Follow @IBMGDPS

#SHAREorg

SHARE
Technology · Connections · Results

Driving towards continuously available applications on System z

Dave Clitherow
IBM
dave.clitherow@uk.ibm.com

Wednesday 14th August 2013
Session 13635

SHARE in Boston

Disclaimer

- IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion.
- Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
- The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including 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 results similar to those stated here

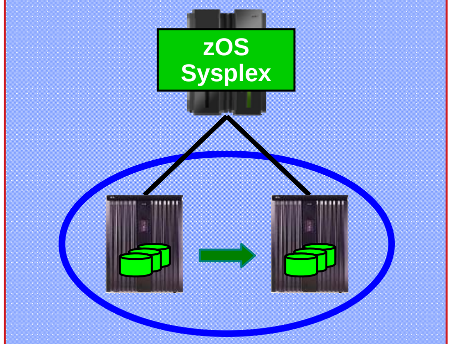
Suite of GDPS service products to meet various business requirements for availability and disaster recovery

Continuous Availability of Data within a Data Center

Single Data Center

Application remain active

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



GDPS/PPRC HM

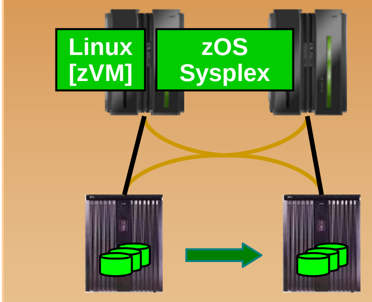
RPO 0 sec & RTO 0 sec

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
active/active,
active/standby configs

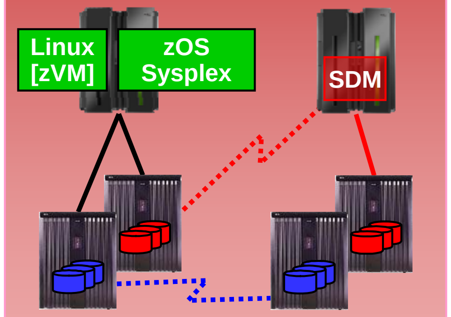
RPO 0 sec & RTO 1-2 min / <1 hr

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

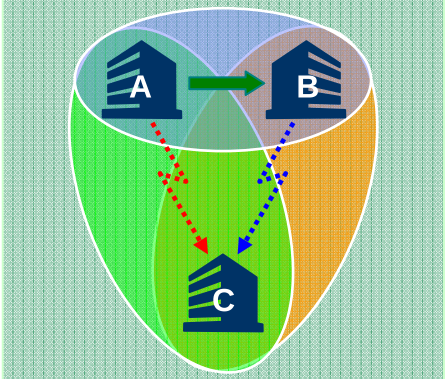
RPO few sec & RTO 1hr

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

RPO 0 sec & RTO 1-2 min / <1 hr
RPO few sec & RTO 1 hr

RPO – recovery point objective (data loss)
RTO – recovery time objective (downtime)

Synch replication →
Asynch replication ⇢

Evolving customer requirements

- Shift focus from failover model to *near-continuous availability* model (RTO near zero)
- Access data from *any site* (unlimited distance between sites)
- 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)

From High Availability to Continuous Availability

GDPS/PPRC	GDPS/XRC or GDPS/GM	GDPS/Active-Active
Failover model	Failover model	Near Continuous Availability model
Recovery time = 2 minutes	Recovery time < 1 hour	Recovery time < 1 minute
Distance < 20 KM	Unlimited distance	Unlimited distance

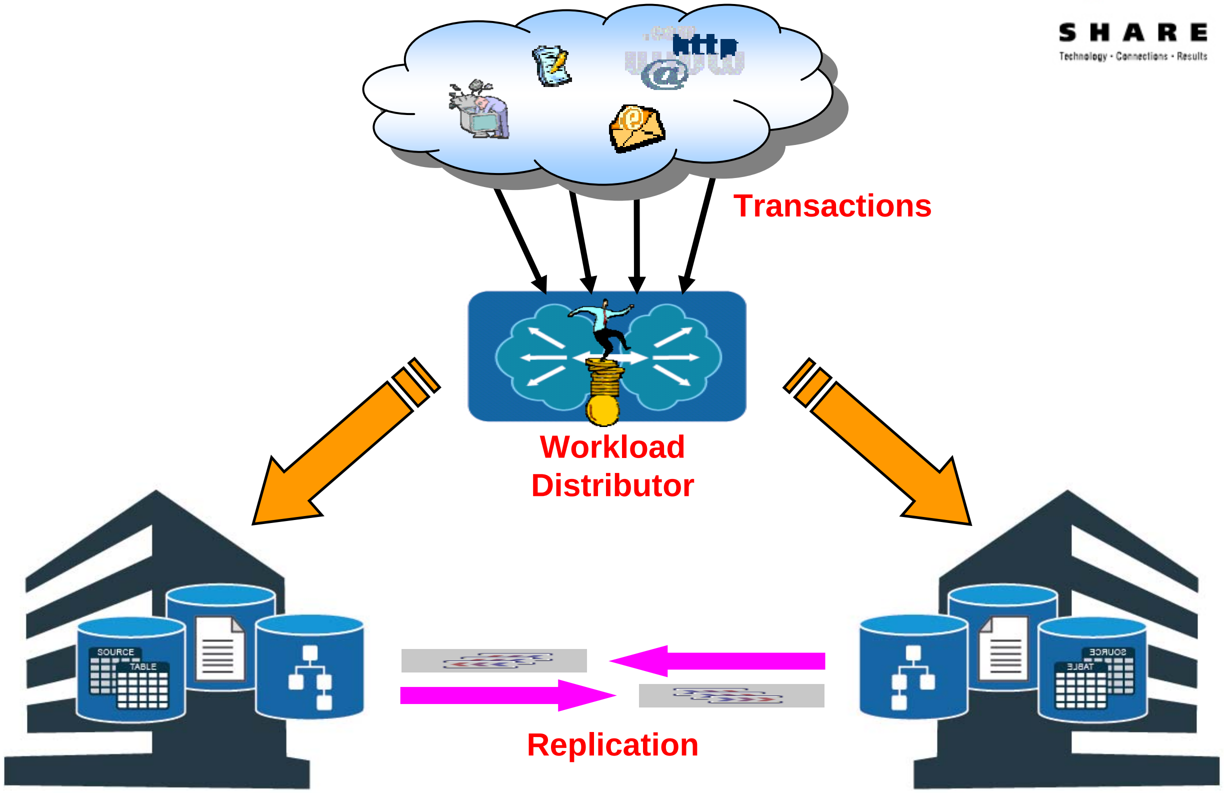
GDPS/Active-Active is for mission critical workloads that have stringent recovery objectives that can not be achieved using existing GDPS solutions.

- RTO approaching zero, measured in seconds for unplanned outages
- RPO approaching zero, measured in seconds for unplanned outages
- Non-disruptive site switch of workloads for planned outages
- At any distance

Active-Active is NOT intended to substitute for local availability solutions such as Parallel SYSPLEX

Active/Active concept

- **Two or more sites, separated by *unlimited* distances, running the same applications and having the same data to provide:**
  - Cross-site Workload Balancing
  - Continuous Availability
  - Disaster Recovery
- **Data at geographically dispersed sites kept in sync via replication**



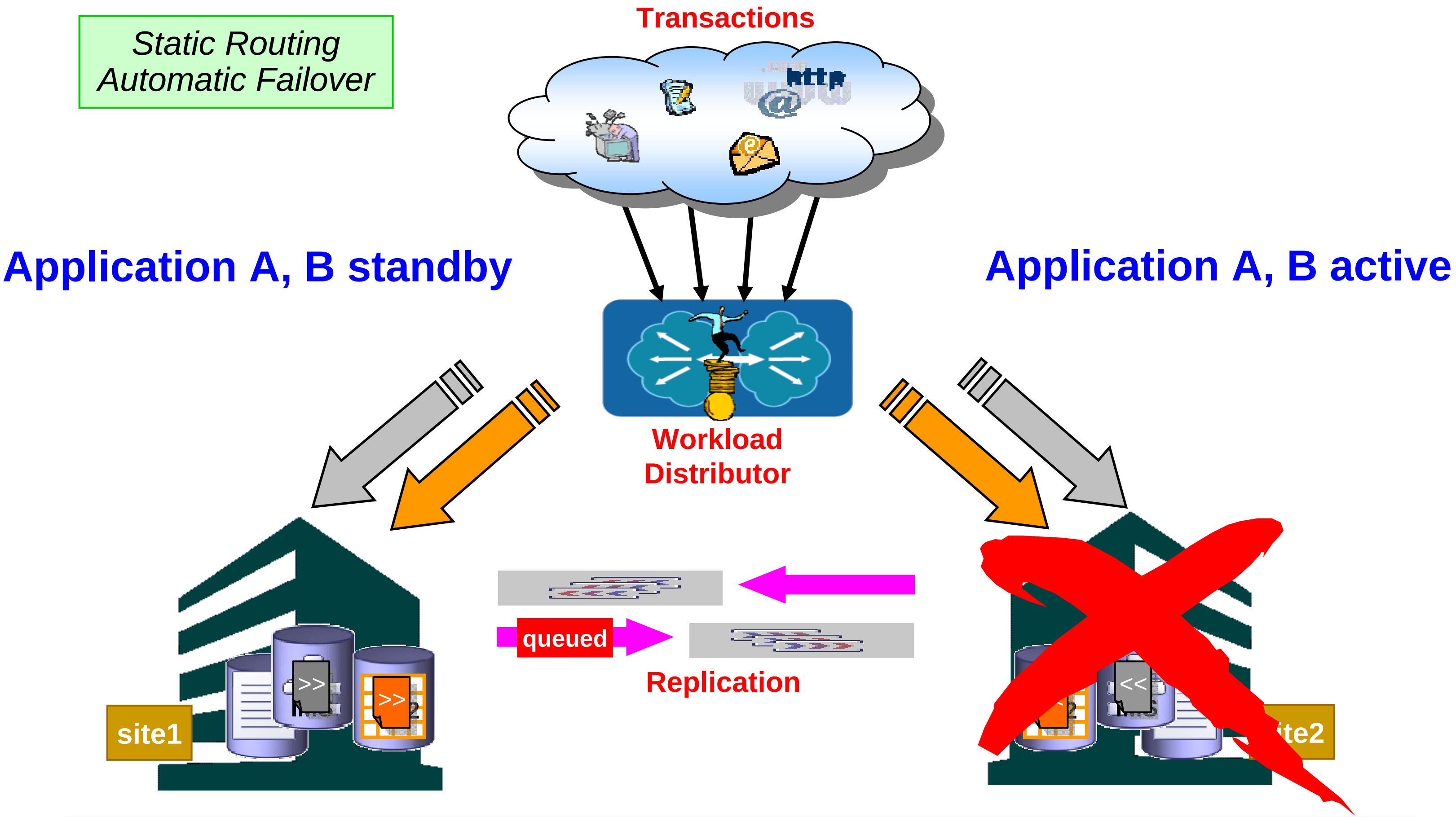
Workloads are managed by a client and routed to one of many replicas, depending upon workload weight and latency constraints; extends workload balancing to SYSPLEXs across multiple sites

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

Active/Active Sites Configurations

- Configurations
  - Active/Standby – GA date 30th June 2011
  - Active/Query – statement of direction
  - Active/Active – intended direction
- A configuration is specified on a workload basis
- A workload is the aggregation of these components
  - ***Software:*** user written applications (eg: COBOL programs) and the middleware run time environment (eg: CICS regions, InfoSphere Replication Server instances and DB2 subsystems)
  - ***Data:*** related set of objects that must preserve transactional consistency and optionally referential integrity constraints (eg: DB2 Tables, IMS Databases)
  - ***Network connectivity*****:** one or more TCP/IP addresses & ports (eg: 10.10.10.1:80)

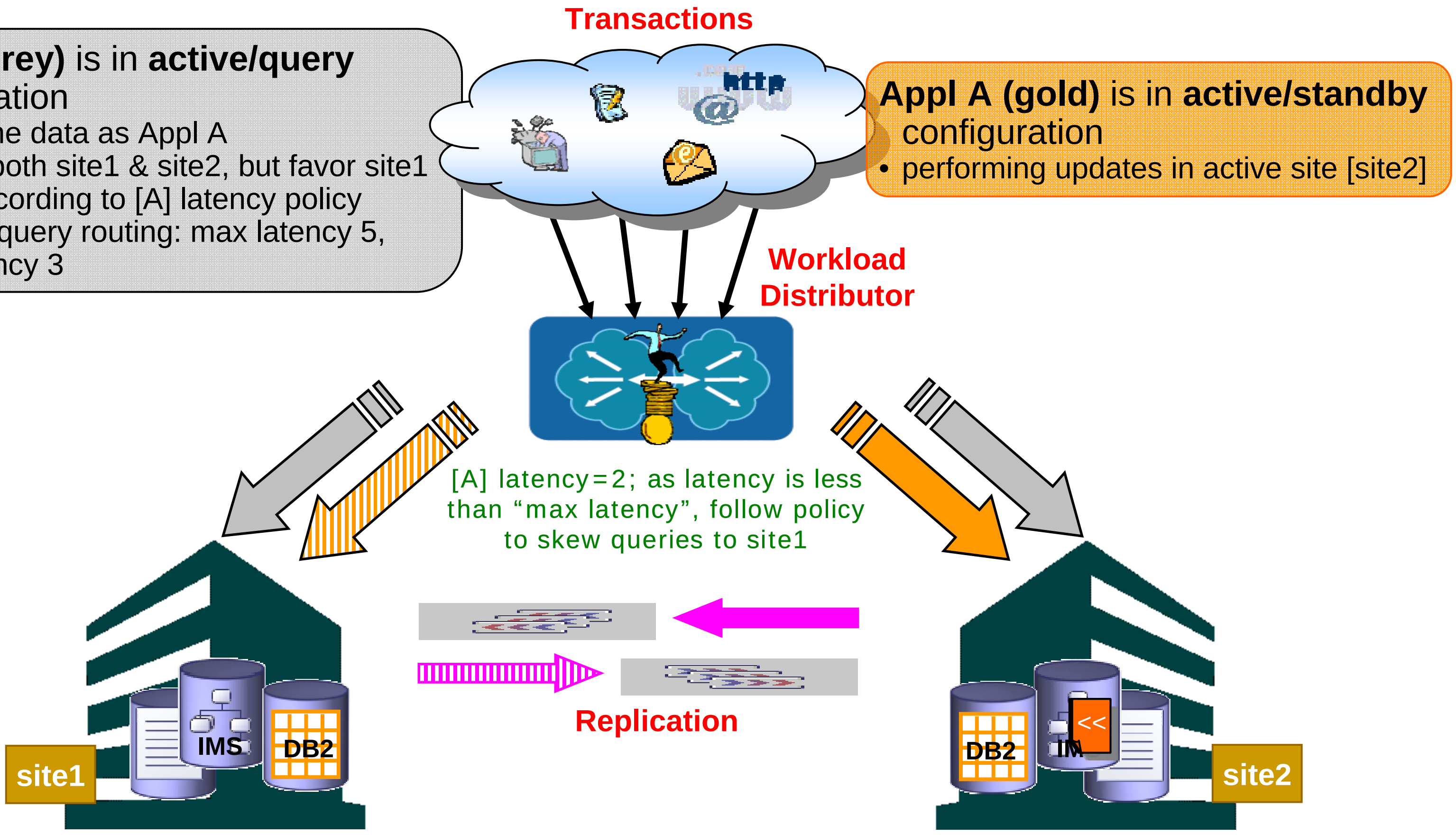
Active/Standby configuration



This is a fundamental paradigm shift from a failover model to a continuous availability model

Active/Query configuration (SOD)

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.



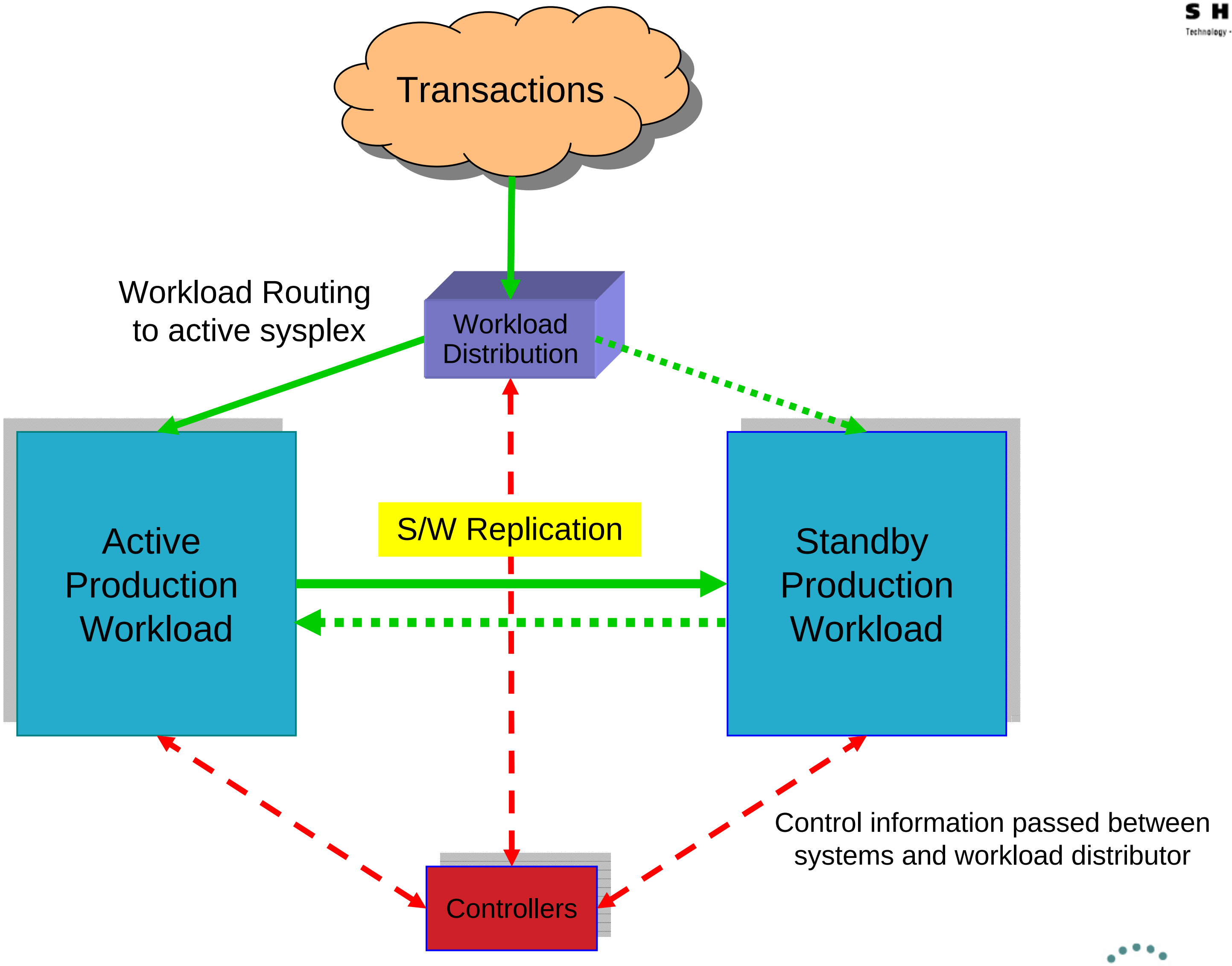
Appl B (grey) is in **active/query** configuration
- using same data as Appl A
- active to both site1 & site2, but favor site1
- routed according to [A] latency policy
- policy for query routing: max latency 5, reset latency 3

Appl A (gold) is in **active/standby** configuration
- performing updates in active site [site2]

[A] latency=2; as latency is less than "max latency", follow policy to skew queries to site1

Read-only or query transactions to be routed to both sites, while update transactions are routed only to the active site

Conceptual view


Transactions
Workload Routing to active sysplex
Workload Distribution
Active Production Workload
S/W Replication
Standby Production Workload
Controllers

Control information passed between systems and workload distributor

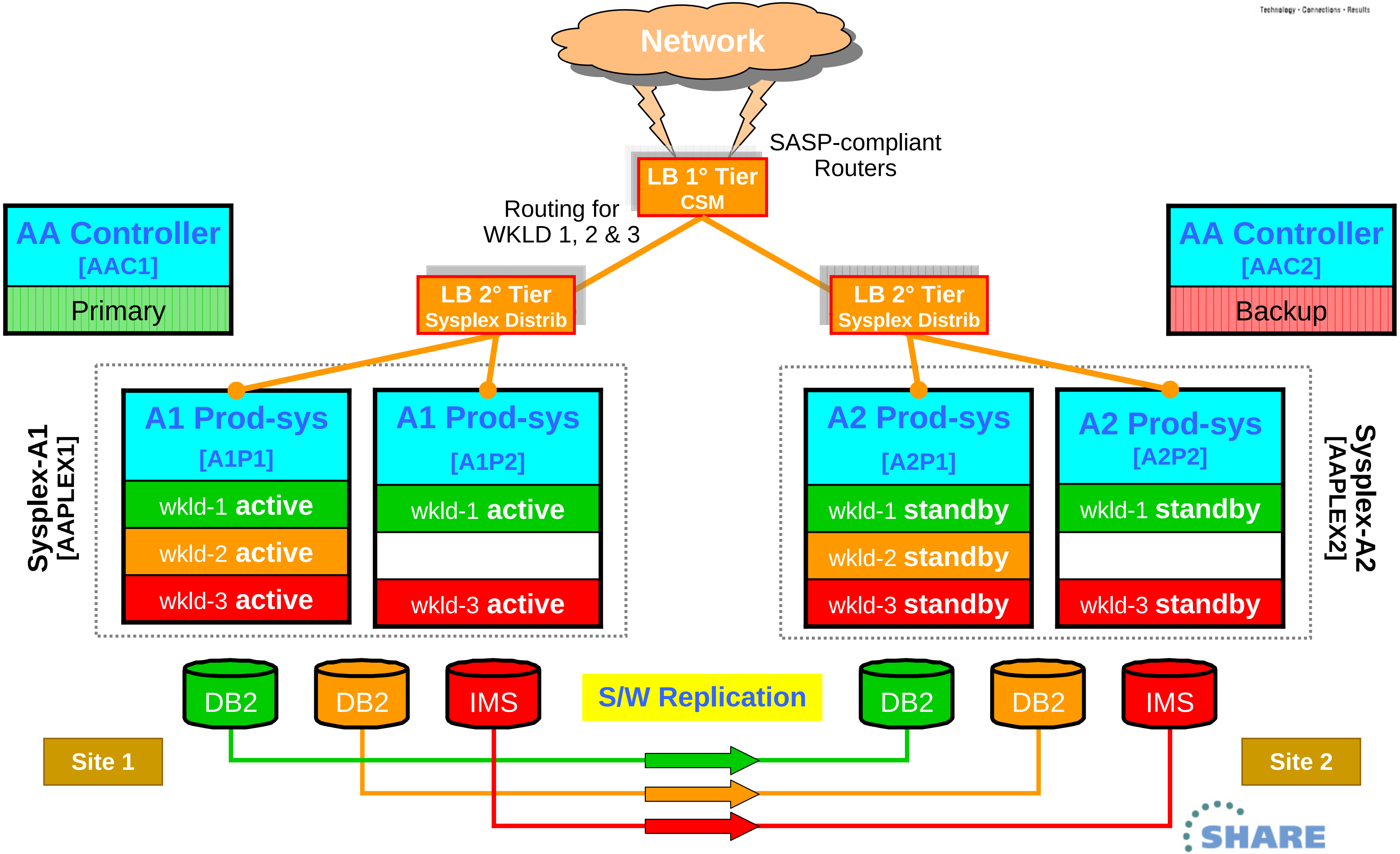
What is a GDPS/Active-Active environment?

- **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 (VSAM SoD)
- **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
  - **What switches/routers are SASP-compliant?**
    - Cisco Catalyst 6500 Series Switch Content Switching Module
    - F5 Big IP Switch
    - Citrix NetScaler Appliance
    - Radware Alteon Application Switch (bought Nortel appliance line)

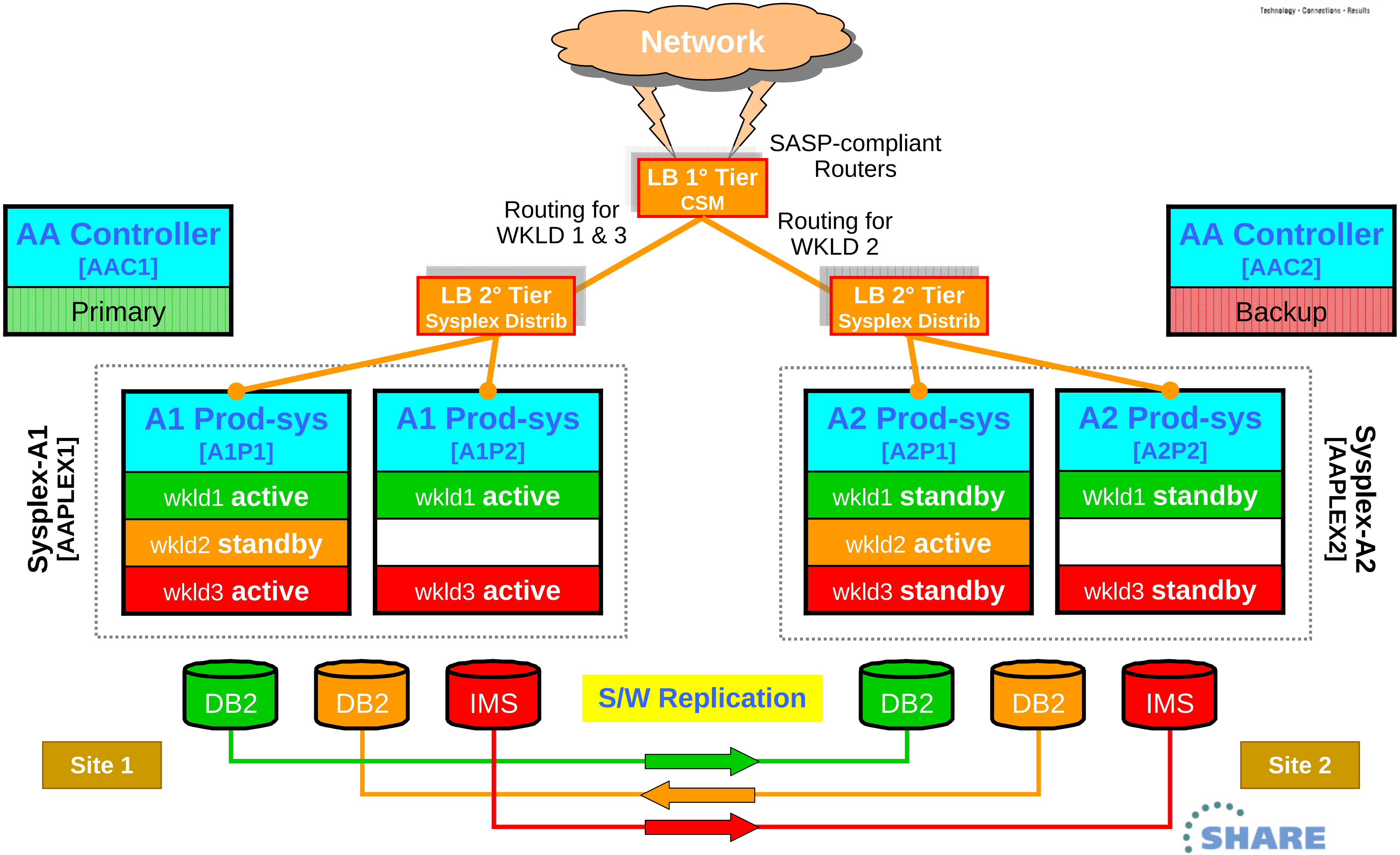
VSAM Statement of Direction

- IBM intends in the future to enhance the IBM Geographically Dispersed Parallel Sysplex™ (GDPS®)/Active-Active continuous availability solution by providing support for software replication of Virtual Storage Access Method (VSAM) data for active-standby and active-query configurations. IBM plans to provide such support for data replication for VSAM data updated by applications that run in CICS and offline in batch mode, using log data provided by CICS Transaction Server for z/OS , V5 or later and CICS VSAM Recovery for z/OS , V5 or later.
- IBM 's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM 's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Sample scenario – all workloads active in one site



Sample scenario – both sites active for individual workloads



What S/W makes up a GDPS/Active-Active environment?

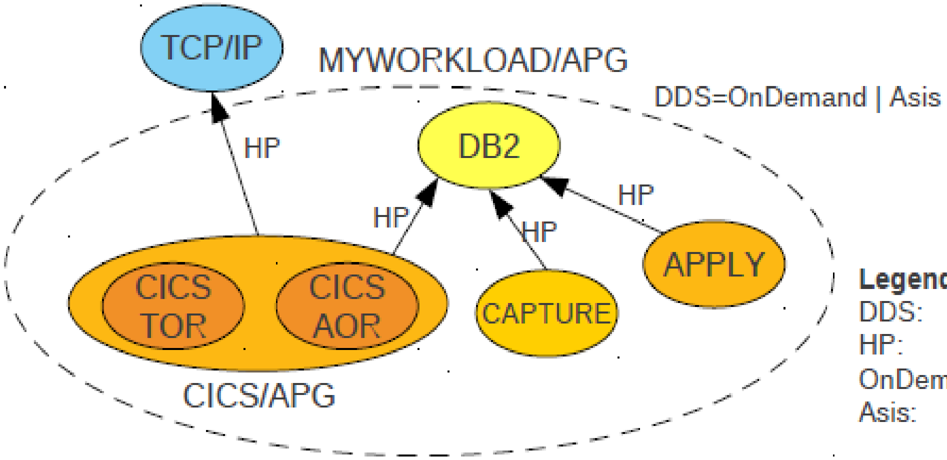
- GDPS/Active-Active
- IBM Tivoli NetView for z/OS
  - IBM Tivoli NetView for z/OS Enterprise Management Agent (NetView agent)
- IBM Tivoli Monitoring
- System Automation for z/OS
- IBM Multi-site Workload Lifeline for z/OS
- Middleware – DB2, IMS, CICS…
- Replication Software
  - IBM InfoSphere Database Replication for DB2 for z/OS
  - IBM InfoSphere IMS Replication for z/OS
- Optionally the Tivoli OMEGAMON XE suite of monitoring products

What is an Active/Active Workload?

- A workload is the aggregation of these components
  - ***Software:*** user written applications (eg: COBOL programs) and the middleware run time environment (eg: CICS regions, InfoSphere Replication Server instances and DB2 subsystems)
  - ***Data:*** related set of objects that must preserve transactional consistency and optionally referential integrity constraints (eg: DB2 Tables, IMS Databases)
  - ***Network connectivity*****:** one or more TCP/IP addresses & ports (eg: 10.10.10.1:80)

Software – deeper insight

- All components of a Workload should be defined in SA* as
  - One or more Application Groups (APG)
  - Individual Applications (APL)
- The Workload itself is defined as an Application Group
- SA z/OS keeps track of the individual members of the Workload's APG and reports a “compound” status to the A/A Controller



Legend:

DDS:	Default Desired Status
HP:	HasParent
OnDemand:	Resource is UNAVAILABLE at IPL time
Asis:	Resource is kept in the state it is at IPL time

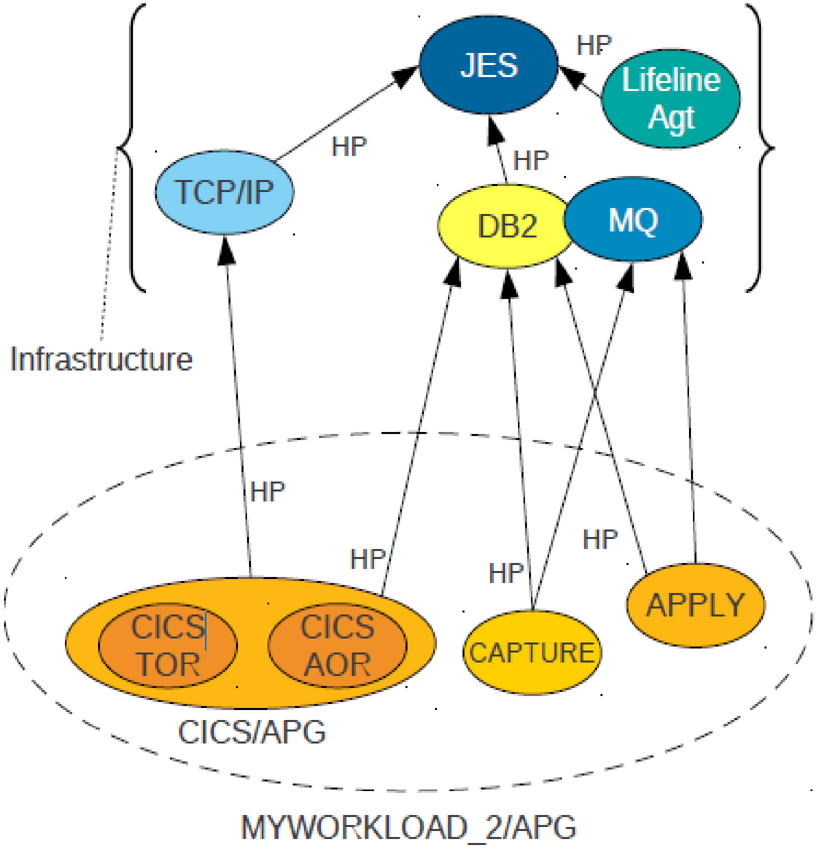
* Note that although SA is required on all systems, you can be using an alternative automation product to manage your workloads.

Software – sharing components between workloads

- Certain components of a workload, for instance DB2, could be also viewed as “infrastructure”

- Relationship(s) from the Workload ensure that the supporting “infrastructure” resources are available when needed

- Infrastructure is typically started at IPL time



Legend:
HP: HasParent

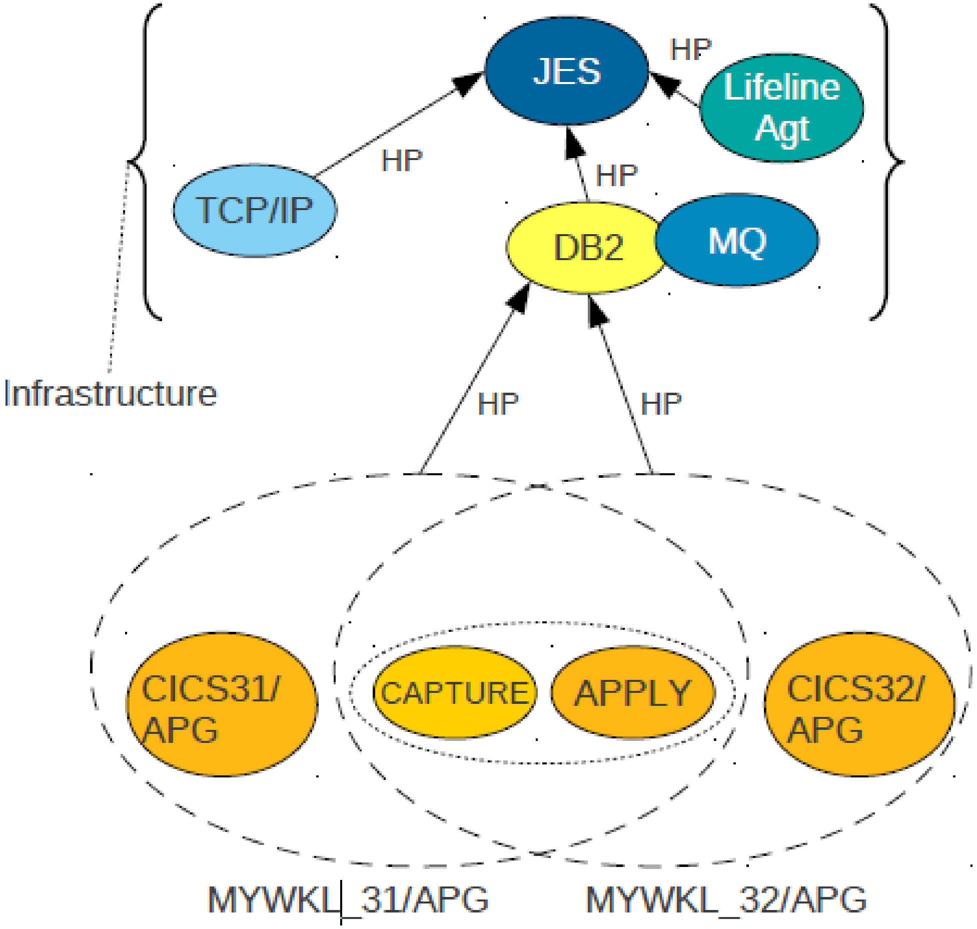
Software – sharing components between workloads

Shared members

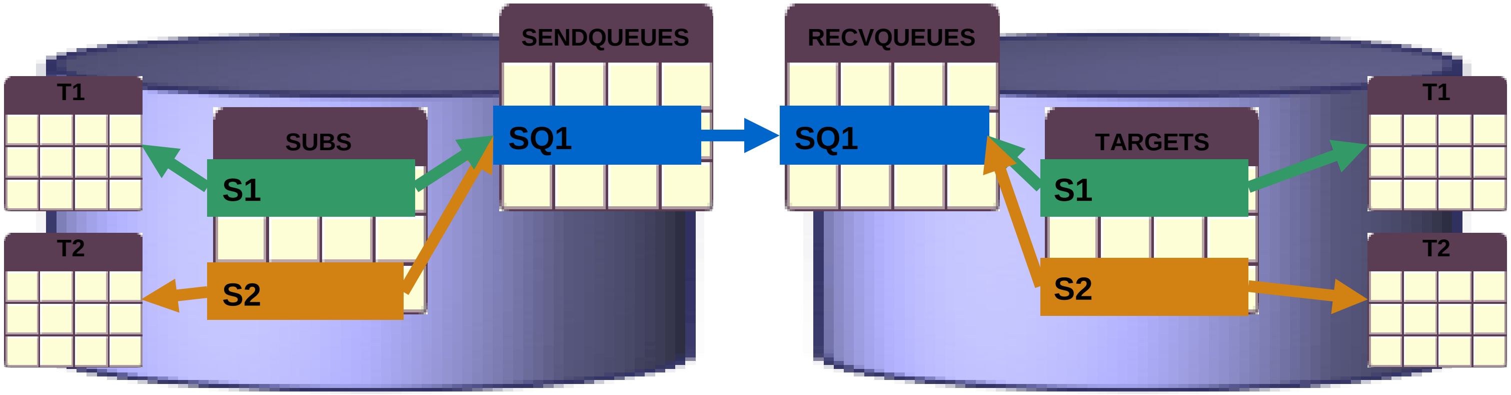
- Other components of a Workload, for instance, capture and apply engines can also be shared
- However, GDPS requires that they are members of the Workload

Rationale

- The A/A Controller needs to know the capture and apply engines that belong to a Workload in order to
  - Quiesce work properly including replication
  - Send commands to them



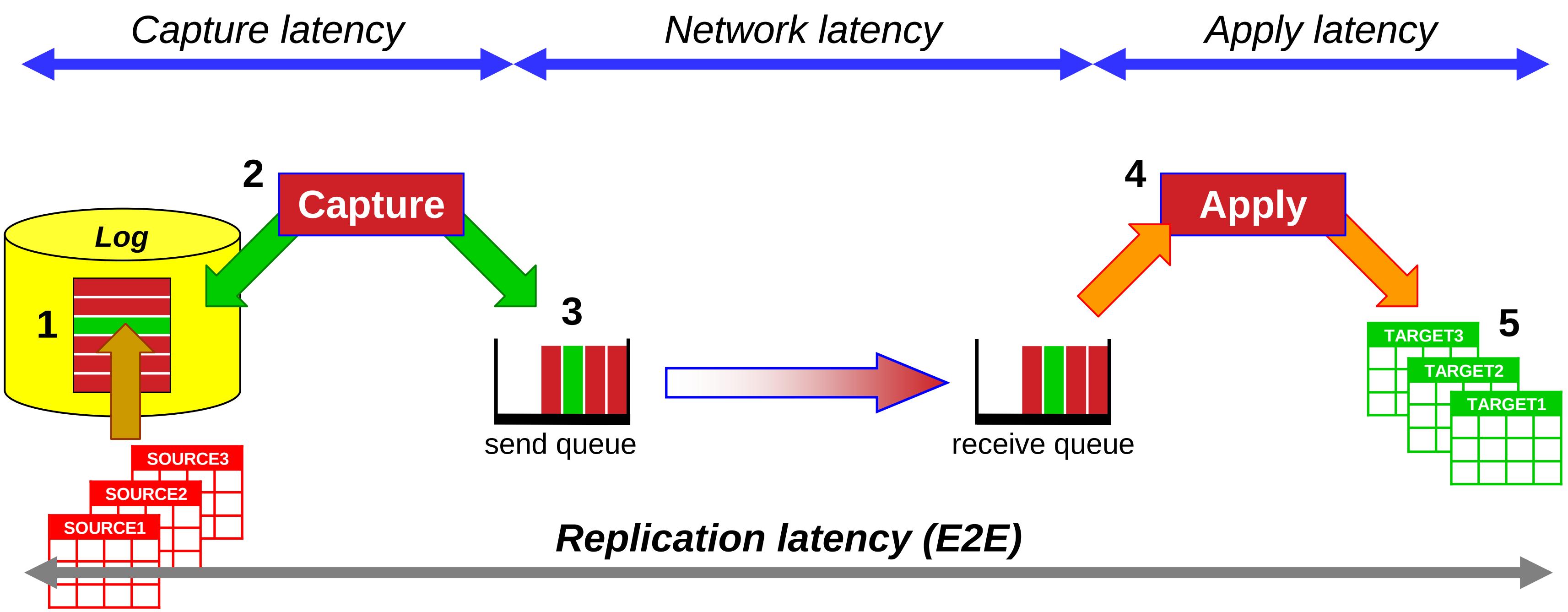
Data – deeper insight



- In DB2 Replication, the mapping between a table at the source and a table at the target is called a ***subscription***
  - Example shows 2 subscriptions for tables T1 and T2
- A subscription belongs to a ***QMap***, which defines the sendq that is used to send data for that subscription
  - Example shows that both subscriptions are using the same QMap (SQ1)

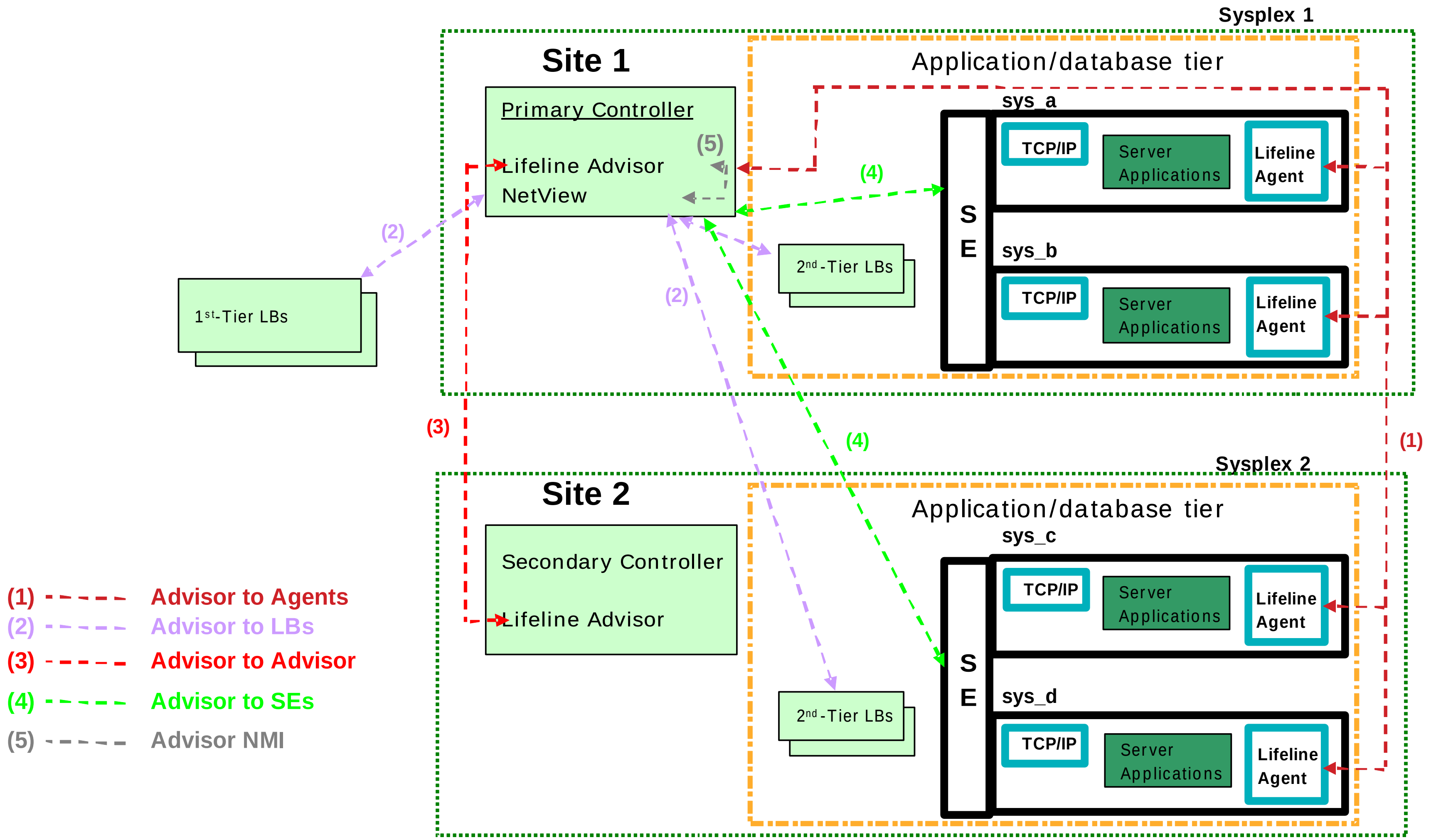
- In IMS Replication, a subscription is a combination of a source server and a target server
  - The subscription is the object that is started/stopped by GDPS/A-A.
  - This corresponds to the QMap in Q Replication
- Each IMS Replication subscription contains a list of replication mappings
  - There is one replication mapping for each IMS database being replicated
  - This corresponds to a subscription in Q Replication

S/W replication technique (for example DB2)



1. Transaction committed
2. Capture read the DB updates from the log
3. Capture put the updates on the send-queue
4. Apply received the updates from the receive-queue
5. Apply copied the DB updates to the target databases

Connectivity – deeper insight



(1) - - - - Advisor to Agents

(2) - - - - Advisor to LBs

(3) - - - - Advisor to Advisor

(4) - - - - Advisor to SEs

(5) - - - - Advisor NMI

Connectivity – deeper insight

- In the IBM Multi-site Workload Lifeline product you must define your workloads:
  - Example:

```
cross_sysplex_list
  {
  10.212.128.151..40000,G0,WORKLOAD_CICSWEB
  10.212.128.118..40000,G1,WORKLOAD_CICSWEB
  10.212.128.151..40001,G0,WORKLOAD_CICSTPCC
  10.212.128.118..40001,G1,WORKLOAD_CICSTPCC
  10.212.128.151..40011,G0,WORKLOAD_IMSTPCC
  10.212.128.118..40011,G1,WORKLOAD_IMSTPCC
  }
```

  - 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
  - Information here must match the definitions in the tier 1 load balancer

Connectivity – deeper insight

- CSS vserver entry – is used by the clients to address the workload. There is one entry for each workload. This is an entry for the workload that listens on port 40000

  Example:

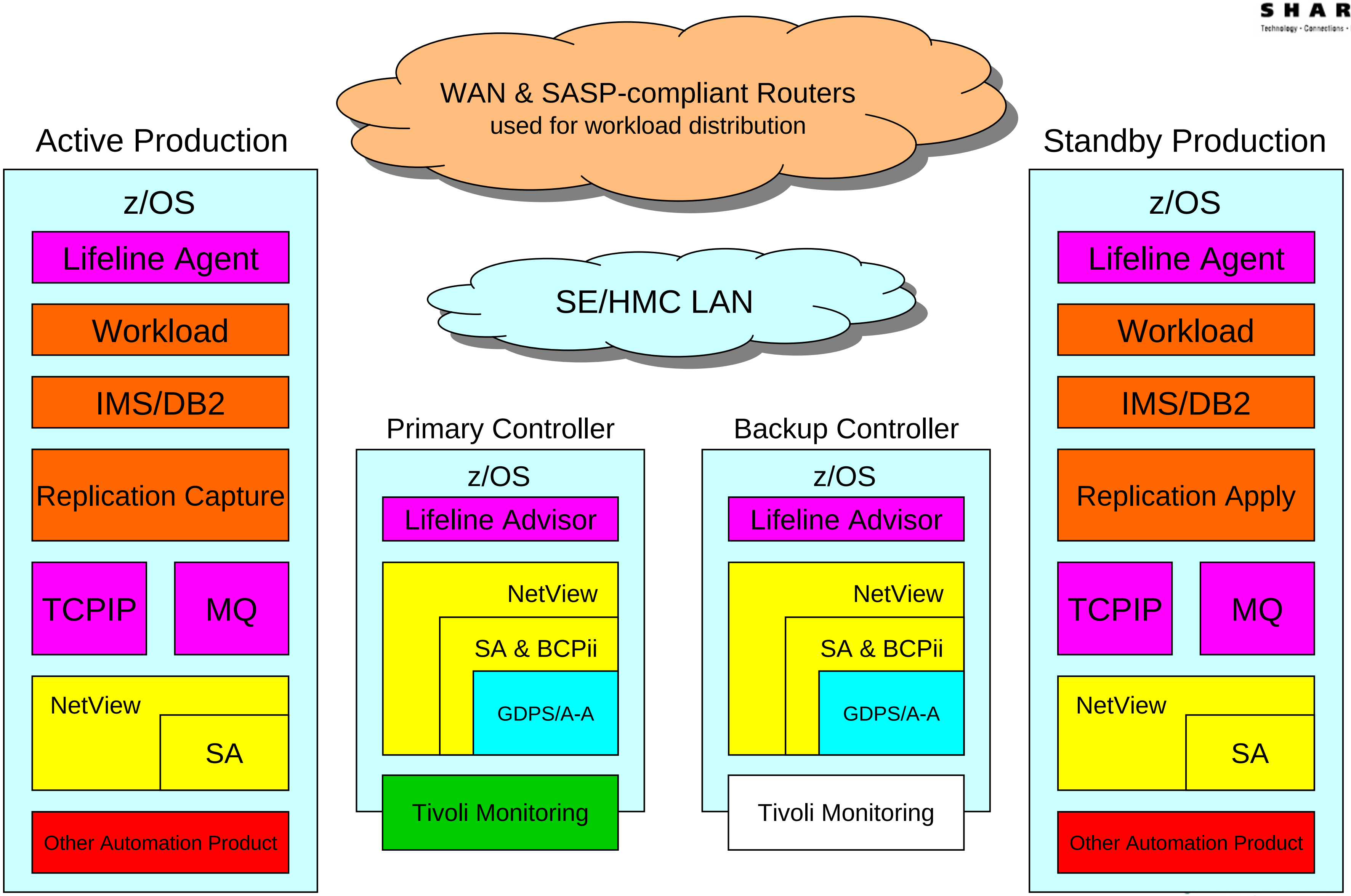
```
vserver GDPS-VSERV-CIC0
    virtual 9.212.135.220 tcp 40000
    no unidirectional
    serverfarm GDPS-FARM-CICS0
    no persistent rebalance
    Inservice
```

- CSS serverfarm entry – is used to specify the IP addresses of the Tier 2 loadbalancers. These are the site specific Sysplex Distributor addresses for the G0 sysplex and G1 Sysplex. This serverfarm entry is referred to in the vserver entry above.

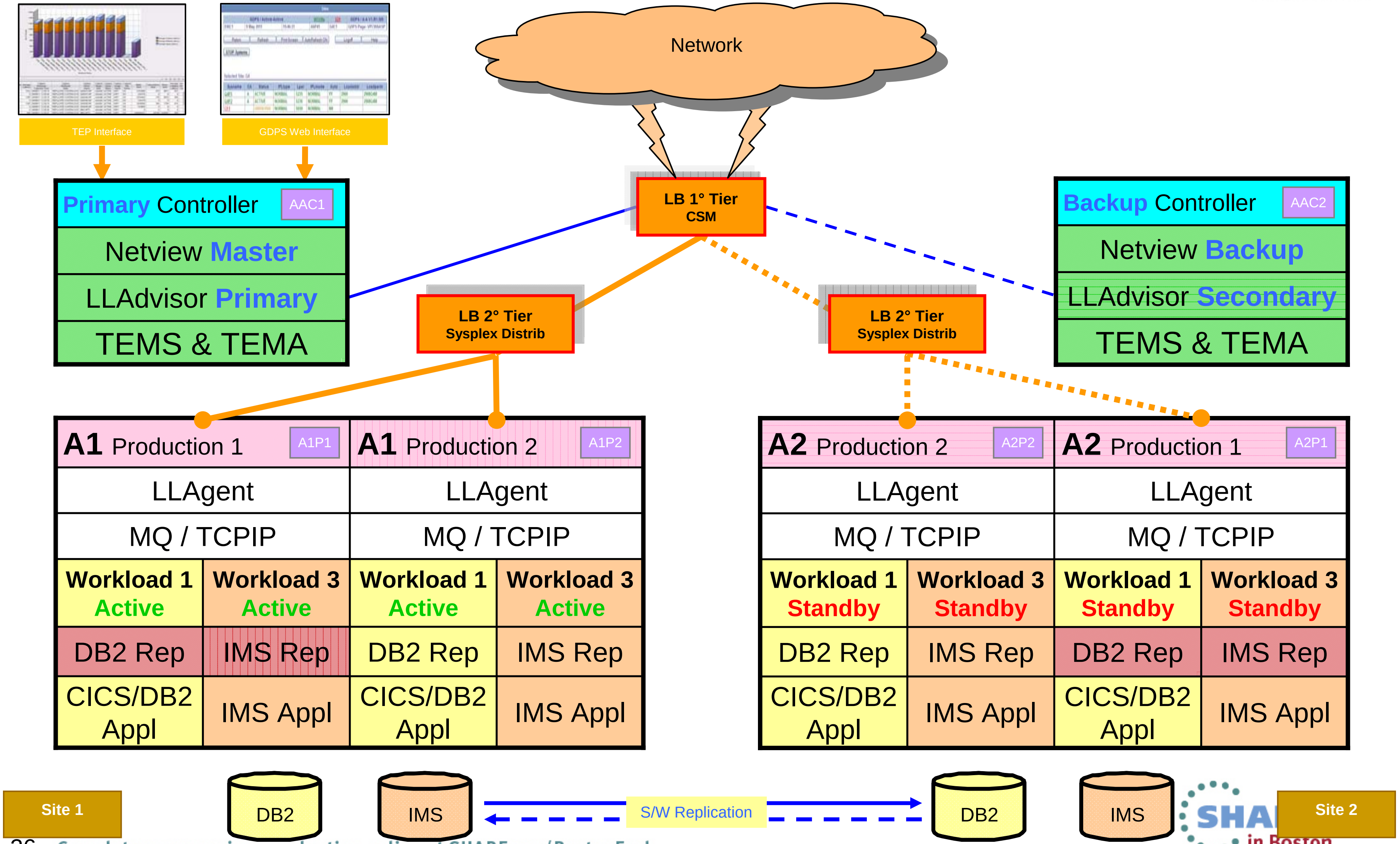
  Example:

```
serverfarm GDPS-FARM-CICS0
    nat server
    nat client CLIENT
    bindid 65520
    real 10.212.128.151
     inservice
    real 10.212.128.118
     Inservice
```

Architectural building blocks



GDPS/A-A configuration



Primary Controller AAC1
Netview Master
LLAdvisor Primary
TEMS & TEMA

Backup Controller AAC2
Netview Backup
LLAdvisor Secondary
TEMS & TEMA

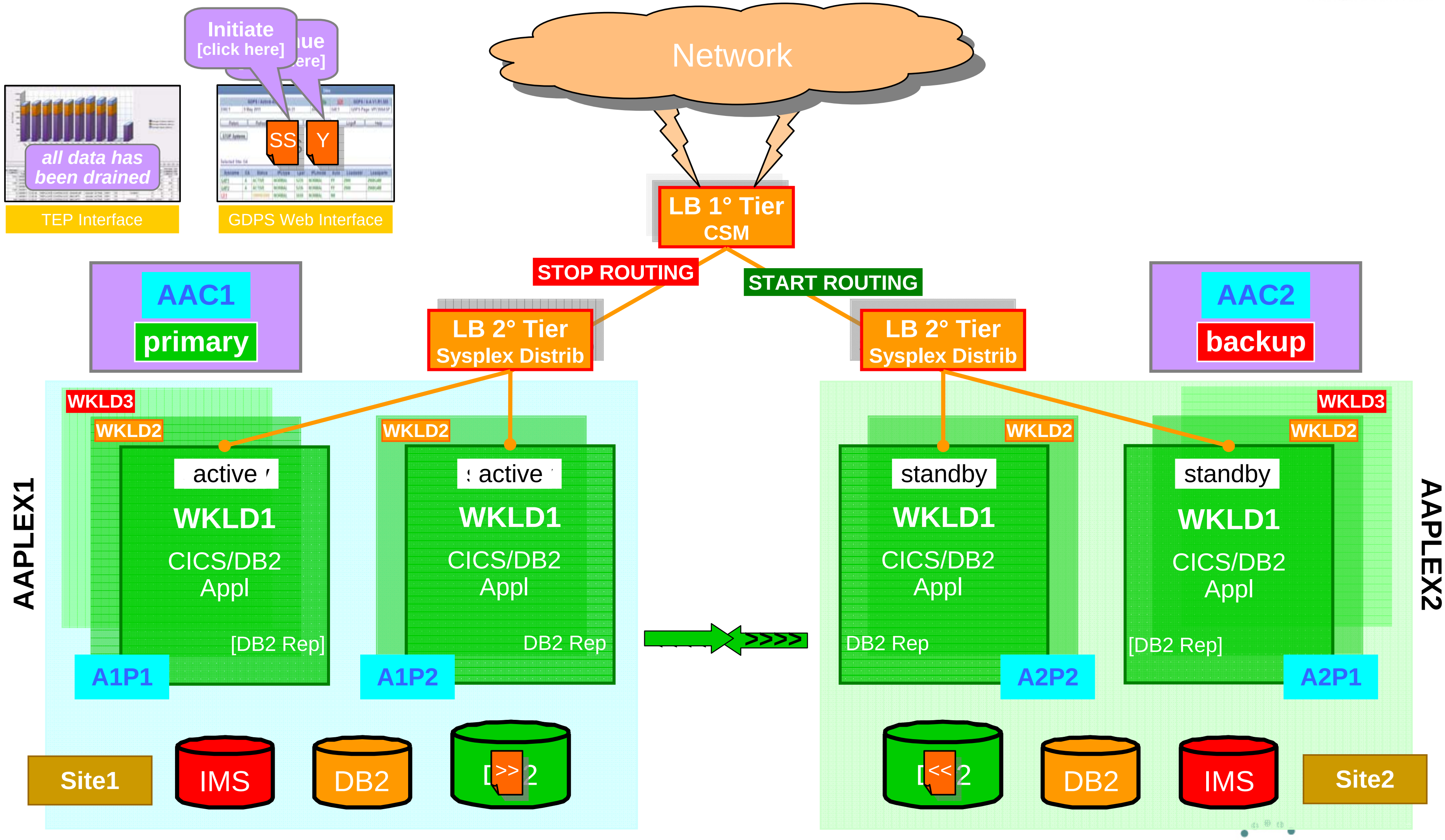
A1 Production 1 A1P1		A1 Production 2 A1P2	
LLAgent		LLAgent	
MQ / TCPIP		MQ / TCPIP	
Workload 1 Active	Workload 3 Active	Workload 1 Active	Workload 3 Active
DB2 Rep	IMS Rep	DB2 Rep	IMS Rep
CICS/DB2 Appl	IMS Appl	CICS/DB2 Appl	IMS Appl

A2 Production 2 A2P2		A2 Production 1 A2P1	
LLAgent		LLAgent	
MQ / TCPIP		MQ / TCPIP	
Workload 1 Standby	Workload 3 Standby	Workload 1 Standby	Workload 3 Standby
DB2 Rep	IMS Rep	DB2 Rep	IMS Rep
CICS/DB2 Appl	IMS Appl	CICS/DB2 Appl	IMS Appl

GDPS/Active-Active (the product)

- **Automation code** is an extension on many of the techniques tried and tested in other GDPS products and with many client environments for management of their mainframe CA & DR requirements
- **Control code** only runs on Controller systems
- **Workload management** - start/stop components of a workload in a given Sysplex
- **Replication management** - start/stop replication for a given workload between sites
- **Routing management** - start/stop routing of transactions to a site
- **System and Server management** - STOP (graceful shutdown) of a system, LOAD, RESET, ACTIVATE, DEACTIVATE the LPAR for a system, and capacity on demand actions such as CBU/OOCoD
- **Monitoring** the environment and **alerting** for unexpected situations
- **Planned/Unplanned situation management and control** - planned or unplanned site or workload switches; automatic actions such as automatic workload switch (policy dependent)
- **Powerful scripting capability** for complex/compound scenario automation

Planned workload/site switch



Note: multiple workloads and needed infrastructure resources are not shown for clarity sake

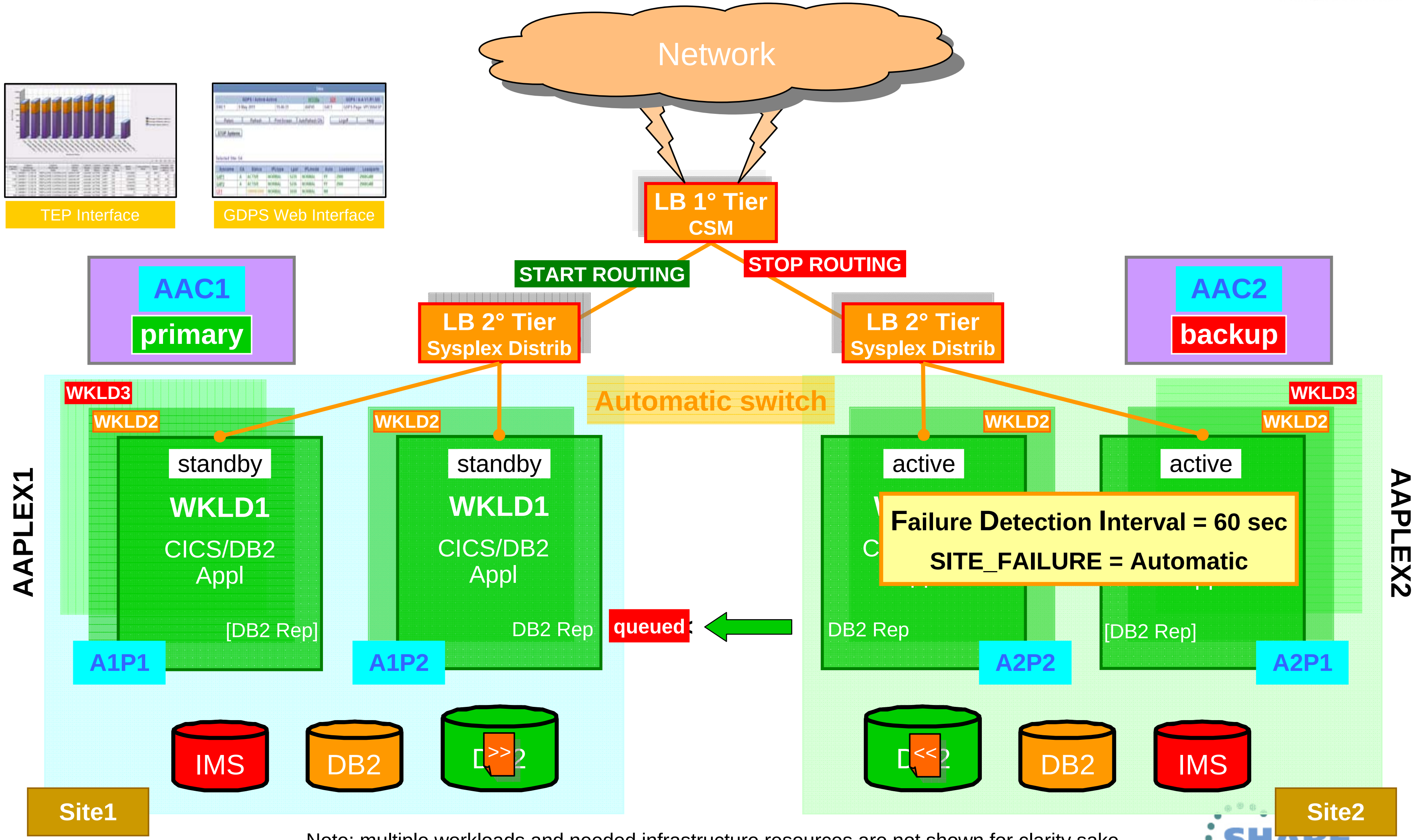
Planned workload/site switch (cont)

```
COMM = 'Switch all workloads to SITE2'
ROUTING = 'STOP WORKLOAD=ALL SITE=AAPLEX1'
ASSIST = 'CHECK ALL WORKLOAD UPDATES REPLICATED'
ROUTING = 'START WORKLOAD=ALL SITE=AAPLEX2'
```

- **Stop routing transactions** to all workloads active to Sysplex AAPLEX1 in Site1
- Wait until all updates on AAPLEX1 are replicated to Sysplex AAPLEX2 in Site2
  - check via the TEP or the Replication Dashboard that all updates have drained from the active to standby site, before stopping replication between the sites
- **Start routing transactions** for workloads previously active in Site1 to Site2
- **Note:** Replication is expected to be active in both directions at all times

The workloads are now processing transactions in Site2 for all workloads with replication from Site2 to Site1

Unplanned site failure



Failure Detection Interval = 60 sec

SITE_FAILURE = Automatic

Note: multiple workloads and needed infrastructure resources are not shown for clarity sake

Go Home scenario

After an unplanned workload/site outage *Note: there is the potential for transactions to have been stranded in the failed site, had completed execution and committed data to the database at the time of the failure, but this data had not been replicated to the standby site.* *Assume the data is still available on the disk subsystems*	**After a planned workload/site outage** *Note: as the process to perform a planned site switch ensures that there are no stranded updates in the active site at the start of the switch, there is no need to start replication in the opposite direction in order to deliver stranded updates.*
Start the **site or workload** that had **failed**	**Start** the **site or workload** that had been **stopped**
Restart replication from the site brought back online to the currently active site - this delivers any stranded changes resulting from the unplanned outage (*)	
Re-synchronize the recovering site with data from the currently active site, by starting replication in the other direction	**Re-synchronize the restarted site or workload with data from the currently active site**, by starting replication from the active to now standby site
Re-direct the workload, once the recovered site is operational and can process workloads	**Re-direct the workload**, once the restarted site is both operational and the data replication has caught up and can now process workloads

(*) attempts to apply the stranded changes to the data in the active site may result in an exception or conflict, as the before image of the update that is stranded will no longer match the updated value in the active site. For IMS replication, the adaptive apply process will discard the update and issue messages to indicate that there has been a conflict and an update has been discarded. For DB2 replication, the update may not be applied, depending on conflict handling policy settings, and additionally an exception record will be inserted into a table.

Testing results*

Configuration:
-**9 * CICS-DB2** workloads + **1 * IMS** workload
-Distance between site 300 miles (≈500kms)

Test1:
Planned site switch

GDPS Active/Active

20 seconds

GDPS/XRC
GDPS/GM

≈ 1-2 hour

Test2:
Unplanned site switch
After a site failure
(Automatic)

GDPS Active/Active

15 seconds

GDPS/XRC
GDPS/GM

≈ 1 hour

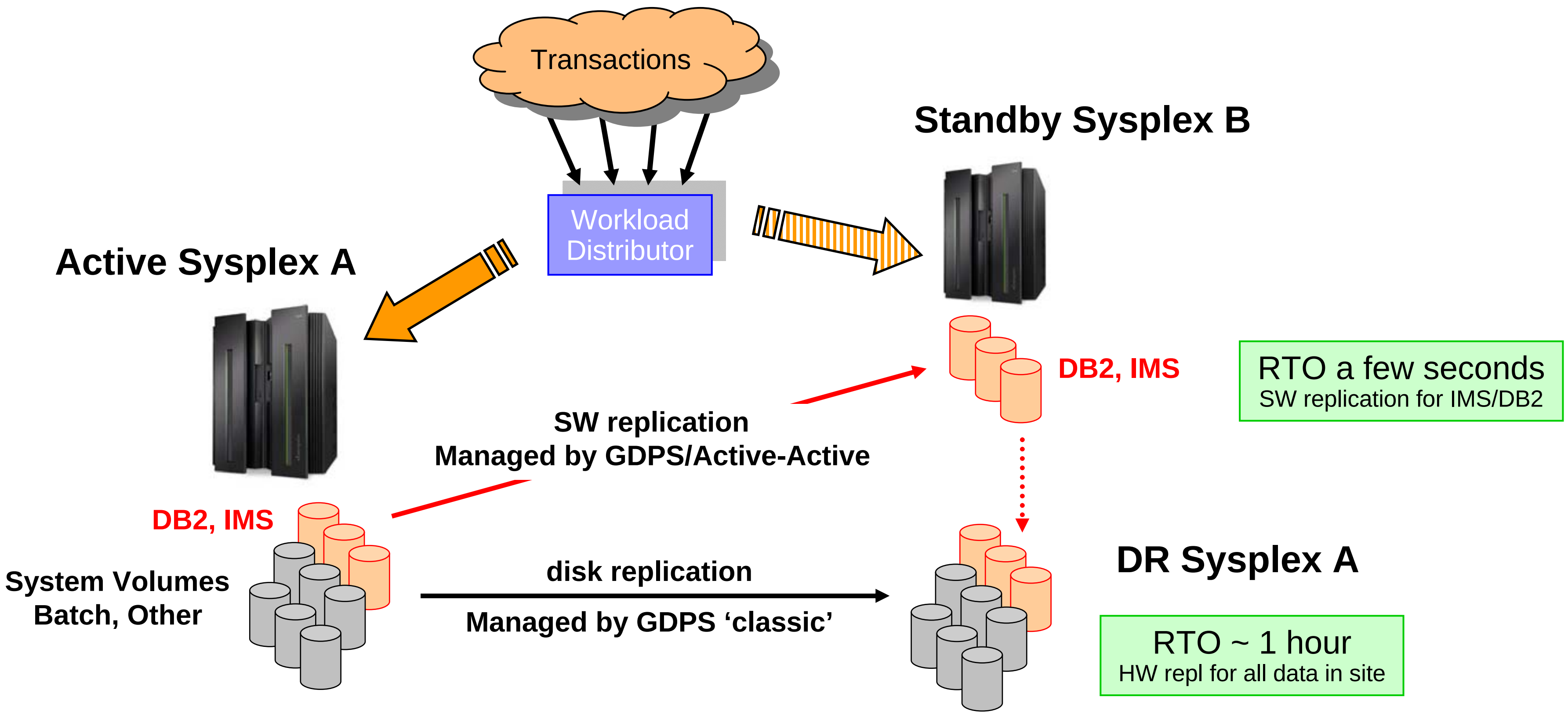
*** IBM laboratory results; actual results may vary.**

Deployment of GDPS/Active-Active

- **Option 1 – create new sysplex environments for active/active workloads**
  - Simplifies operations as scope of Active/Active environment is confined to just this or these specific workloads and the Active/Active managed data
- **Option 2 – Active/Active workload and traditional workload co-exist within the same sysplex**
  - Still will need new active sysplex for the second site
  - Increased complexity to manage recovery of Active/Active workload to one place, and remaining systems to a different environment, from within the same sysplex
  - Existing GDPS/PPRC customer will have to understand operational interactions between GDPS/PPRC and GDPS/Active-Active

No single right answer – will depend on your environment and requirements/objectives

Disk Replication and Software Replication with GDPS



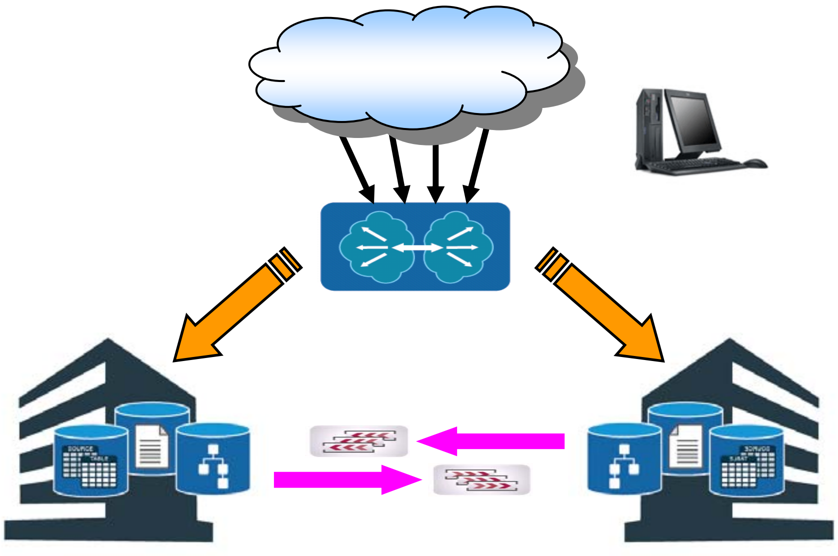
Two switch decisions for Sysplex A problems …

Workload Switch – switch to SW copy (B); once problem is fixed, simply restart SW replication

Site Switch – switch to SW copy (B) <u>and</u> restart DR Sysplex A from the disk copy

Summary

- Manages availability at a workload level
- Provides a central point of monitoring & control
- Manages replication between sites
- Provides the ability to perform a controlled workload site switch
- Provides near-continuous data and systems availability and helps simplify disaster recovery with an automated, customized solution
- Reduces recovery time and recovery point objectives – measured in seconds
- Facilitates regulatory compliance management with a more effective business continuity plan
- Simplifies system resource management



GDPS/Active-Active is the next generation of GDPS

QR Code for Evaluations

Backup Charts

Pre-requisite products

- **IBM Multi-site Workload Lifeline v1.1**
  - Advisor – runs on the Controllers & provides information to the external load balancers on where to send transactions and information to GDPS on the health of the environment
    - There is one primary and one secondary advisor
  - Agent – runs on all production images with active/active workloads defined and provide information to the Lifeline Advisor on the health of that system
- **IBM Tivoli NetView for z/OS v6.1**
  - Runs on all systems and provides automation and monitoring functions. The NetView Enterprise Master normally runs on the Primary Controller
- **IBM Tivoli Monitoring v6.2.2 FP3**
  - Can run on the Controllers, on zLinux, or distributed servers – provides monitoring infrastructure and portal plus alerting/situation management via Tivoli Enterprise Portal, Tivoli Enterprise Portal Server and Tivoli Enterprise Monitoring Server

Pre-requisite products

- **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 or higher**
  - Runs on all images. Provides a number of critical functions:
    - BCPii
    - Remote communications capability to enable GDPS to manage sysplexes from outside the sysplex
    - System Automation infrastructure for workload and server management
- **Optionally the OMEGAMON suite of monitoring tools to provide additional insight**

Pre-requisite software matrix

Pre-requisite software [version/release level]	GDPS Controller	A-A Systems	non A-A Systems
Operating Systems			
z/OS 1.11 or higher	YES	YES	YES
Application Middleware			
DB2 for z/OS V9 or higher	NO	YES [1]	as required
IMS V11	NO	YES [1]	as required
Websphere MQ V7	NO	MQ is only required for DB2 data replication	as required
Replication			
InfoSphere Replication Server for z/OS V10.1	NO	YES [1]	as required [2]
InfoSphere IMS Replication for z/OS V10.1	NO	YES [1]	as required [2]
Management and Monitoring			
GDPS/A-A V1.1	YES	NO	NO
Tivoli NetView for z/OS V6.1	YES	YES	YES
Tivoli System Automation for z/OS V3.3 + SPE APARs	YES	YES	YES
Multi-site Workload Lifeline Version for z/OS 1.1	YES	YES	NO
Tivoli Monitoring V6.2.2 Fix Pack 3	YES	YES	NO

[1] workload dependent

[2] can use Replication Server instances, but not the same instances as the A-A workloads

Pre-requisite software matrix (cont)

Pre-requisite software [version/release level]		GDPS Controller	A-A Systems	non A-A Systems
Optional Monitoring Products				
	IBM Tivoli OMEGAMON XE on z/OS V4.2.0	YES	YES	as required
	IBM Tivoli OMEGAMON XE for Mainframe Networks V4.2.0	YES	YES	as required
	IBM Tivoli OMEGAMON XE for Storage V4.2.0	YES	YES	as required
	IBM Tivoli OMEGAMON XE for DB2 Performance Expert (or Performance Monitor) on z/OS v4.2.0	NO	YES 1)	as required
	IBM Tivoli OMEGAMON XE on CICS for z/OS v4.2.0	NO	YES 1)	as required
	IBM Tivoli OMEGAMON XE on IMS v4.2.0	NO	YES 1)	as required
	IBM Tivoli OMEGAMON XE for Messaging v7.0	NO	YES 1)	as required

1) workload dependent

Note: Details of cross product dependencies are listed in the PSP information for GDPS/Active-Active which can be found by selecting the **Upgrade:GDPS** and **Subset:AAV1R1** at the following URL: http://www14.software.ibm.com/webapp/set2/psearch/search?domain=psp&new=y