Avoiding Perplexation about Sysplex
It's as Easy as A, B, CF!

CF

David Raften
Raften@us.ibm.com

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*
CICS*
DB2*
DFSMSrmm
DFSORT*
Domino
e-business logo*
e-business on demand
Enterprise Storage Server*
ESCON*
FICON
FICON Express

FlashCopy*
HiperSockets
HyperSwap
IBM*
IBM e(logo)server*
IBM eServer
IBM logo*
IMS
iSeries
Language Environment*
Lotus*

OS/390*
Multiprise*
Parallel Sysplex*
Performance Toolkit for VM
PR/SM
RMF
S/390*
Tivoli*
TotalStorage*
VSE/ESA
WebSphere*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries

UNIX is a registered trademark of The Open Group in the United States and other countries.

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation in the United States, other countries, or both.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

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

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.

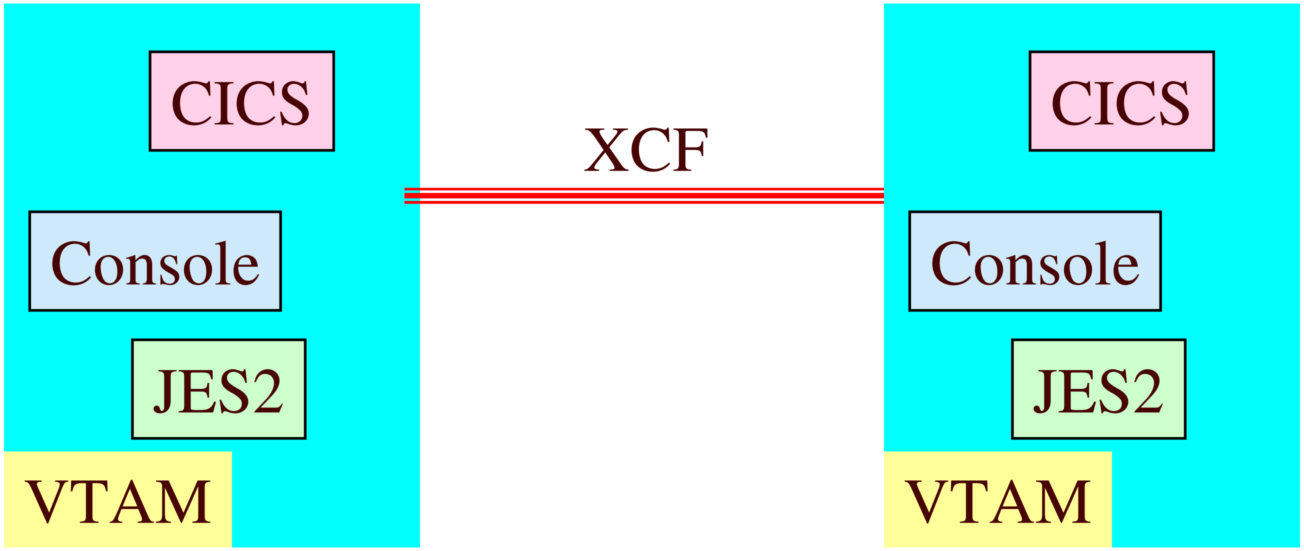
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda

- **Base Sysplex**
  - WLM
- **Parallel Sysplex Overview**
- **Parallel Sysplex Software**
- **Parallel Sysplex Hardware**
  - Coupling Facility
  - System z Exploitation
- **Server Time Protocol (STP)**

Sysplex History

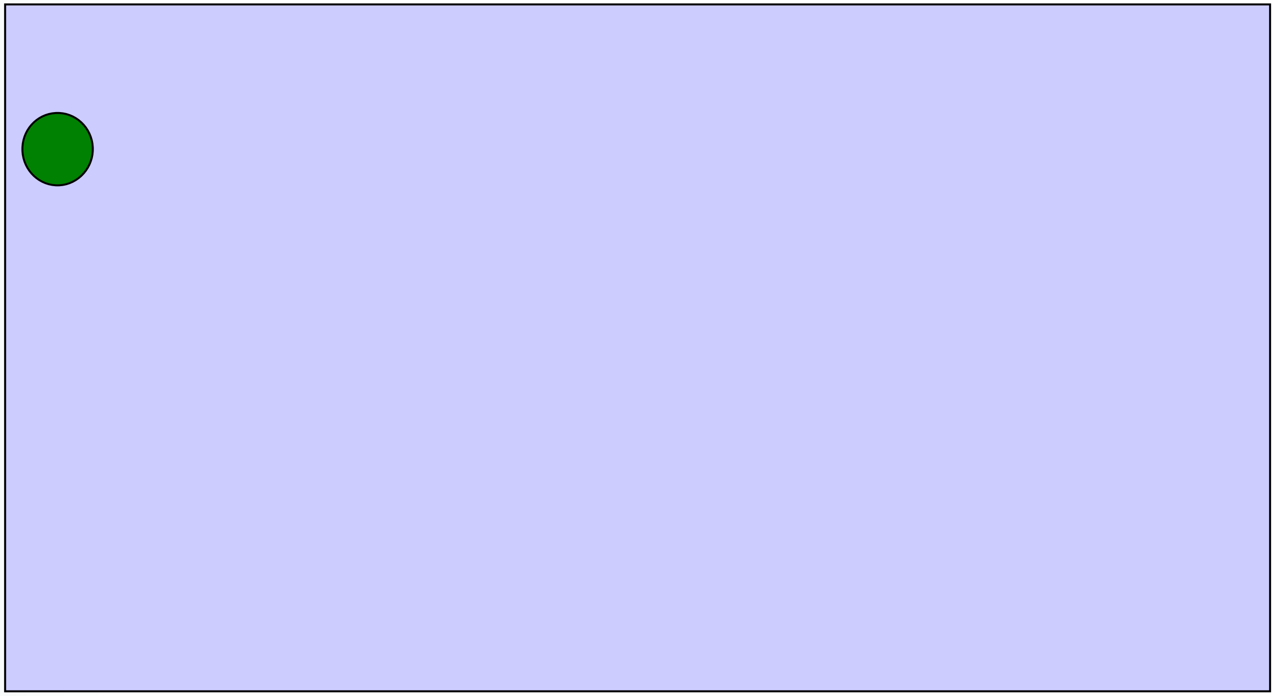
- **(Base) Sysplex – MVS V4.1**
  - 1990
  - XCF - Allows communication between authorized programs



(Base) Sysplex Users

Consoles	Multi-system Consoles
GRS	Dynamic RNL
OPC/ESA	Hot Standby
CICS	MRO communication
JES2	Automatic reset of Checkpoint
RACF	RVARY and SETROPTS command
PDSE	PDSE sharing
DAE	Multi-system DAE
VTAM	Avoid dedicated CTCs
zFS	zFS sharing
Workload Manager (WLM)	
Sysplex Failure Manager (SFM)	
Automatic Restart Manager (ARM)	

Managing Multiple Workloads



High Priority Transactions

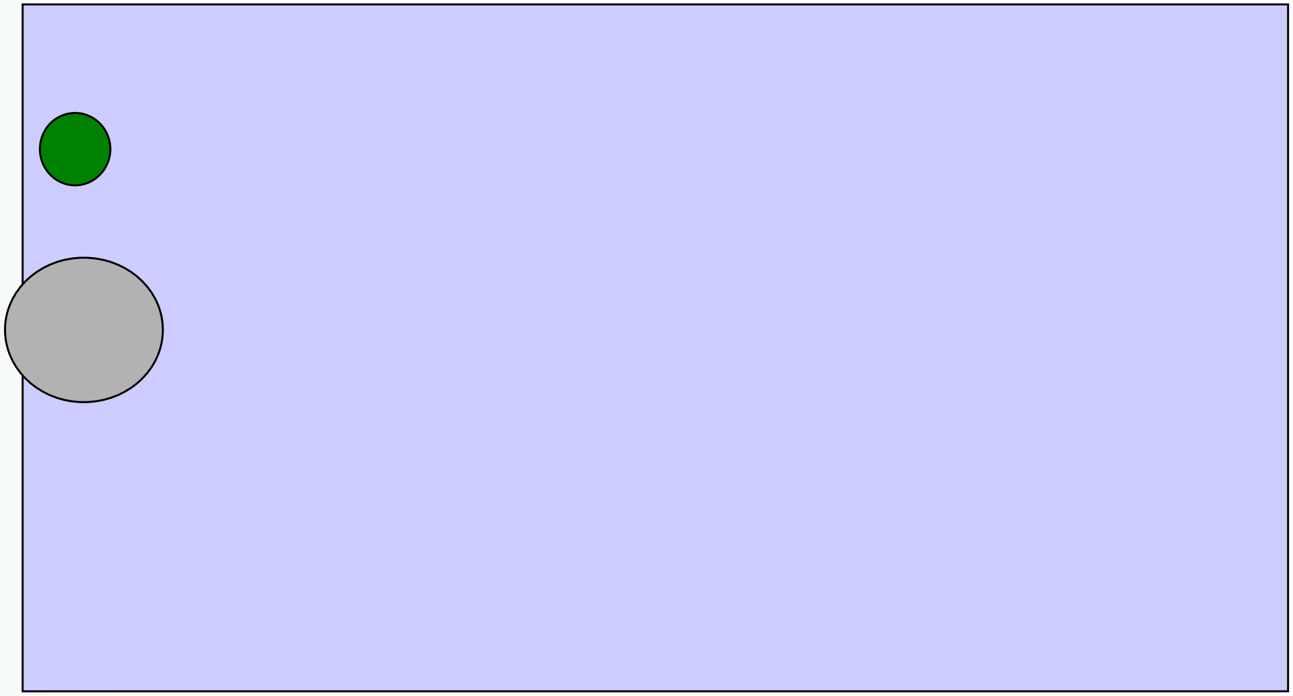
Medium Priority Analysis

Low Priority Batch

Workloads can affect one another. A long running lower priority workload might affect higher priority workloads.

IBM

Managing Multiple Workloads



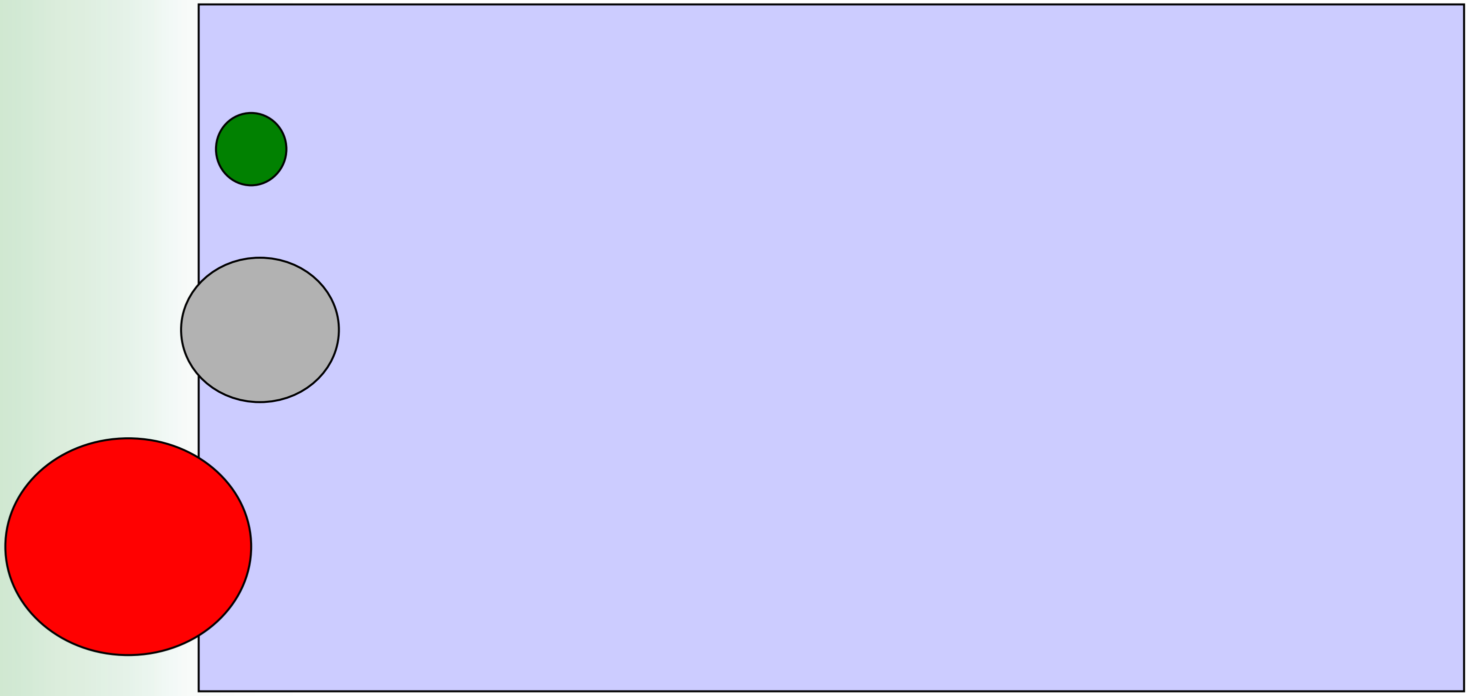
High Priority Transactions

Medium Priority Analysis

Low Priority Batch

Workloads can affect one another. A long running lower priority workload might affect higher priority workloads.

Managing Multiple Workloads



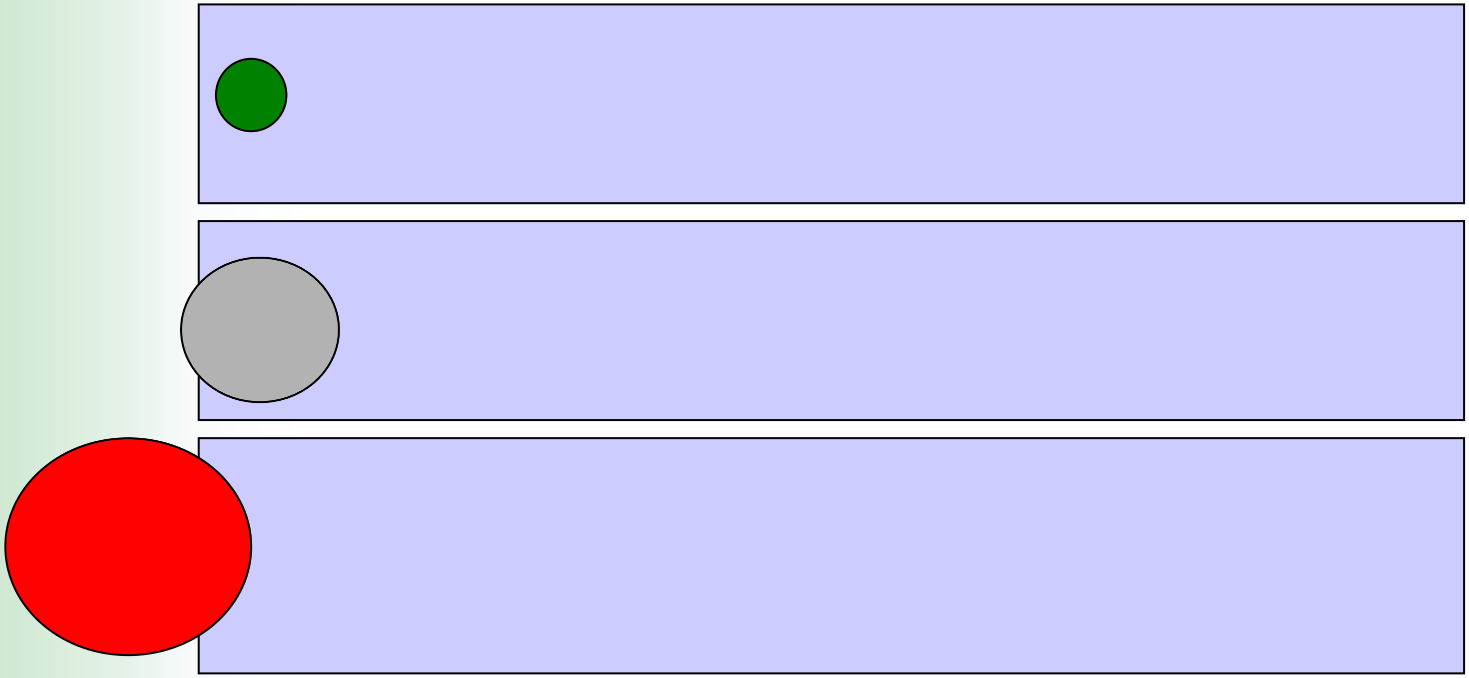
High Priority Transactions

Medium Priority Analysis

Low Priority Batch

Workloads can affect one another. A long running lower priority workload might affect higher priority workloads.

Managing Multiple Workloads



High Priority Transactions

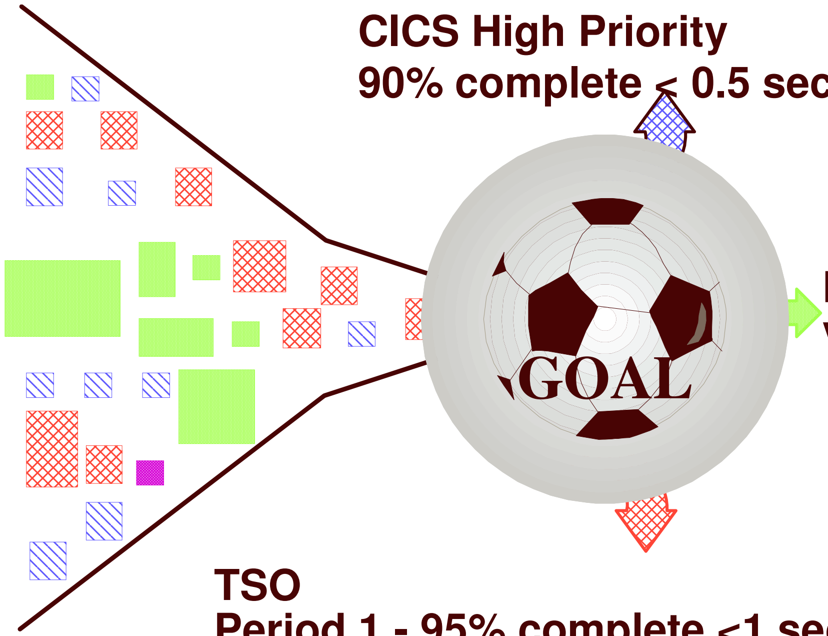
Medium Priority Analysis

Low Priority Batch

Workloads can affect one another. A long running lower priority workload might affect higher priority workloads.

Business Goal Oriented

- **Transaction Type**
  - Web "buy" vs "browse"
  - B2B
  - Batch payroll
  - Test
- **User / User type**
  - Top clients
  - Typical clients
  - Executive
  - Design team
- **Time Periods**
  - Prime shift
  - Off shift weekday
  - Weekends
  - End of quarter



CICS High Priority
90% complete ≤ 0.5 sec

Batch
Velocity 40

TSO
Period 1 - 95% complete <1 sec
Period 2 – 50% complete < 5 sec

WLM

- **Goal Types**
  - Response time – Average or percentile response time
  - Velocity – % without being delayed for processor or storage
  - Importance – 1 (highest) □ 5 □ Discretionary (lowest)

- **Resources Managed**
  - CPU (Dispatching Priority)
  - I/O Priority
  - Storage allocation

WLM Involvement

- Address Spaces dispatching
- WLM Managed Batch Initiators
- Resource Affinities
- Sub-capacity Licensing
- Parallel Access Volumes (PAV)
- Intelligent Resource Director (IRD)
- zIIP / zAAP dispatching

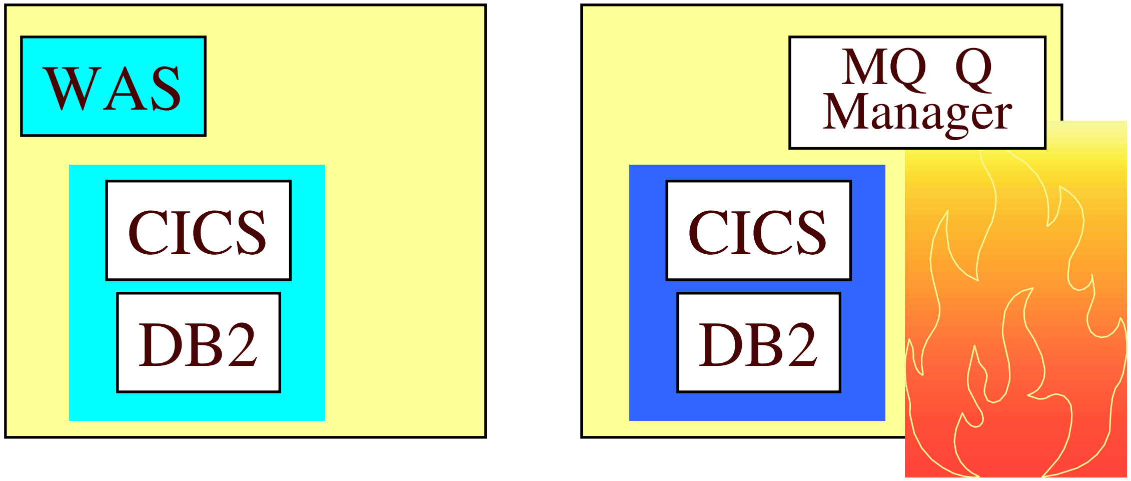
- Dynamic Workload Balancing recommendations

Sample Service Class Definitions

Srvclass	Descript.	Workload	RG	Per	Dur	Imp	Goal
APPN	APPN/MVS users	ASCH	101	1	500	2	80% .5 sec
			102	2		4	Velocity 30
OMVS	OMVS users	OMVS	103	1	500	2	80% .5 sec
			104	2		4	Velocity 30
ONLPRDHI	Prod High	ONLINE		1		1	90% .5 sec
ONPRDMD	Prod Med	ONLINE		1		2	80% 3.0 sec
ONPRDLO	Prod Lo	ONLINE		1		3	50% 10.0 sec
ONLTEST	Test	ONLINE		1			Discretionary
PRDBATHI	Batch High	PRDBAT		1		2	Velocity 30
PRDBATLO	Batch Low	PRDBAT		1			Discretionary
TSOPRD	TSO users	TSO	105	1	500	2	80% .5 sec
			106		2000	3	80% 2.0 sec
			107			5	50% 10.0 sec

Automatic Restart Manager

- **Minimized outage time**
  - Not message driven
  - No operator intervention required
- **Awareness of the state of the sysplex**
- **Groups restarted to system with most available storage**
- **Assists automation products**
- **ARM Wrapper available**



Sysplex Failure Manager (SFM)

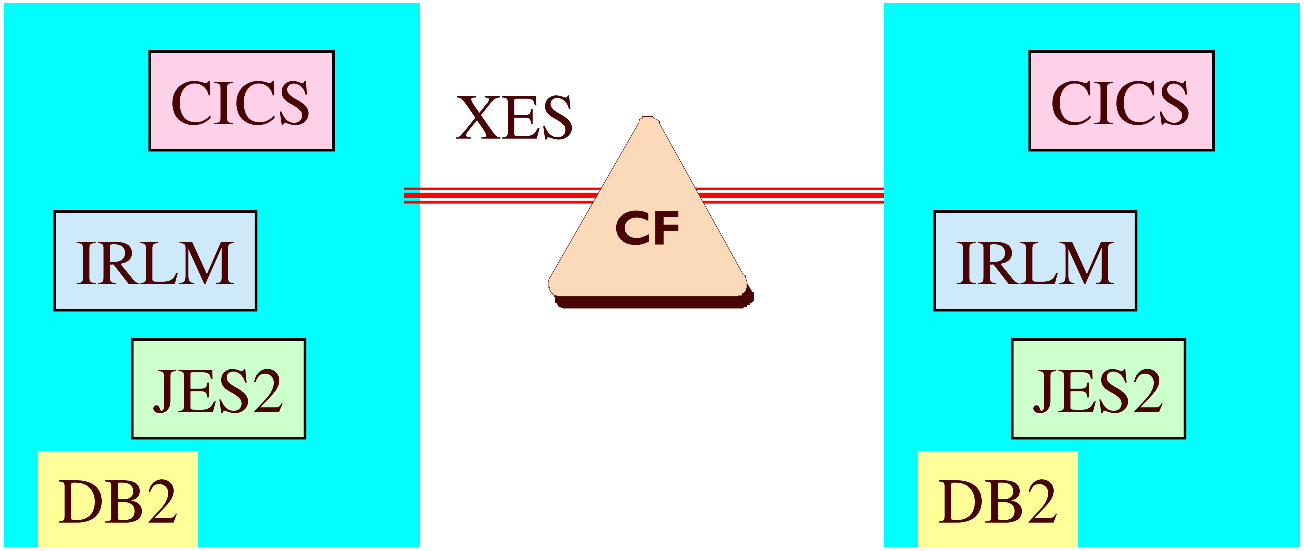
- Automate the planned and unplanned removal of z/OS systems from the sysplex:
  - VARY XCF,sysname,OFFLINE Command
  - I/O Reset
  - System Cleanup (CDS, Locks, …)

- System Failures
  - Status Update Missing condition (Missing heartbeat)
  - Loss of intersystem signaling connectivity
- Enables the REBUILDPERCENT function of CFRM.
- Reconfigures storage to backup LPAR after failed system removed

Sysplex History …

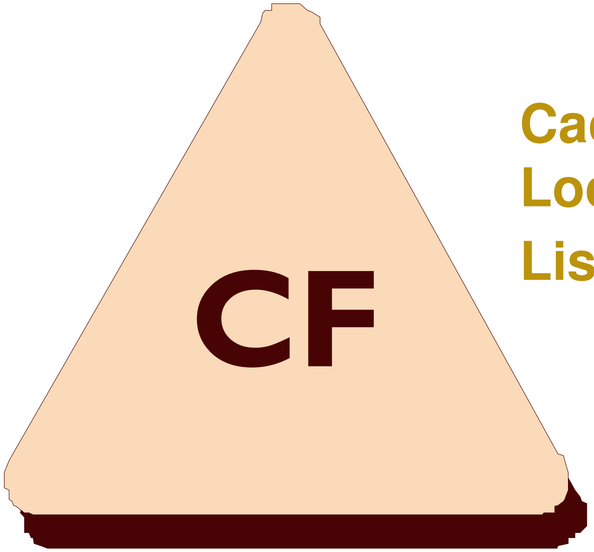
- **(Base) Sysplex – MVS V4.1 (1990)**
  - XCF - Allows communication between authorized programs

- **Parallel Sysplex – MVS V5.1 (1994)**
  - XES – Allows communication between authorized program and CF



Coupling Facility

- **Coupling Facility**
  - Just an LPAR
  - Runs CFCC “LICC”
  - CFCC LPAR can be on stand-alone server or with other LPARS (ICF)
  - Manages structured storage “Structures”



Cache - Used to manage local buffers (DB2, IMS, CAS)

Lock - Used by lock managers (IRLM, GRS)

List - Message passing (XCF, JES, CICS, Logger)

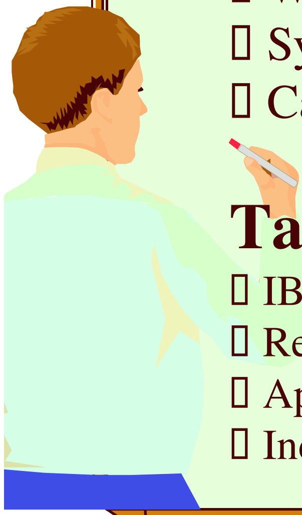
Why Parallel Sysplex?

ES9000

- **ES9000 (1990)**

The System z® Parallel Sysplex Clustering Solution

- Dynamic workload balancing
- Continuous application availability
- Incremental growth



Strategic benefits – The Ultimate

□ Application availability (Planned / Unplanned)

□ Data Accessibility with responsiveness

□ Scalability

□ Workload Management

□ Systems Operation - SSI

□ Capacity

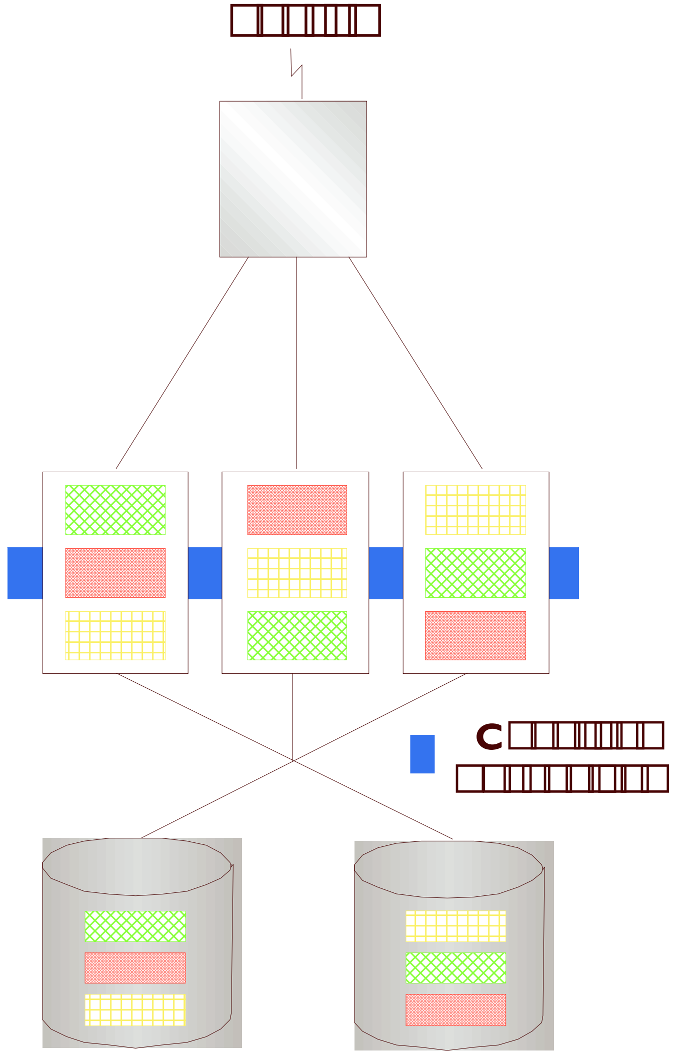
Tactical Benefits

□ IBM Software license savings

□ Reduced cost of ownership

□ Application scalability

□ Industry direction

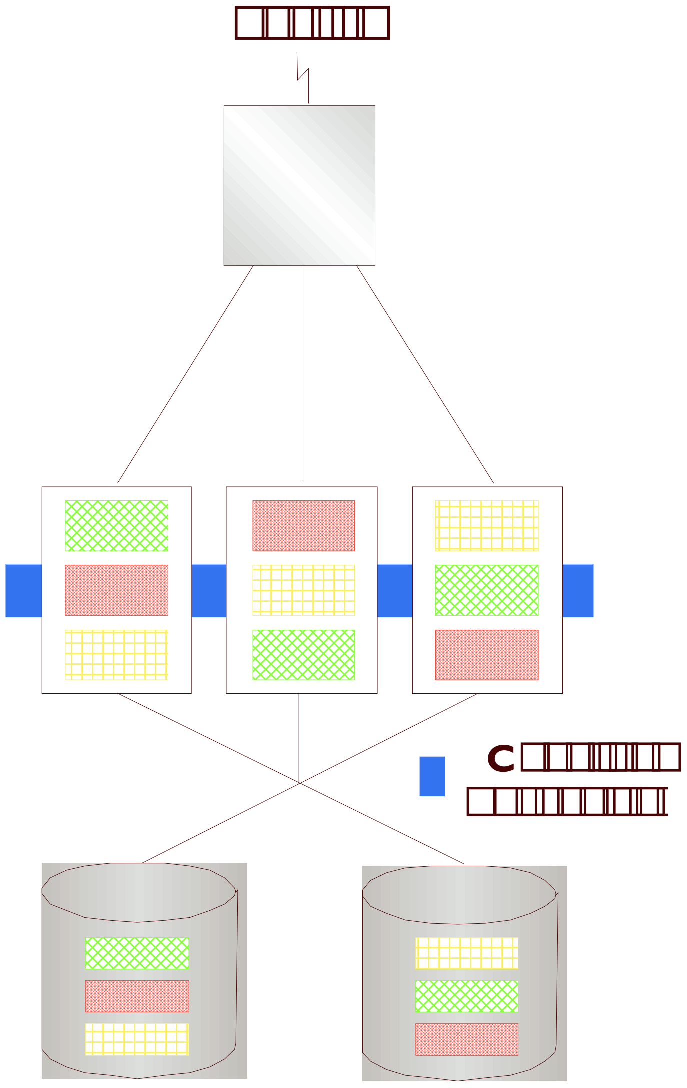


9672 – G1

- **9672 – G1 (Parallel Transaction Server)**
- **Compared to 3090-400E**
  - More capacity
  - 98% less energy
  - 93% less floor space
  - 84% less to maintain ($15,700/month)

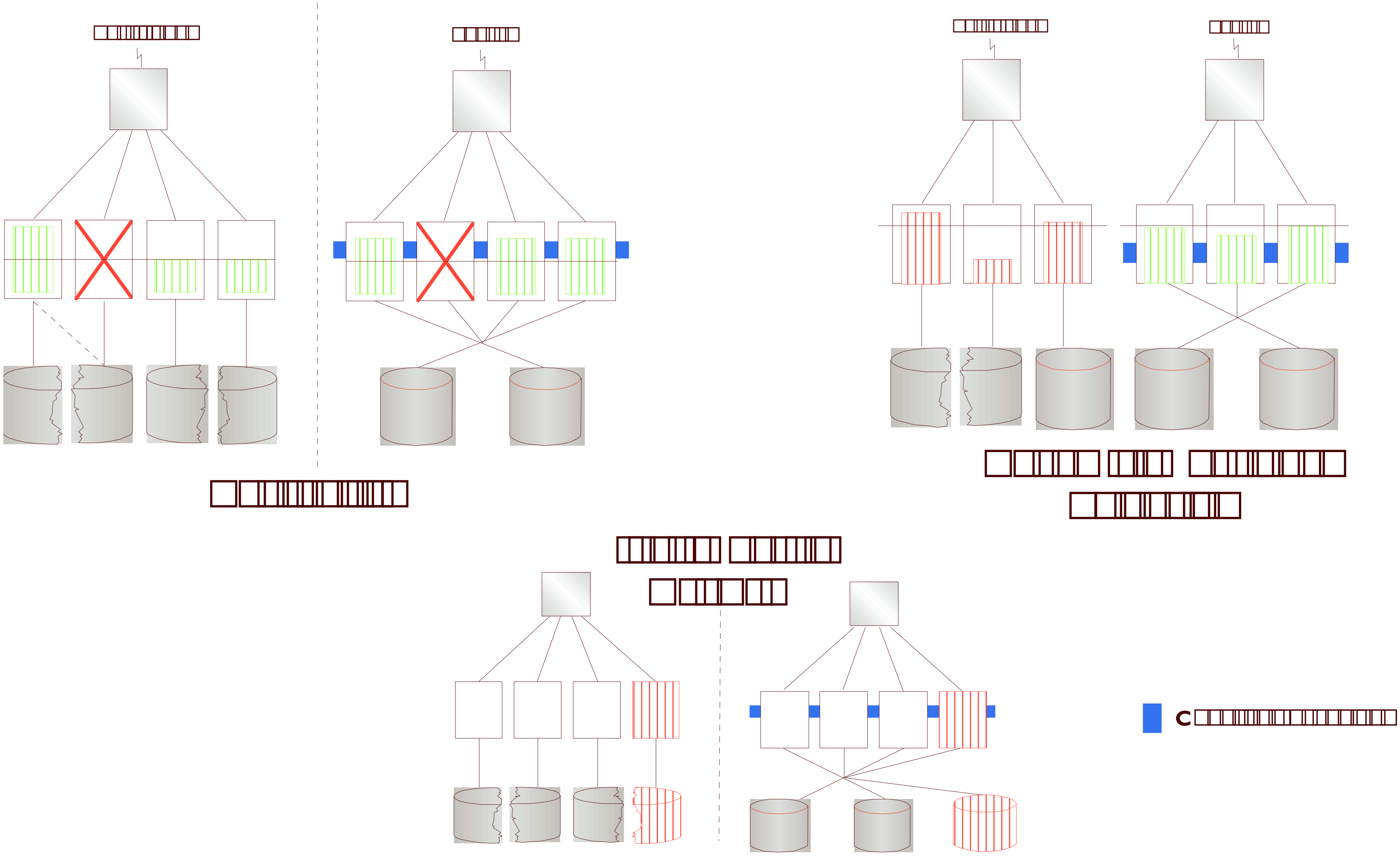
Parallel Sysplex Advantages

- **Availability (End user to Data and back)**
  - Goal: No Single Points of Failure (SPOF)
  - Planned or Unplanned outages
- **Capacity**
  - Span across multiple servers for very large applications
- **Single-system image**
  - Systems and operations management
  - End users
  - Application developers
- **Automatic, dynamic workload balancing**
- **Near linear scalability**
- **Resource Sharing**
  - Performance, System Management, Reduction of hardware resources

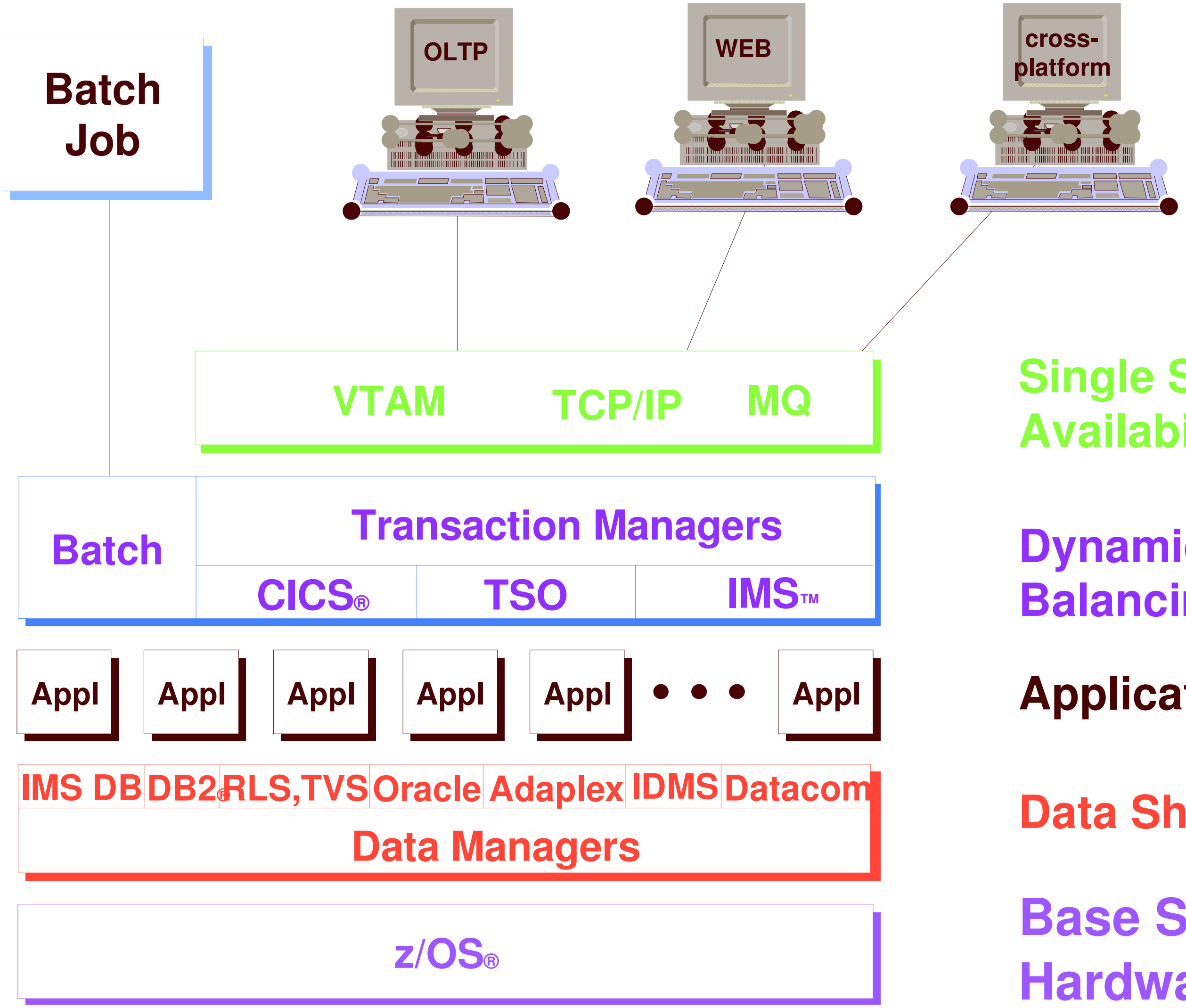


www.ibm.com/systems/z/pso

Benefits



Parallel Sysplex Software Structure



Batch Job

VTAM TCP/IP MQ

Batch

Transaction Managers

CICS® TSO IMS™

Appl Appl Appl Appl Appl • • • Appl

IMS DB DB2® RLS,TVS Oracle Adaplex IDMS Datacom

Data Managers

z/OS®

Single System Image & High Availability Connections

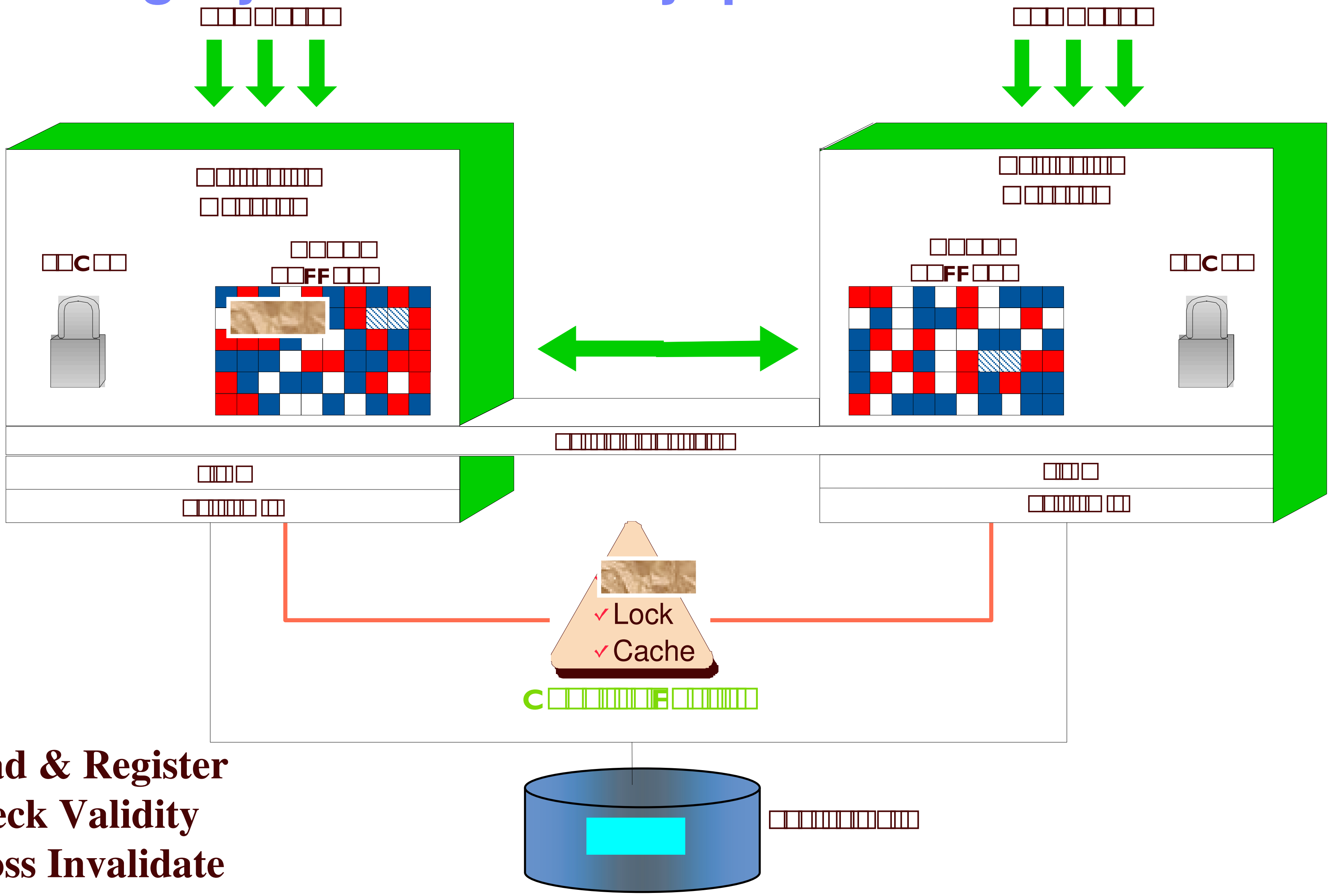
Dynamic Workload Balancing

Applications Unchanged

Data Sharing

Base Services
Hardware Interfaces

Data Integrity in a Parallel Sysplex Cluster



- Read & Register
- Check Validity
- Cross Invalidate

What is a CICS Affinity

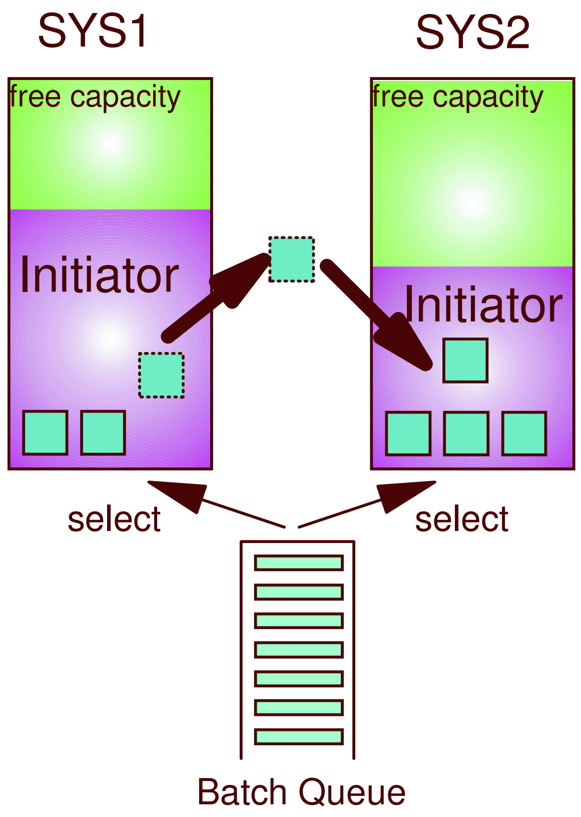
- **Two or more CICS transactions exchange data**
  - Transaction ends, leaving state data for subsequent transaction

- **Global**
  - All transactions in a group must execute in same AOR
- **LU NAME**
  - All instances in a group from same terminal must execute in same AOR
- **User ID**
  - All transactions in a group from same USERID must execute in same AOR

- **Safe, Unsafe, Suspect coding techniques**

Batch Workload Balancing
z/OS 1.4

- **Performance**
- **“Move” initiators to images with capacity**
  - Reduce number on constrained systems
  - Starting new ones on less constrained systems
  - Recheck every 10 sec.



Batch Workload Balancing

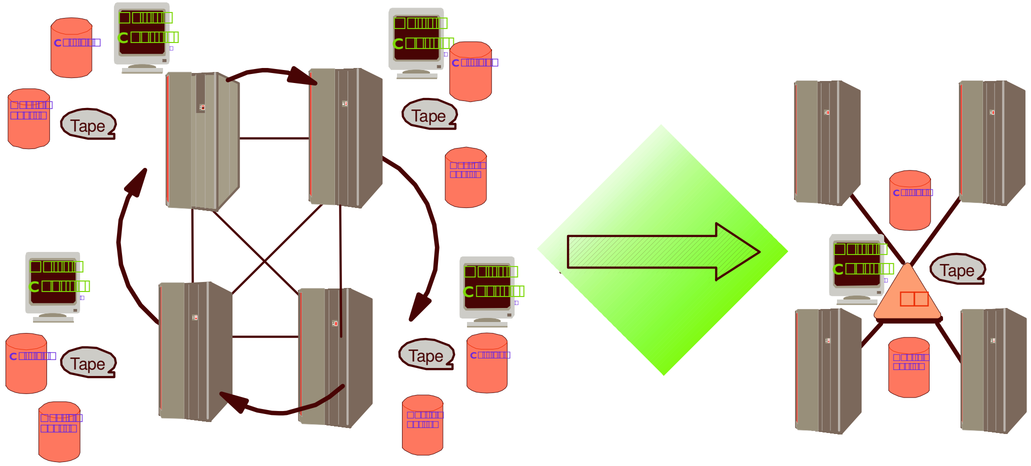
Resource Sharing

MQ Series

Operlog / Log Rec

XCF Star

GRS Star



Tape Switching

JES2 Checkpoint

- ✓ *System Management*
- ✓ *Performance*
- ✓ *Reduced HW requirements*

Shared Catalog

RACF - Security Server

IRD

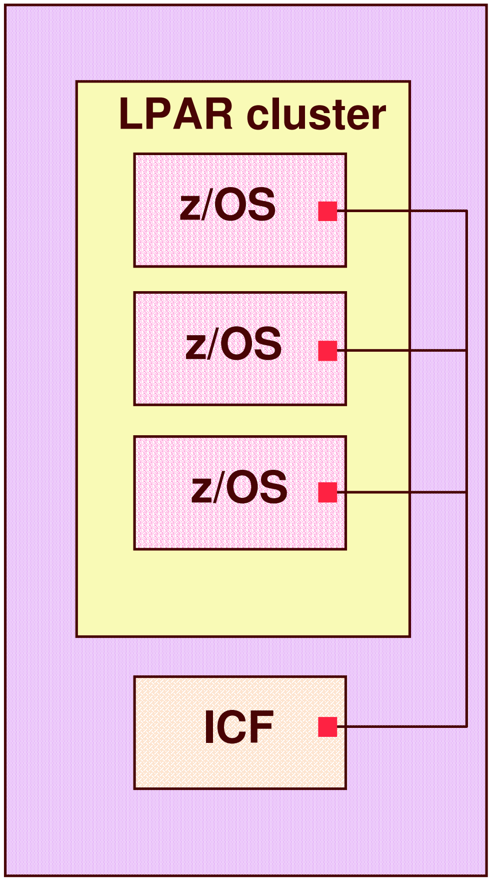
HFS / zFS

DFSMShsm

Intelligent Resource Director

- **Policy managed resources in a single CEC**
  - Processors and I/O
- **Integration of**
  - Parallel Sysplex
  - PR/SM™
  - Workload Manager
- **Directs physical resources to logical workload**
- **Handle unpredictable workloads**
- **Increase resource efficiencies**

zSeries IRD scope



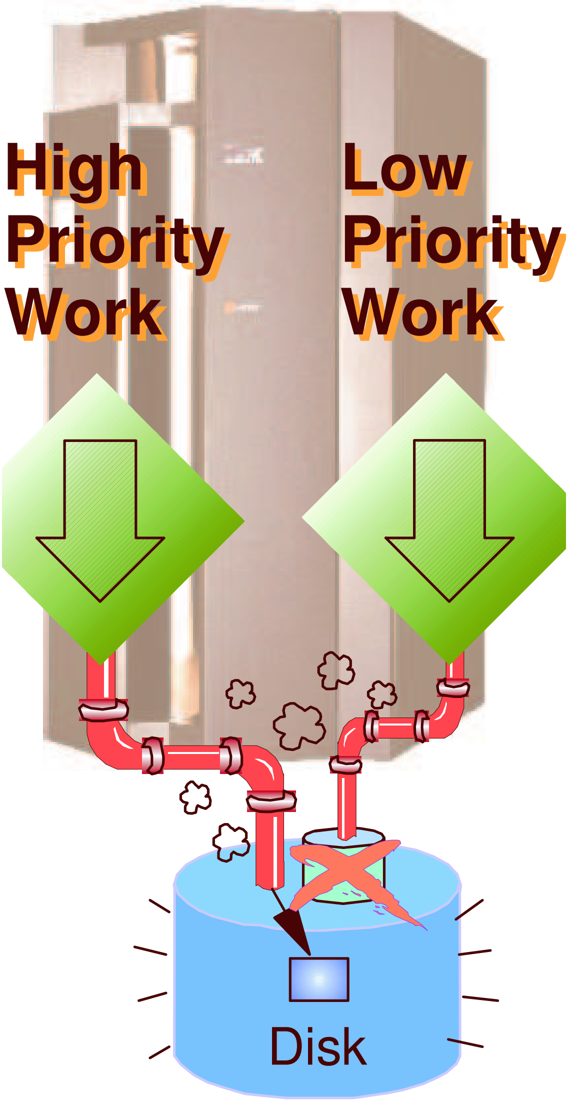
Intelligent Resource Director
LPAR CPU Management

- **Description**
  - ➢ LPAR Weight Management
  - ➢ Vary Logical CPU Management
- **Benefits**
  - ➢ Manages CPU resources across LPARs in accordance with workload goals.
  - ➢ Prevent or mitigate possible capacity problems
  - ➢ Balances multiprocessing level with processing speed for each workload
  - ➢ Helps Reduce LPAR overhead

- **Can manage Linux (native and under z/VM)**

Intelligent Resource Director
Channel Subsystem Priority Queuing

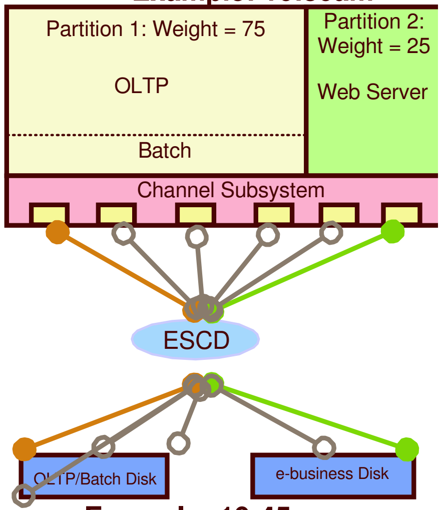
- **Description**
  - Prioritizes I/O within an LPAR cluster
  - Basic I/O Priority Queuing works within LPAR
- **Benefits**
  - Allows better channel resource management with MIF
    - High priority work is given preferential access to the channel
    - Can reduce channel requirements



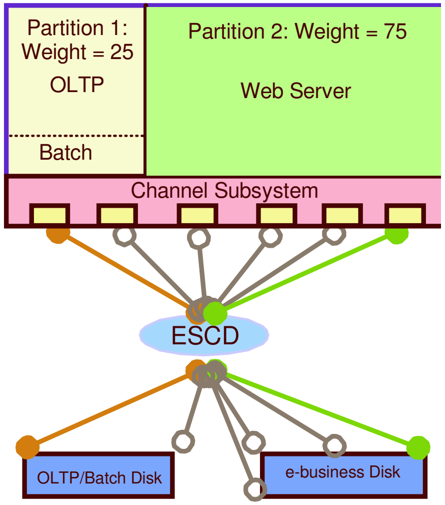
Intelligent Resource Director
Dynamic Channel Path Management

- **Description**
  - Dynamically manage channel paths
  - Moves bandwidth to subsystem(s) based on workload requirements
  - Optimized with Channel Subsystem Priority Queuing
- **Benefits**
  - More efficient use of hardware resource
  - Reduces channel requirements
  - Simplifies I/O configuration planning and definition
  - Dynamically balances I/O connectivity based on workload demand

Example: 10:30am



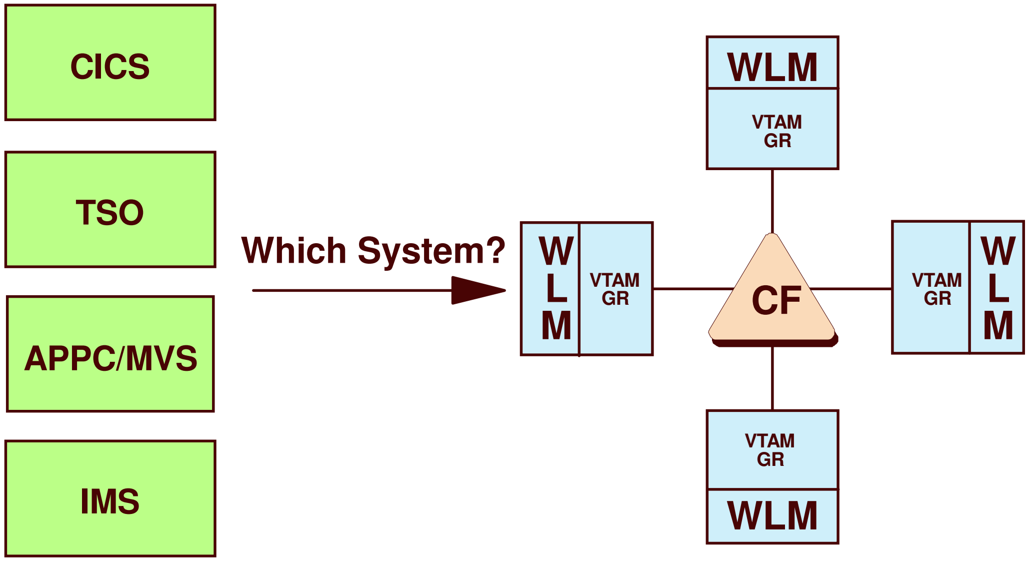
Example: 10:45am



SNA Support

- **VTAM Generic Resources**
  - Based on CPU capacity

- **Multi-Node Persistent Sessions**
  - Avoids reestablishing VTAM connection
  - Optionally track CICS or IMS sessions

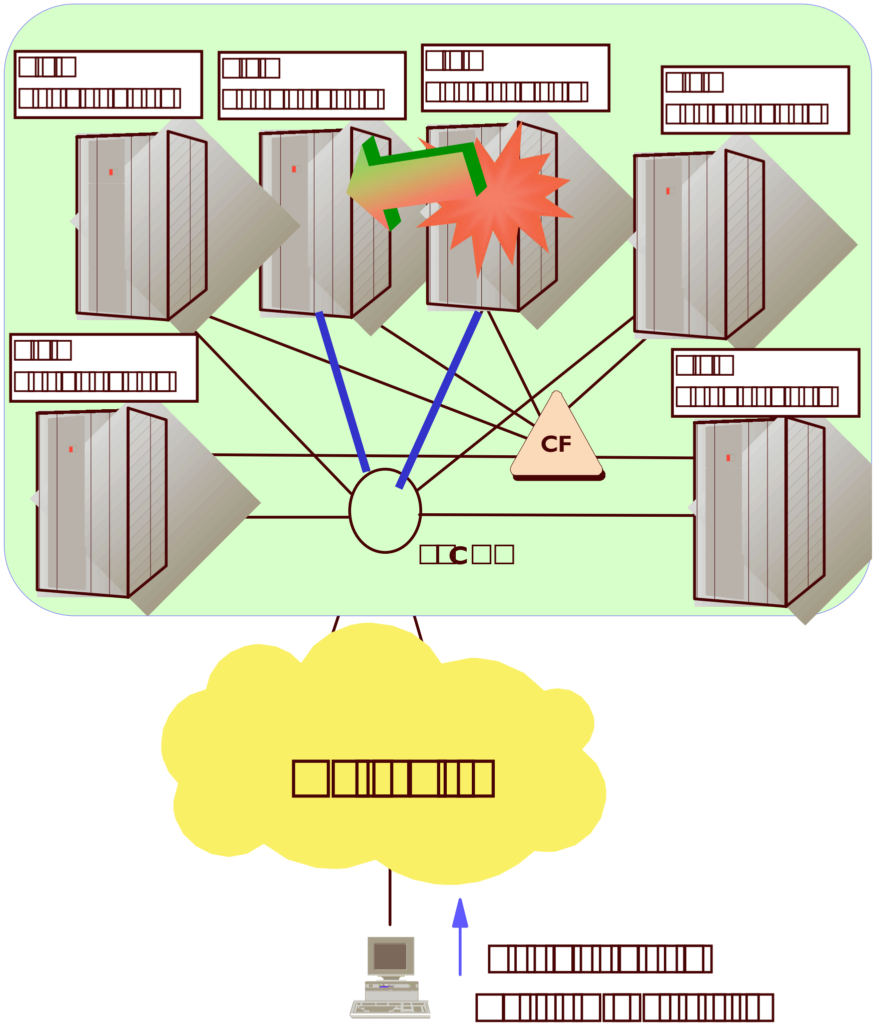


TCP/IP Workload Balancing

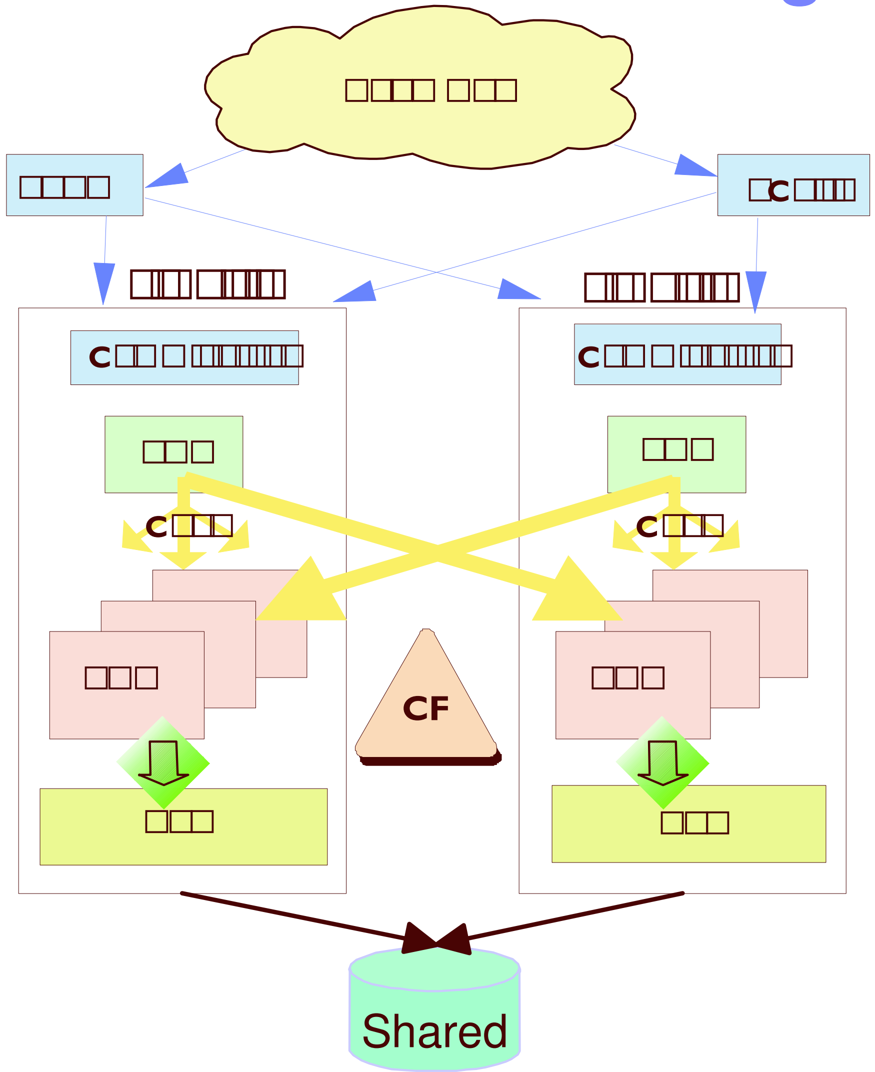
- **Spraying**
  - “Dumb” round robin
- **DNS/WLM**
  - Domain Name Server (URL) is resolved to an IP Address
  - WLM consulted, and request routed to best host to balance workload
- **Network Distributor**
  - External box. Requires connectivity to each host
  - Routes based upon WLM, user, application, QoS, etc.
  - Similar to Cisco Multi-Node Load Balancer
- **Sysplex Distributor**
  - No external box required. Connects to a node within Sysplex,
  - Routes to host based upon WLM, user, application, QoS, etc.
  - Removes SPOF of external box
  - Removes complexities of multiple LPARs in a CEC w/ OSA

Dynamic VIPA / VIPA Takeover

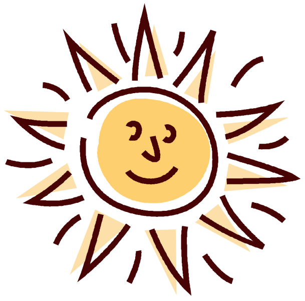
- **Single System Image to IP Network**
- **VIPA Takeover**
  - Stack may be moved to another host automatically
  - No configuration changes to routers
  - Coordinated with application dependencies
- **VIPA Takeback**
  - Non-disruptive movement of stack to another host
  - Prior to planned outage
  - After original host back online



CICS/DB2 Data Sharing Example



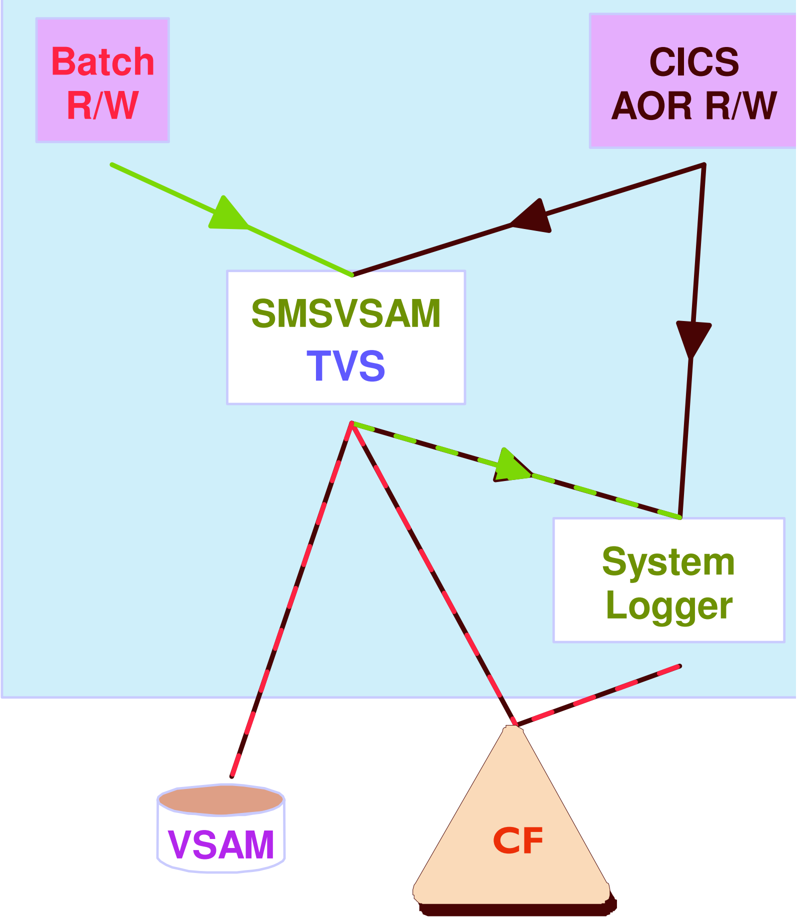
- Removes SPOF of:
  - Server
  - LPAR
  - Subsystems
- Planned and Unplanned Outages
- Single System Image
- Dynamic Session Balancing
- Dynamic Transaction Routing

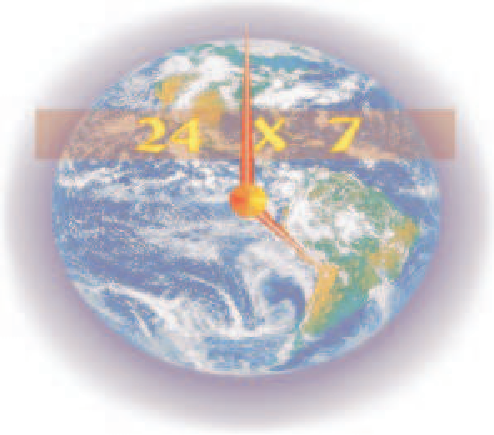


Improved Application Availa□ility

Transactional VSAM

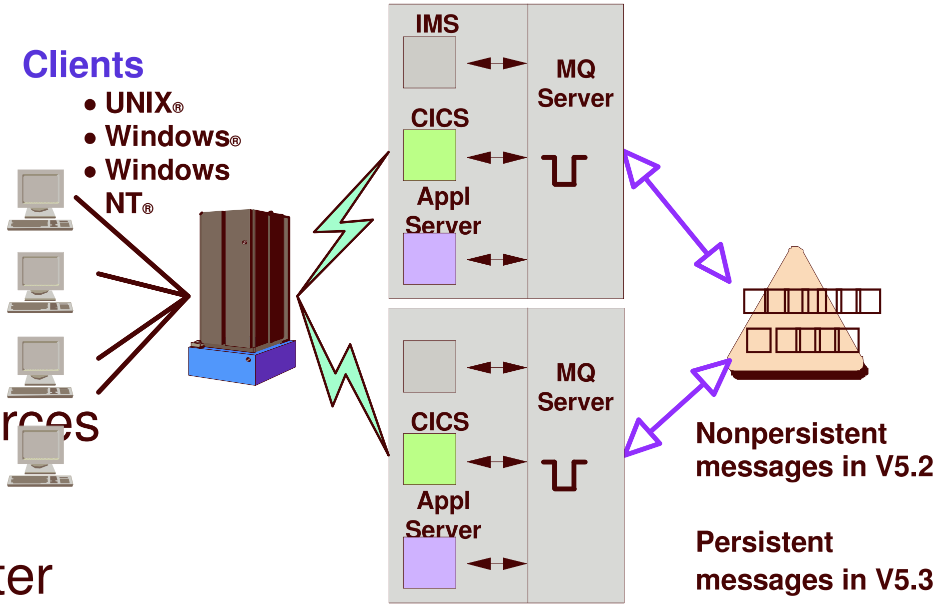
- Addresses the batch window for CICS
- Batch updates concurrent with CICS on-line
- Multiple concurrent batch updates against same files
- Enables 24 x 7 availability





WebSphere MQ for z/OS

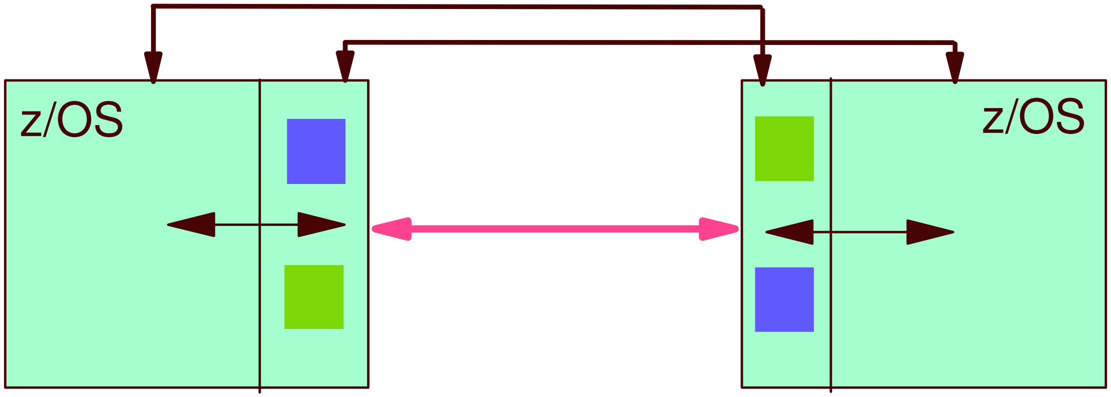
- Availability
  - Workload Balancing
  - Planned maintenance easier
- Administration
  - Simple, scalable administration
  - Single name space to describe resources
  - Fewer resources to define
  - Single system to control and administer
- ARM Support
- System-Managed CF Structure Duplexing Support

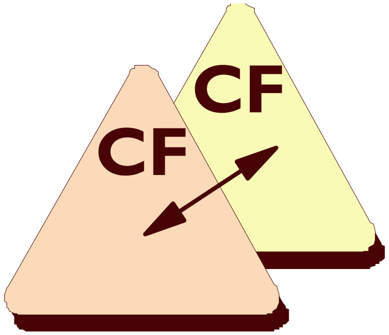


Nonpersistent messages in V5.2

Persistent messages in V5.3

System-Managed Coupling Facility (CF) Structure Duplexing



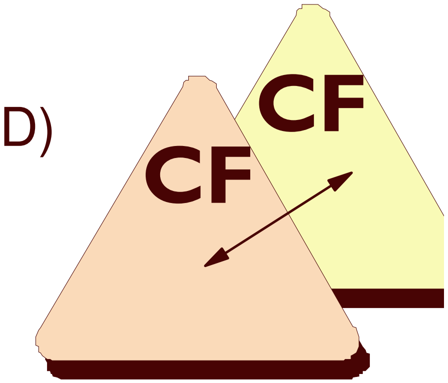


- **Can Improve availability by providing:**
  - Enables "all-ICF" configuration
  - Basic recovery for structures
  - Consistent recovery mechanism - Reduced complexity
  - Faster than structure rebuild
- **Technical paper** (zsw01975usen) **available at ibm.com**/server/eserver/zSeries/pso

Robust failure recovery capability

System-Managed Coupling Facility (CF) Structure Duplexing Exploiters

CICS	Shared TS, CF data tables, named counter
CommServer (TCP/IP)	G/R, MNPS (VTAM), SWSA, Sysplex Ports
DB2	SCA
DFSMS	RLS Lock (VSAM), Common Recall Queue
IMS	CQS, EMH, VSO structures
IRLM	Lock (DB2 and IMS)
JES2	Checkpoint
MQ	Shared queues
WLM	Shared enclaves, LPAR Clusters (IRD)
BatchPipes	
System Logger	



Coupling Facility (CF) Level of Support

CF Level	Function	G3, G4	G5/G6	Z800	z900	z890 z990
1	Dynamic Alter support CICS temporary storage queues System logger	X X X	X X X	X X X	X X X	X X X
2	DB2 performance VSAM RLS 255 Connectors / 1023 structures for IMS Batch DL1	X X X	X X X	X X X	X X X	X X X
3	IMS shared message queue base	X	X	X	X	X
4	Performance optimization for IMS & VSAM RLS Dynamic CF Dispatching Internal Coupling Facility IMS shared message queue extensions	X X X X X	X X X X X	X X X X X	X X X X X	X X X X X
5	DB2 cache structure duplexing DB2 castout performance improvement Dynamic ICF expansion into shared CP pool	X X X	X X X	X X X	X X X	X X X
6	ICB & IC TPF support	X X	X X	X X	X X	X X
7	Shared ICF partitions on server models DB2 Delete Name optimization	X X	X X	X X	X X	X X

Detailed information regarding CF levels can be found in “Coupling Facility Level (CFLevel) Considerations at URL: **ibm.com**/servers/eserver/zseries/pso/cftable.html

Coupling Facility (CF) Level of Support

CF Level	Function	G3, G4	G5/G6	z800	z900	z890 z990
8	Systems-Managed Rebuild Dynamic ICF Expansion into shared ICF pool	X	X X	X X	X X	X X
9	MQSeries Shared Queues WLM Multi-System Enclaves Intelligent Resource Director IC3 / ISC3 / ICB3 peer mode		X X	X X X X	X X X X	X X X X
10	z900 GA2 Level				X	
11	SM Duplexing support for 9672 G5/G6/R06		X			
12	64-bit CFCC addressability Message Time Ordering SM Duplexing support for zSeries CFs			X X	X X X	X X X
13	DB2 Castout Performance			X	X	X
14	CFCC Dispatcher Enhancements					X

Detailed information regarding CF levels can be found in "Coupling Facility Level (CFLevel) Considerations at URL: **ibm.com**/servers/eserver/zseries/pso/cftable.html

Configuring CF Links

Server	IC	ICB-4	ICB-3	ICB	ISC-3	Max # Links
z800	32	-	5 6 (0CF)	-	24	26 + 32
z900-100 CF	32	-	16	16	32 42 w/ RPQ	64
z900	32	-	16	8 16 w/ RPQ	32	64
z890	32	8	16	-	48	64
z990	32	16	16	8	48	64
z9	32	16	16	-	48 Peer Mode Only	64

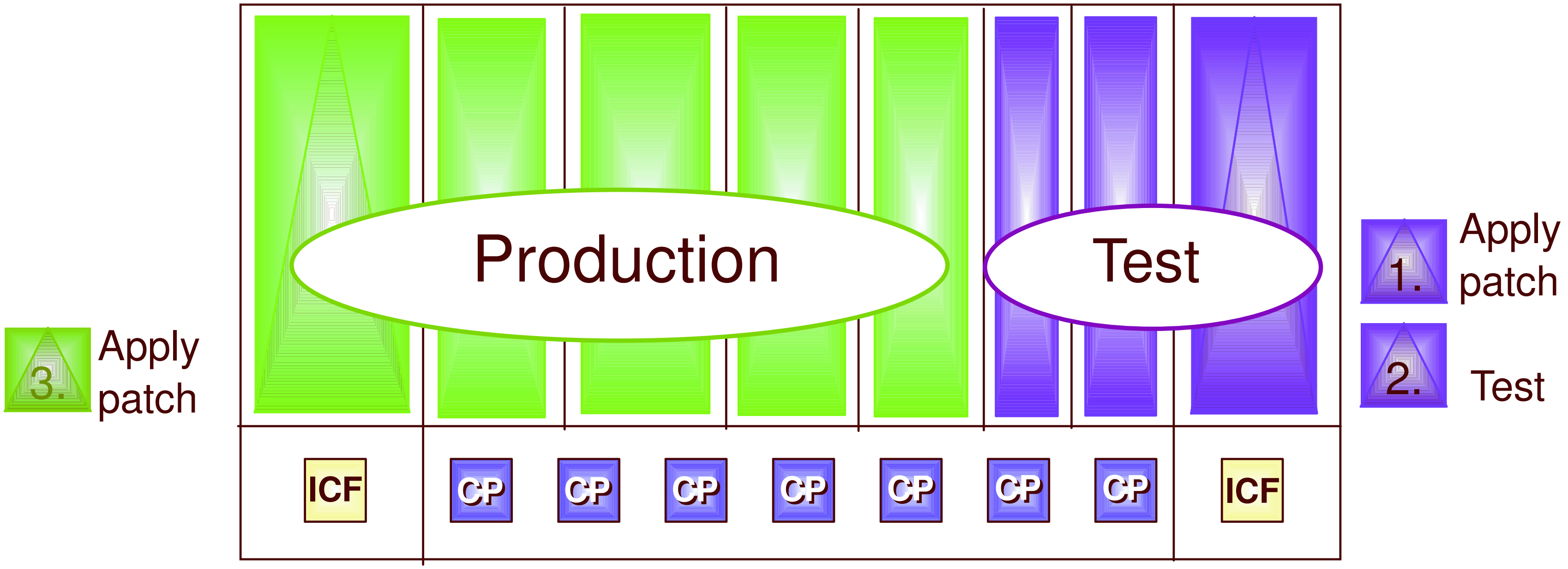
zSeries CF Link Speeds

Model	IC	ICB-4	ICB-3	ICB	ISC-3	ISC
9672 G5/G6	700 MB/sec	-	-	250 MB/sec	-	100 MB/sec
z800	1125 MB/Sec	-	500 MB/sec	-	✓200 MB/sec ✓100 MB/Sec beyond 10km ✓100 MB/Sec Compat Mode	n/a
z890	MB/sec	1500 MB/sec	500 MB/sec	-	Same as z800	n/a
z900	1400 MB/sec	-	500 MB/sec	250 MB/sec	Same as z800	n/a
z990	3500 MB/sec	1500 MB/sec	500 MB/sec	250 MB/sec	Same as z800	n/a

- **Peer mode supports**
  - Improved throughput, increasing coupling efficiency and improving response times
  - Merging of Sender and Receiver links, reducing number of links required
  - Increase from 2 to 7 subchannels per buffer sets, reducing number of links required
  - Larger data buffers and improved protocols improving long distance performance
  - zSeries connected to 9672s must use compatibility mode

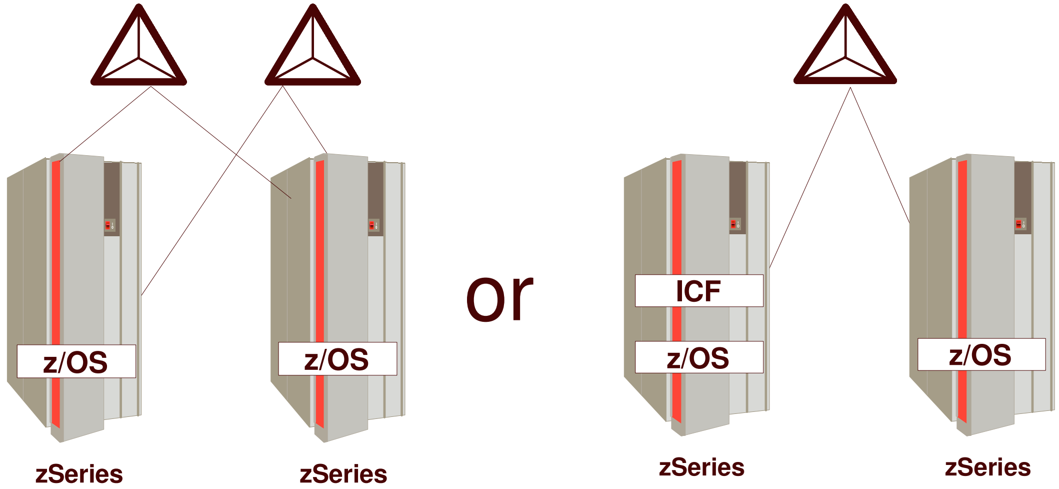
Non-disruptive CFCC Patch Apply

z890, z990

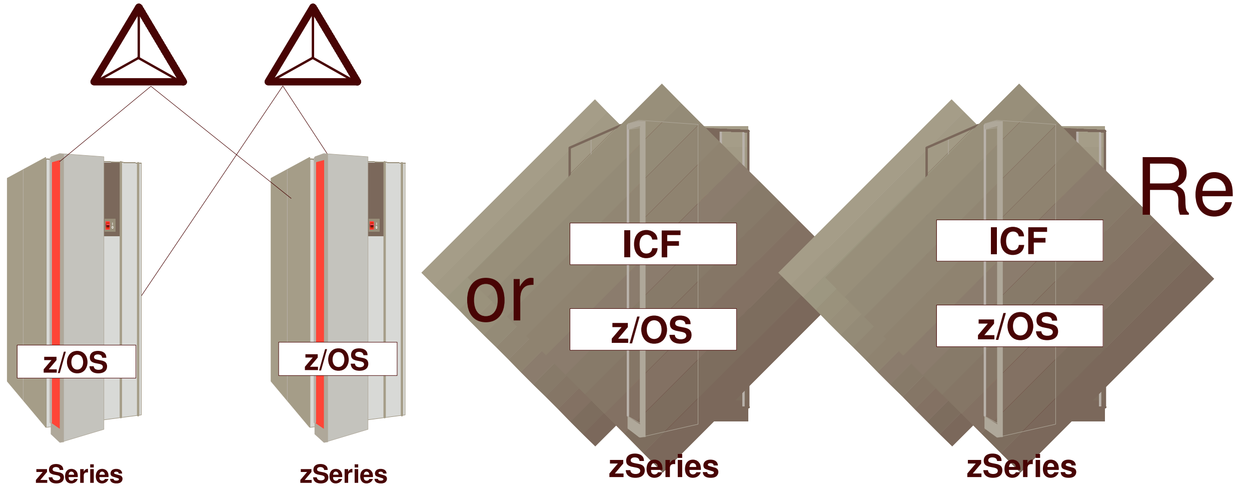


- **Removes disruption to entire CEC for previously disruptive CFCC patches**
  - Disruption occurs one CFCC LPAR at a time
  - Allows rolling CFCC maintenance across CF LPARs
  - Similar to rolling z/OS maintenance across OS images
  - Reduces requirement to isolate test CFs from production OS and CF images
  - CFLevel upgrades will still be disruptive to the entire box

Traditional CF Configuration Recommendations



Data Sharing



Resource Sharing

Which CF Configuration is Right for Me?

- **It Depends !**
- **On the factors that are most important to your business**
  - cost
  - availability
  - system management
- **Much less on the technical factors associated with your Parallel Sysplex implementation**
  - link technology
  - max. size of sysplex
  - etc.

Performance

- **“Typical” Observed Performance (all IBM HW)**
  - Multisystem Management - 3%
  - Resource Sharing - 3%
  - Application data sharing - <10%
  - Incremental cost of adding an image – 0.5%

Server Time Protocol –
Time synchronization for the next generation

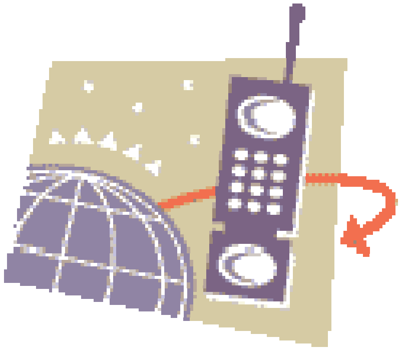
What is Server Time Protocol (STP)?

- Time synchronization using a Coordinated Timing Network (CTN)
  - Similar to Network Time Protocol standard
- Uses CF links
- IBM System z9 EC, z9 BC, IBM eServer™ zSeries® 990 and 890 (z990, z890)

- **Benefits**
  - Improved time synchronization
  - Can scale with distance
  - Supports up to 100 km
  - Potentially reduces the cross-site connectivity
  - Concurrent migration from ETR network
  - Coexistence with ETR network

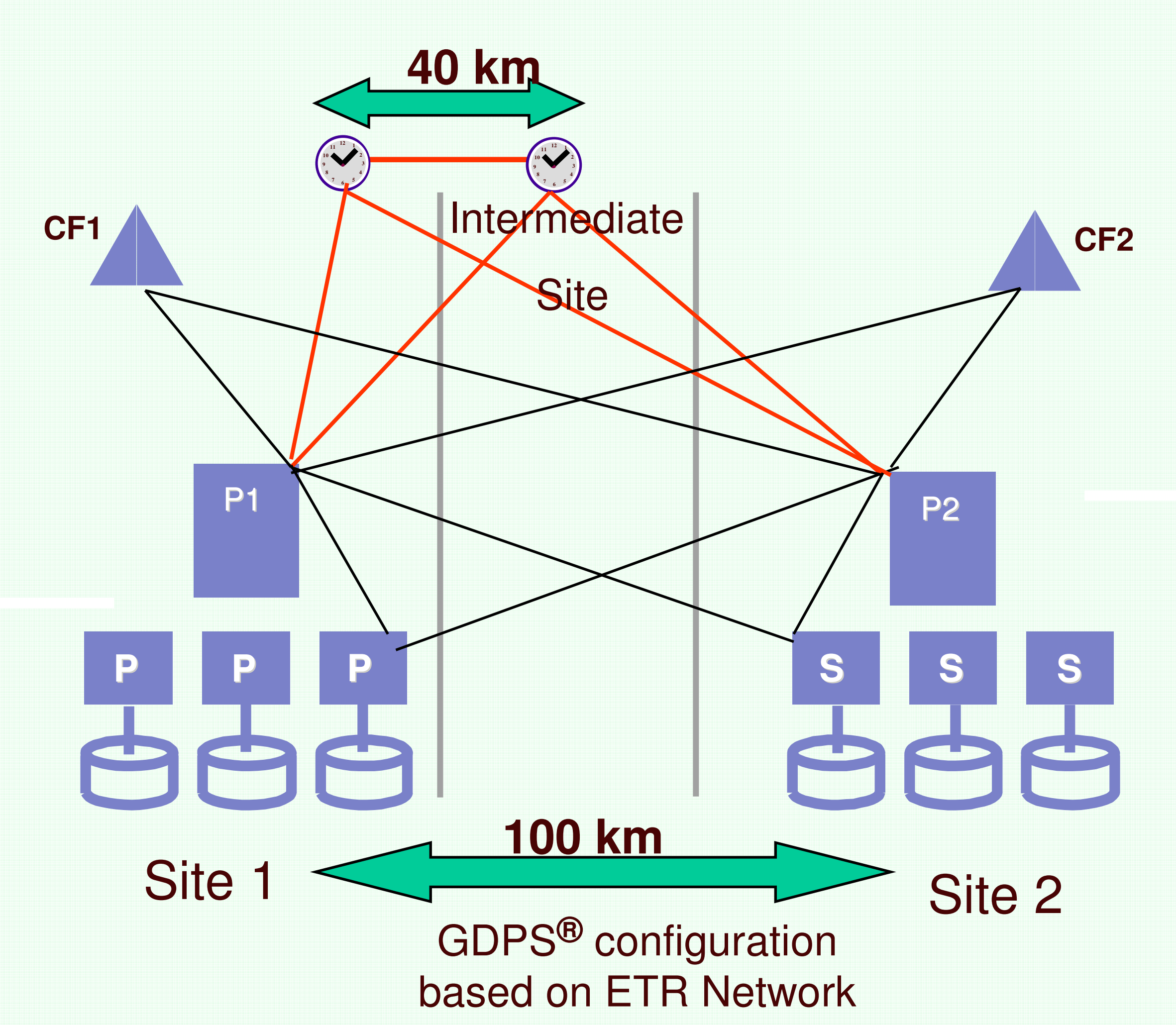
Key Attributes

- Allows
  - Use of dial-out time services to within +/- 100 ms of UTC
    - NIST Automated Computer Time Service (ACTS)
    - NRC Canadian Time Service (CTS)
    - IEN Telephone Date Code (CTD)
  - Scheduling of dial-outs so that CST can be steered to UTC
  - Setting of local time parameters
    - Time zone offset
    - Daylight Saving Time offset with automatic update
    - Leap Seconds offset
  - Adjustment of CST up to +/- 60 seconds



STP Enhancements over ETR Network

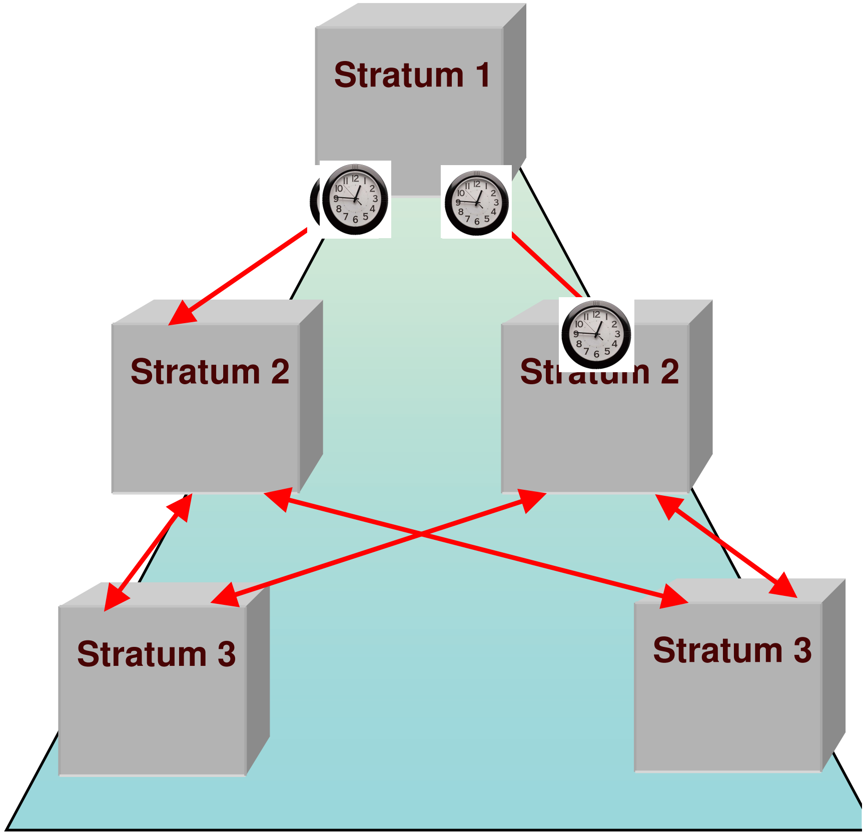
- STP supports a multi-site timing network of up to 100 km without requiring an intermediate site
- Fiber distance between Sysplex Timers cannot exceed 40 km
  - **Intermediate site to locate second timer recommended to avoid a single point of failure, if data centers more than 40 km apart**



GDPS® configuration based on ETR Network

Terminology

- STP transmits timekeeping information in layers or Stratums
- Stratum 1 (S1)
  - ➢ Highest level in the timing network
- Stratum 2 (S2)
  - ➢ Server/Coupling Facility (CF) synchronizing to Stratum 1
- Stratum 3 (S3)
  - ➢ Server/Coupling Facility (CF) synchronizing to Stratum 2
- STP supports configurations up to S3



Time message will find a new path if needed