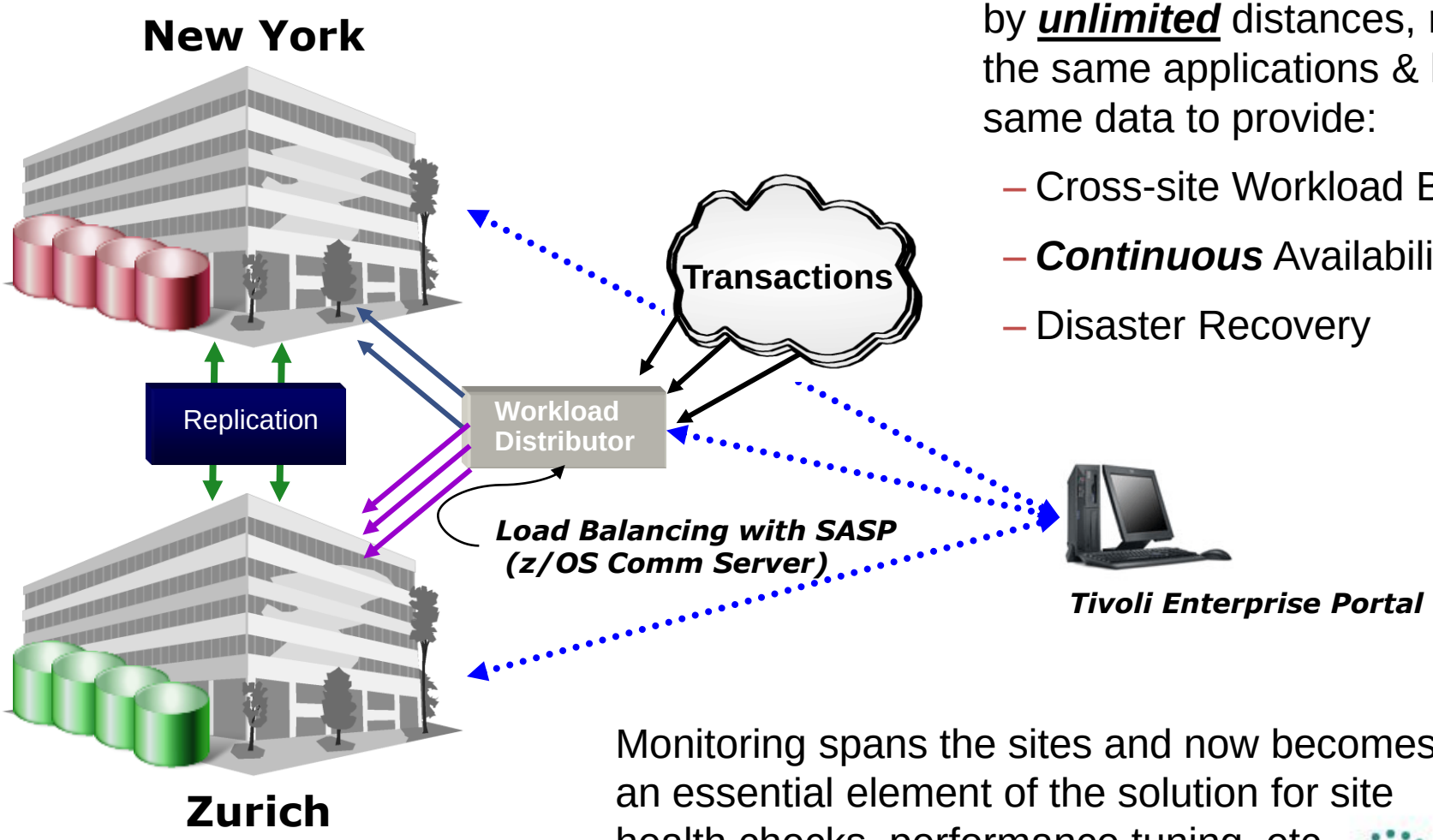
Two or more sites, separated by ***unlimited*** distances, running the same applications & having the same data to provide:

- Cross-site Workload Balancing
- ***Continuous*** Availability
- Disaster Recovery

Workloads are managed by a client and routed to one of many replicas, depending upon workload weight and latency constraints … extends workload balancing to SYSPLEXs across multiple sites!

Data at geographically dispersed sites are kept in sync via replication

Active/Active concepts



Two or more sites, separated by ***unlimited*** distances, running the same applications & having the same data to provide:

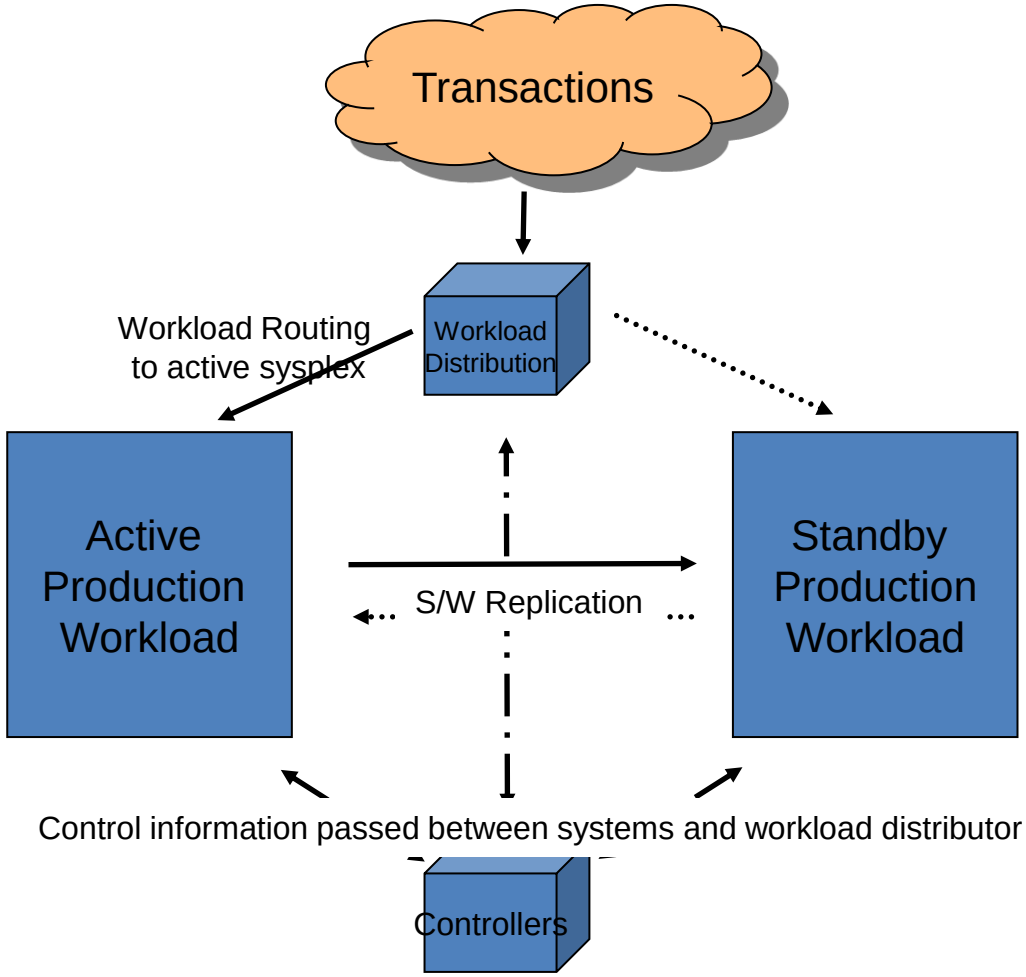
- Cross-site Workload Balancing
- ***Continuous*** Availability
- Disaster Recovery

Monitoring spans the sites and now becomes an essential element of the solution for site health checks, performance tuning, etc.

Hardware and software supporting SASP

GDPS ACTIVE-ACTIVE REQUIREMENTS

Conceptual view of GDPS Active-Active



GDPS/Active-Active Software Components

- Integration of a number of software products
  - z/OS 1.11 or higher
  - IBM Multi-site Workload Lifeline v1.1
  - IBM Tivoli NetView for z/OS v6.1
  - IBM Tivoli Monitoring v6.2.2 FP3
  - IBM InfoSphere Replication Server for z/OS v10.1
  - IBM InfoSphere IMS Replication for z/OS v10.1
  - System Automation for z/OS v3.3
  - GDPS/Active-Active v1.1
  - Optionally the OMEGAMON suite of monitoring tools to provide additional insight

Replication

- Not hardware based mirroring
  - IBM InfoSphere Replication Server for z/OS v10.1
    - Runs on production images where required to capture (active) and apply (standby) data updates for DB2 data. Relies on MQ as the data transport mechanism (QREP).
  - IBM InfoSphere IMS Replication for z/OS v10.1
    - Runs on production images where required to capture (active) and apply (standby) data updates for IMS data. Relies on TCPIP as the data transport mechanism.
  - System Automation for z/OS v3.3
    - Runs on all images. Provides a number of critical functions:
      - Remote communications capability to enable GDPS to manage sysplexes from outside the sysplex
      - System Automation infrastructure for workload and server management

GDPS/Active-Active Hardware components

- Two Production Sysplex environments (also referred to as sites) in different locations
  - One active, one standby – for each defined workload
  - Software-based replication between the two sysplexes/sites
    - IMS and DB2 data is supported
- Two Controller Systems
  - Primary/Backup
  - Typically one in each of the production locations, but there is no requirement that they are co-located in this way
- Workload balancing/routing switches
  - Must be Server/Application State Protocol compliant (SASP)
    - RFC4678 describes SASP

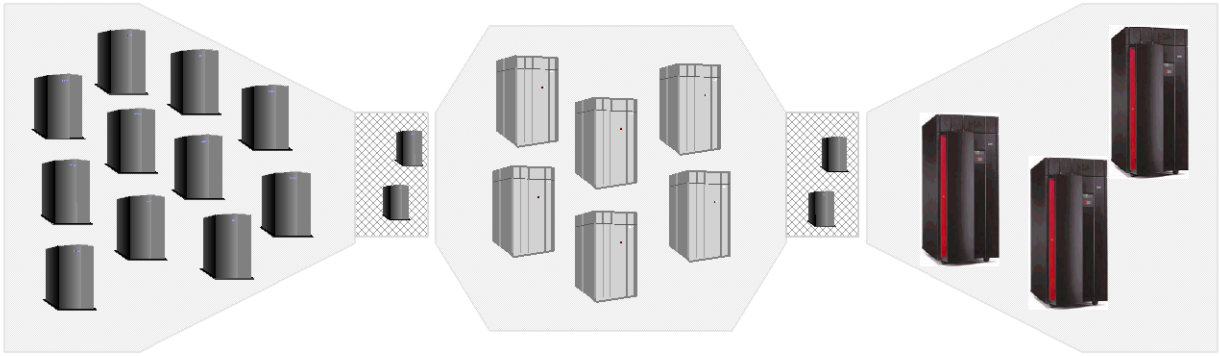
Basics of how it works and how workloads get balanced

SASP

Agenda

- Server Application State Protocol (SASP) Overview
  - Motivation and high level overview of the protocol
- Overview of IBM solutions based on SASP
  - z/OS Sysplex Clusters (z/OS Load Balancing Advisor)
  - Active/Active - Next generation of IBM's disaster recovery technology (Multi-site Workload Lifeline product)

Traffic Weight recommendations to Load Balancers



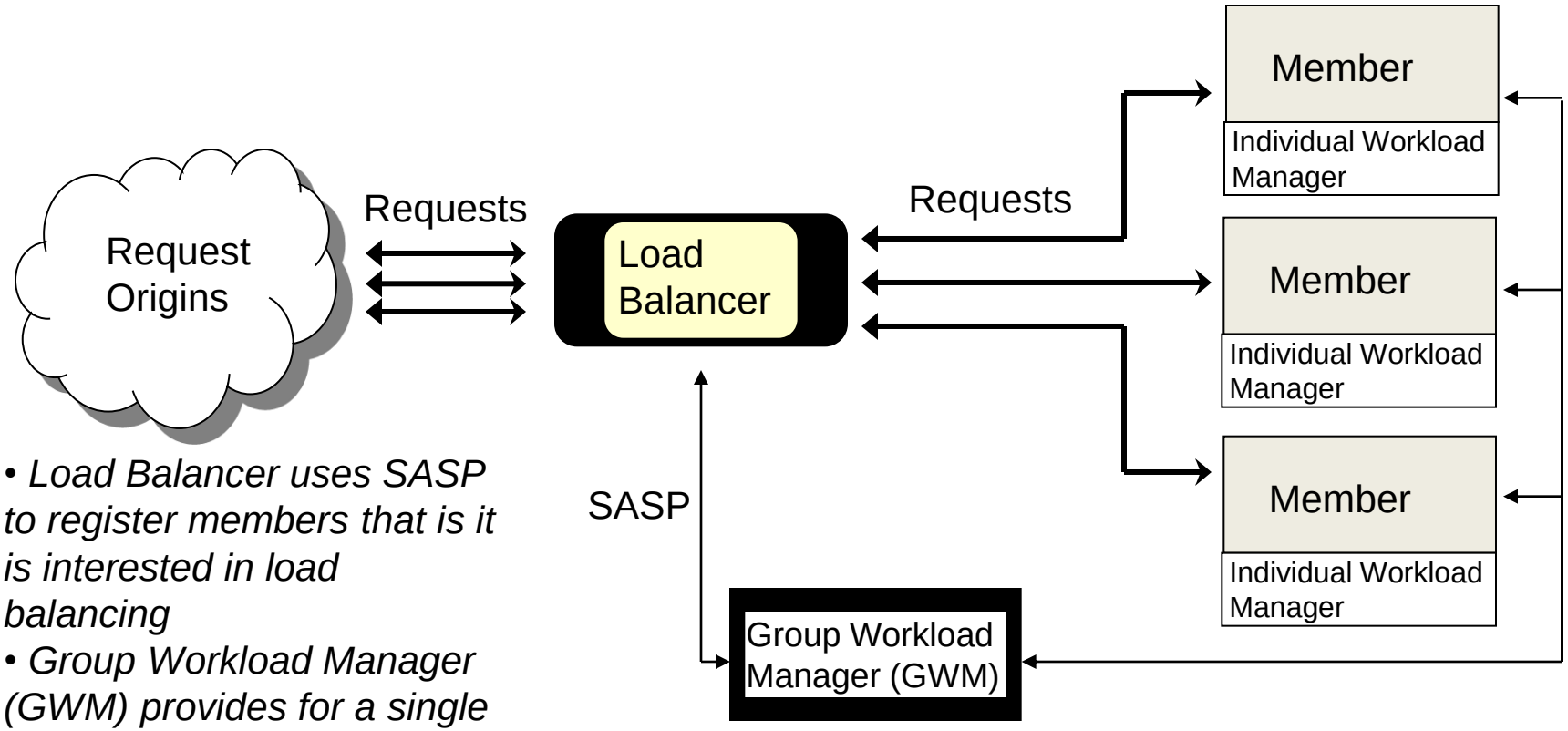
The ability to distribute work across equal servers...

- factoring in the availability of server resources,
- factoring in the *business importance* of the work,
- estimating the liklihood of meeting objectives,
- avoiding over-utilized servers, where possible,
- factoring in down-stream server dependencies.

Server/Application State Protocol (SASP) Objectives

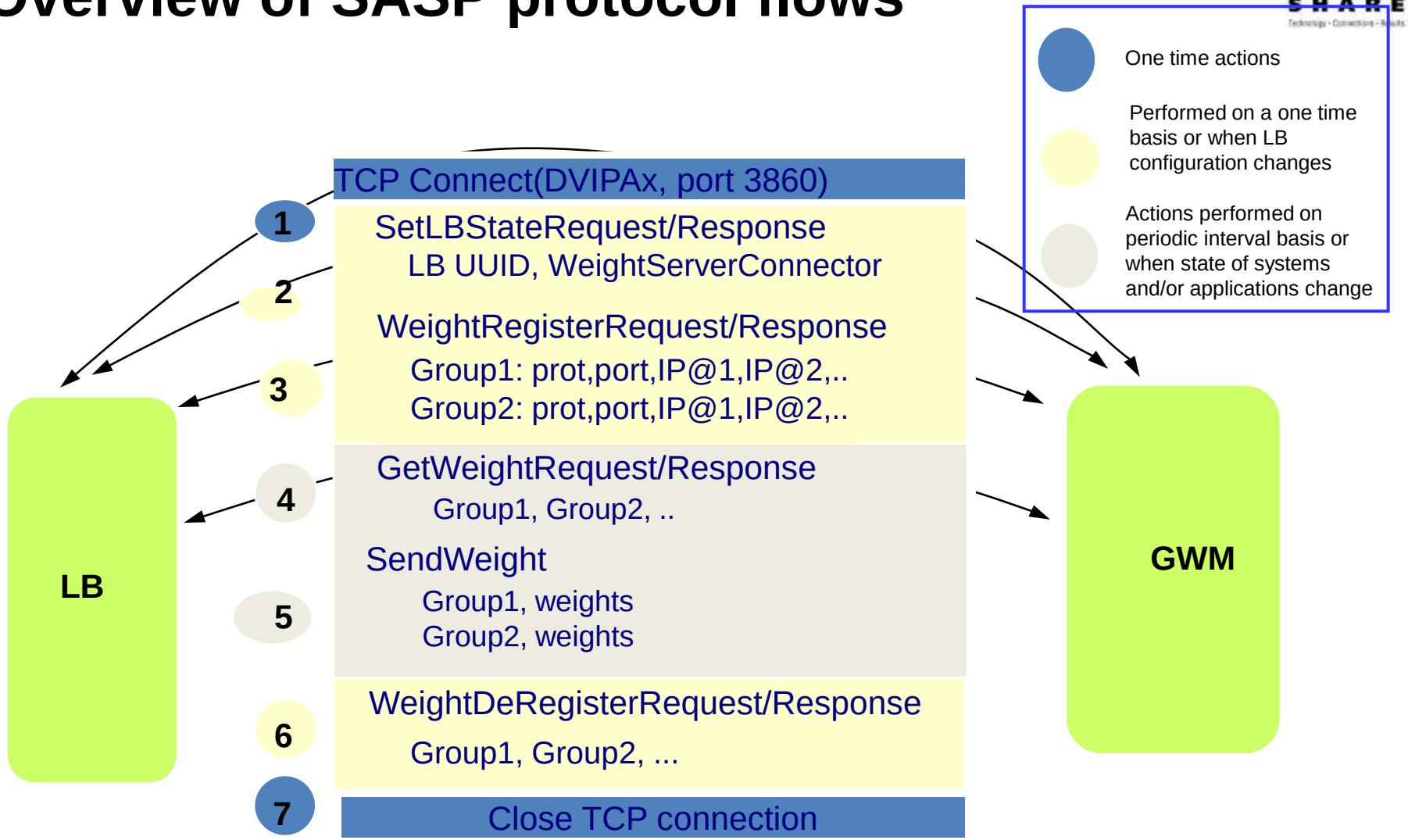
- Provide a mechanism for workload managers to give distribution recommendations to load balancers.
- Must be lightweight – minimal:
  - implementation complexity
  - processing overhead
  - additional user configuration
- Must be extensible
- *SASP will not handle the transport or actual distribution of work, only give recommendations*
- Open protocol documented in RFC4678:
  - http://www.faqs.org/rfcs/rfc4678.html

SASP - High Level Architecture



- *Load Balancer uses SASP to register members that is it is interested in load balancing*
- *Group Workload Manager (GWM) provides for a single point of contact for the entire cluster*

Overview of SASP protocol flows



Registering interest in target applications using SASP

- Load Balancer registers Groups of clustered servers it is interested in load balancing
  - Each group designates an application cluster to be load balanced
- Each group consists of a list of members (i.e. target servers)
  - System-level cluster: A list of target Systems identified by IP address (in lieu of individual application servers)
    - Recommendations returned in this scenario are also at a System-level
    - No specific target application information returned in this case
  - Application-level cluster: A list of applications comprising the “load balancing” group
    - Allows specific recommendations to be provided for each target socket application
      - *vs providing the same recommendation for all application instances running on the same host*
    - Identified by protocol (TCP/UDP), IP address of the target system they reside on, and the port the application is using.
    - SASP allows for target servers in a load balancing group to use different ports (and even different protocols TCP/UDP)
      - *Probably not applicable for real application workloads*
  - Supports IPv4 and IPv6
    - Both for identifying members (target servers) and for the actual communications to the GWM (i.e. SASP connections)

Frequency of SASP communications

- SASP supports both a "push" and a "pull" model for updating the load balancer with workload recommendations
  - Load balancer tells GWM which model it wants to use
- "Pull" model
  - GWM "suggests" a polling interval to the load balancer
    - z/OS Load Balancing Advisor uses the configurable update_interval value for this purposcte
  - Load balancer has the option to ignore this value
    - Load balancer requests updates each polling interval
- "Push" model
  - GWM sends updated information to the load balancer on an interval basis
    - z/OS Load Balancing Advisor uses the configurable update_interval value for this purpose
  - GWM may send data more frequently than the interval period
  - Active/Active Multi-site Workload Lifeline product requires “push” to be enabled
- Load balancer determines whether it wants information about all members it registered or only changed information about its registered members

Basic Protocol Components

5 Basic Components are used throughout the protocol

- **SASP Header**
  - Version
  - Message Length
  - Message ID
- **Member Data**
  - Protocol
  - Port
  - IP Address
  - Label length
  - Label
- **Weight Entry**
  - Member Data
  - Contact Flag
  - Quiesce Flag
  - Registration Flag
  - Weight
- **Member State Instance**
  - Member Data
  - Opaque State
  - Quiesce Flag
- **Group Data**
  - LB UID Length
  - LB UID
  - Group Name Length
  - Group Name

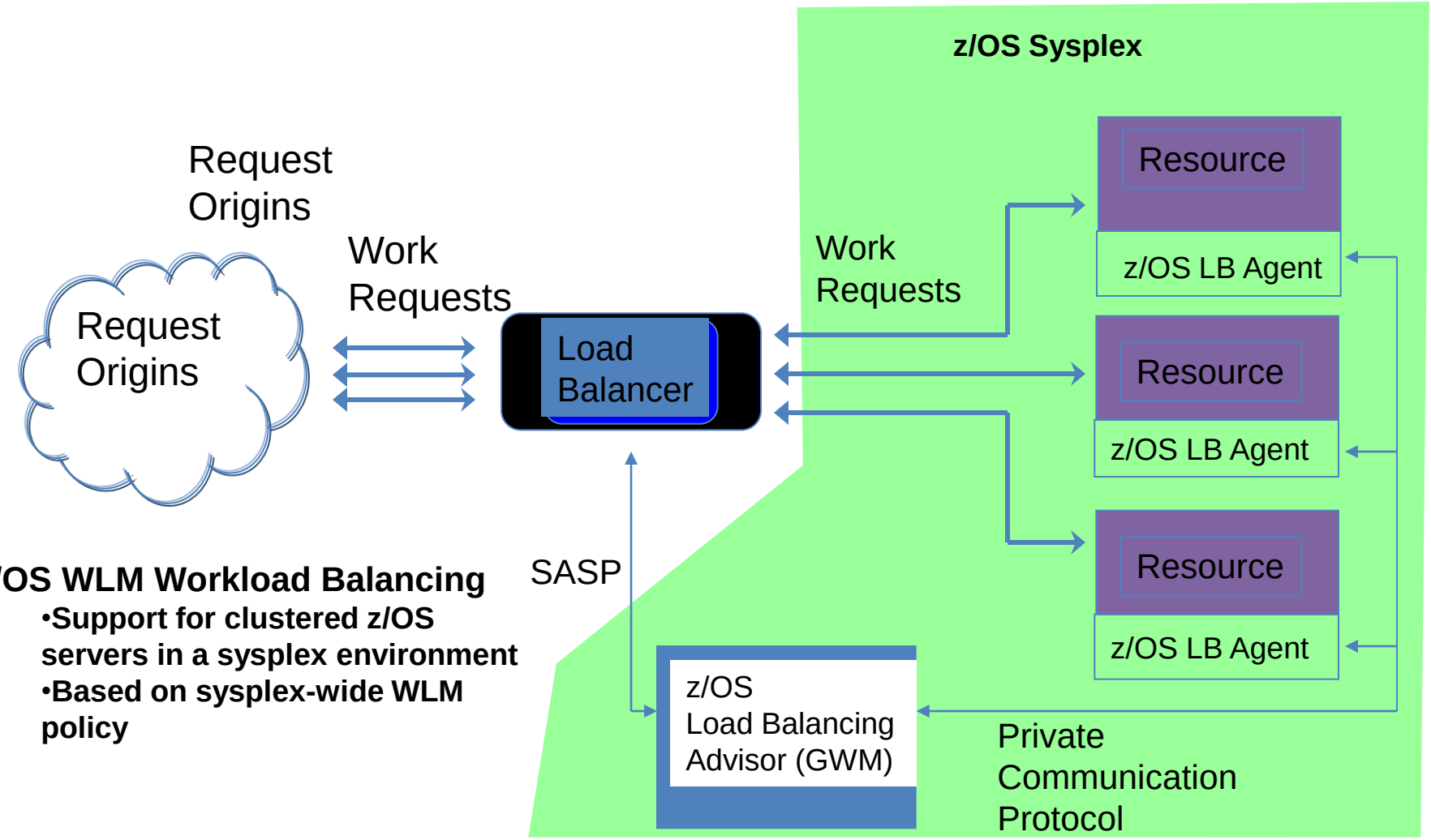
Group Protocol Components

3 Group Components are used throughout the protocol

- **Group of Member Data**
  - **Group Data**
  - **Member Data Count**
  - **Array of Member Data Components**
- **Group of Weight Data**
  - **Group Data**
  - **Weight Entry Count**
  - **Array of Weight Entry Components**
- **Group of Member State Data**
  - **Group Data**
  - **Resource State Instance Count**
  - **Array of Resource State Instances**

IBM solutions supporting SASP

z/OS Load Balancing Advisor



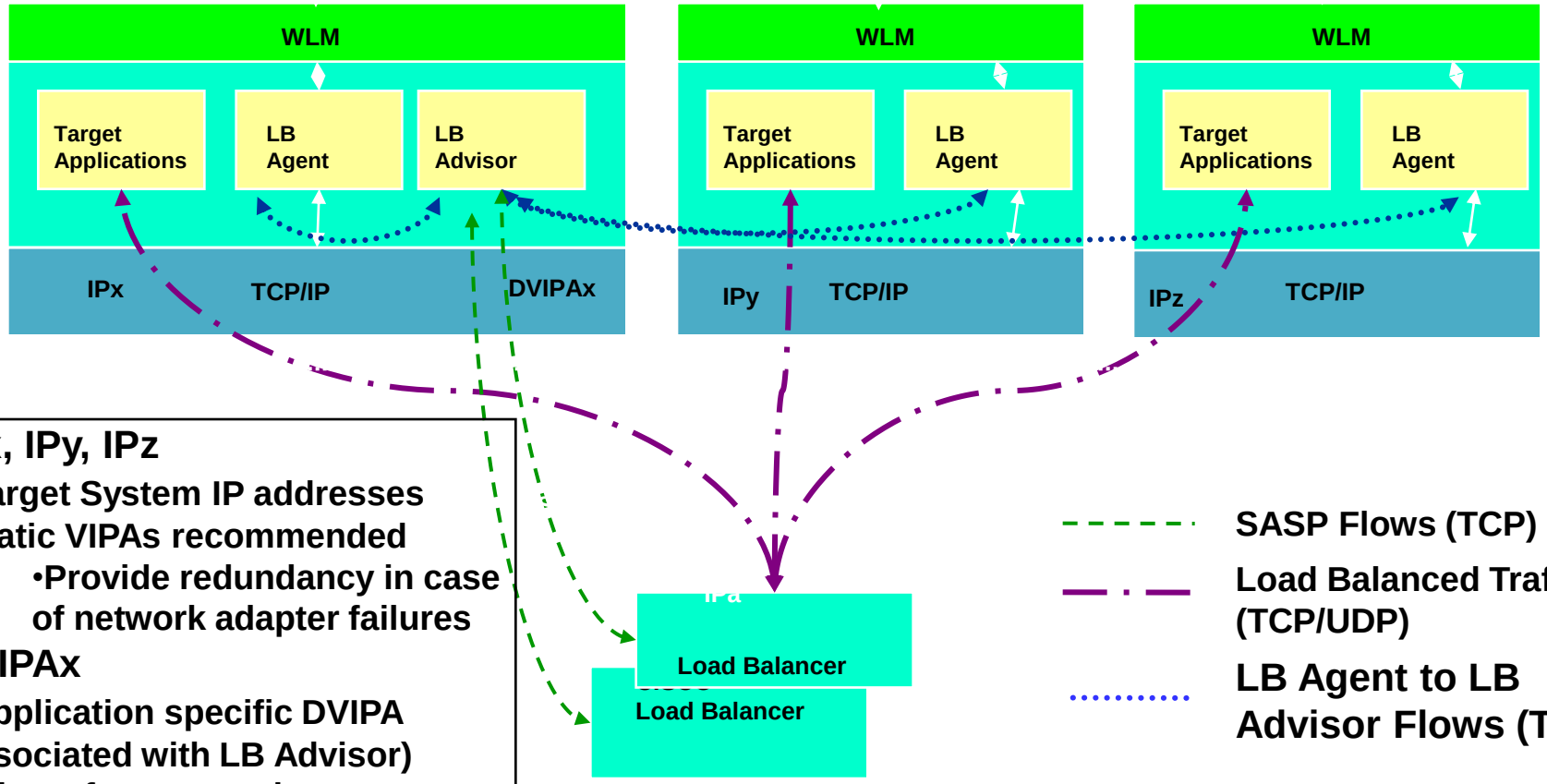
- **z/OS WLM Workload Balancing**
  - **Support for clustered z/OS servers in a sysplex environment**
  - **Based on sysplex-wide WLM policy**

How does the z/OS LBA calculate weights?

- The weights are composed of several components:
  - Available CPU capacity for each target system
  - Displaceable capacity for each target application and system
    - For systems with high CPU utilization, what is the relative importance of the workload being load balanced to other workloads on that system
      - *If a lot of lower important work is running on that system, it can be displaced by this higher priority workload*
  - Is the target application server meeting user specified WLM (Workload Manager) goals for the workload?

- Health of the application – What if CPU capacity exists but the application infrastructure is experiencing other constraints and abnormal conditions?
  - When these conditions are detected, the weights for the "unhealthy" target application instance is reduced appropriate
  - Health conditions monitored:
    - Monitoring of TCP backlog queue. What is the application is falling behind in accepting new connections?
    - Application reported health
      - *APIs are available on the platform that allow applications to report any abnormal conditions that result in sub optimal workload processing*
        - *memory constraints, thread constraints, resources unavailable*
        - *The next tier server or resource manager is not available (e.g. Database server)*

z/OS LBA Solution Overview



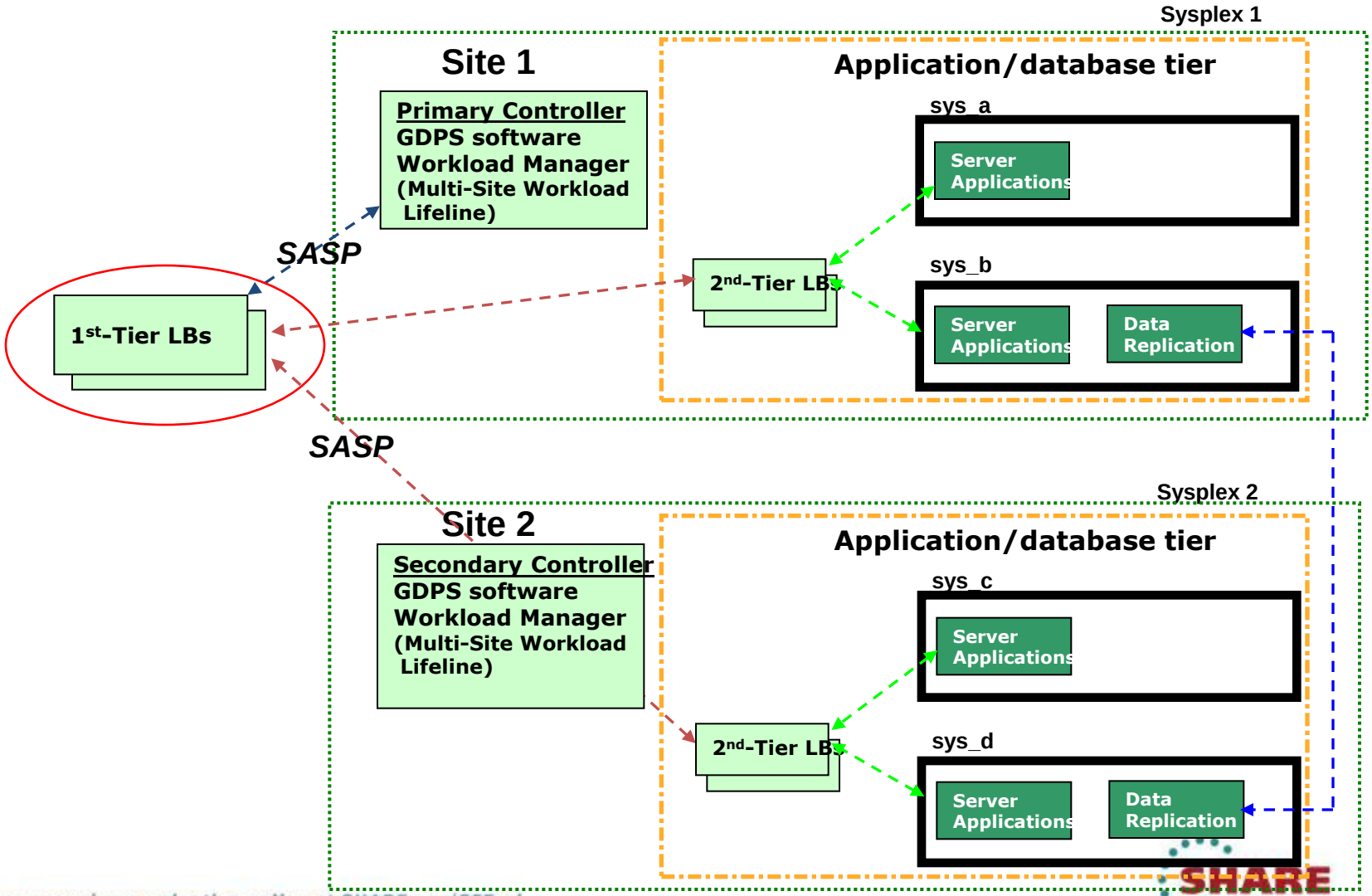
IPx, IPy, IPz

- Target System IP addresses
- Static VIPAs recommended
  - Provide redundancy in case of network adapter failures

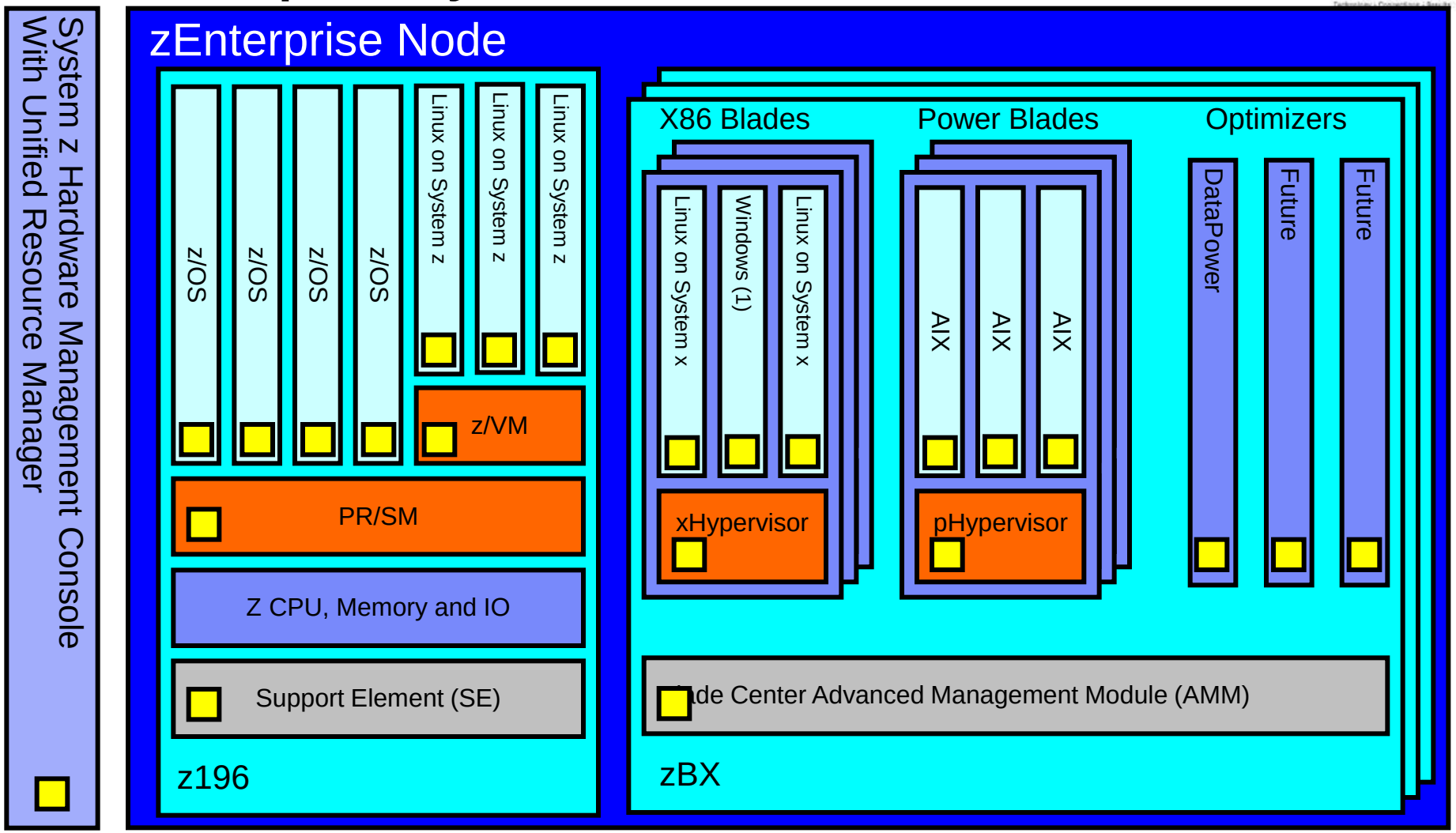
DVIPAx

- Application specific DVIPA (associated with LB Advisor)
- Allows for automatic movement in failure scenarios

Active/Active Sites use case – Multi-Site Workload Lifeline



IBM zEnterprise System Overview



Connecting the pieces with zManager (aka. Unified Resource Manager)!

CONCLUSION AND QUESTIONS?