Parallel Sysplex Basics

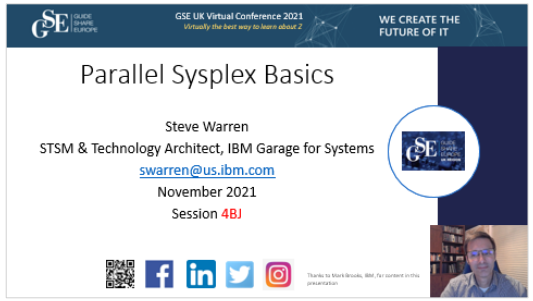
Steve Warren
STSM & Technology Architect, IBM Garage for Systems
swarren@us.ibm.com

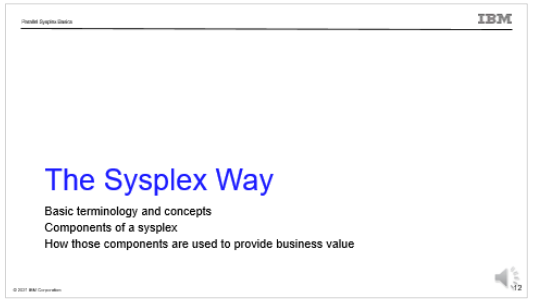
November 2021

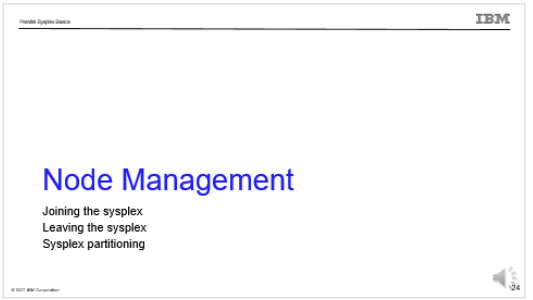
Session 4BJ

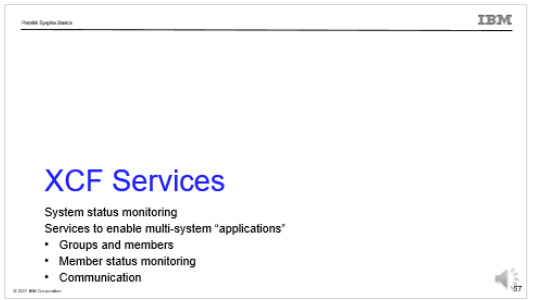
Thanks to Mark Brooks, IBM, for content in this presentation

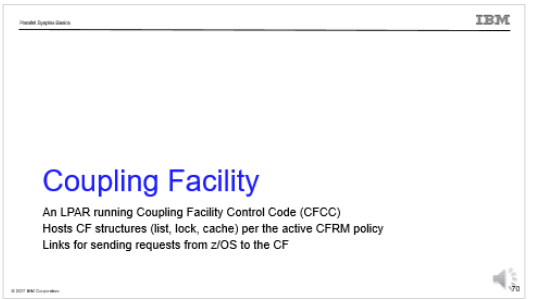
Parallel Sysplex Basics Agenda

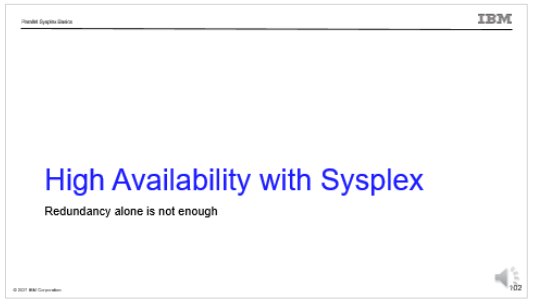












Introduction and positioning

Why sysplex?

Availability

Capacity and growth

What on earth is a sysplex?

- A group of coupled yet distinct clustered instances of z/OS images that cooperate to provide a virtual single z/OS instance view to applications with the goal for high availability and capacity by providing:
  - Messaging between the various images in the sysplex
  - Data sharing between all the images

Why sysplex? Availability of business services!

Planned Outages (90%)

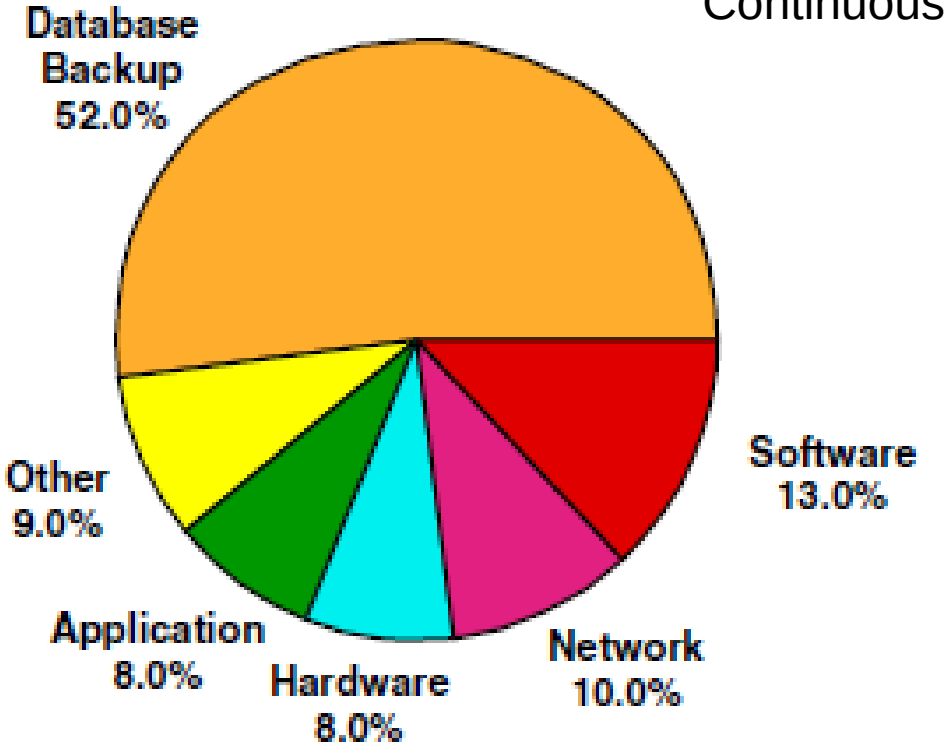
Continuous Operations (CO)

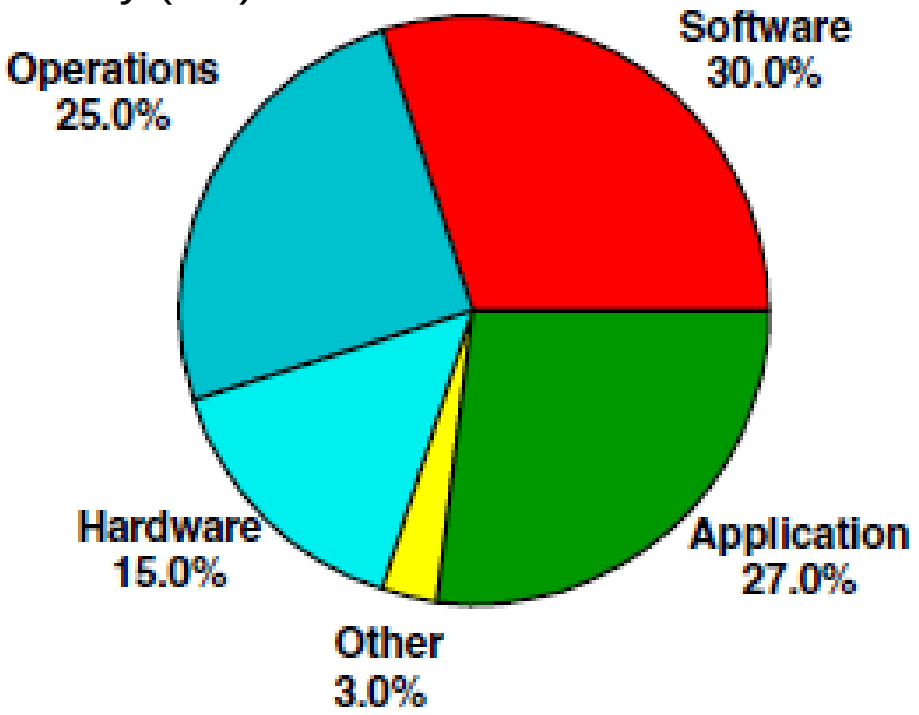
Unplanned Outages (10%)

High Availability (HA)

Parallel Sysplex addresses both

Continuous Availability (CA)

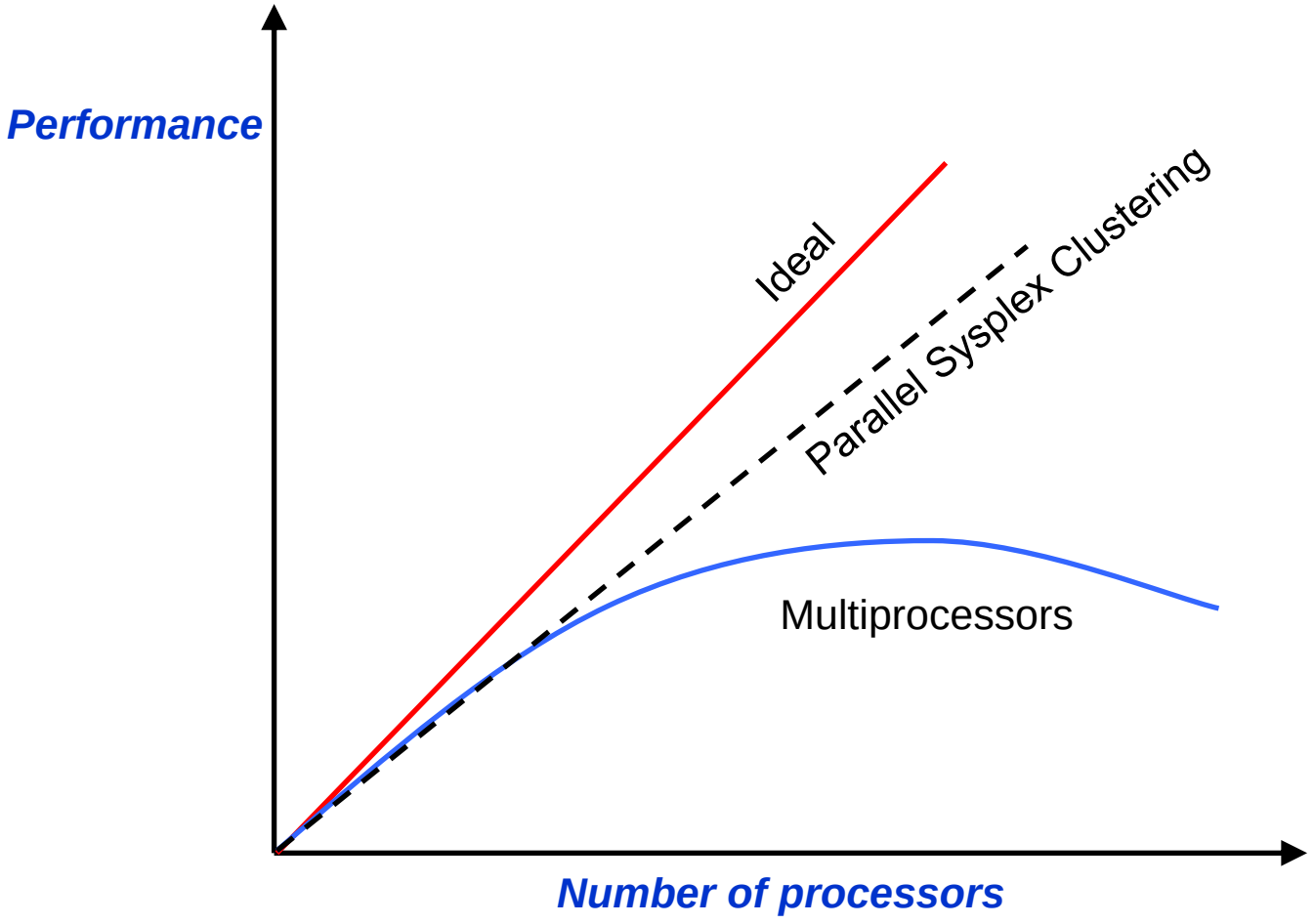




Companies face increasing pressure to deliver business services 24x7.

A well managed, properly configured parallel sysplex can provide near continuous availability.

Why sysplex? More capacity!



"*It doesn't take an extremely large number of CPUs before a single-image system will deliver less effective capacity than a sysplex configuration of two systems, each with half as many CPUs*."
-- Bob Rogers, IBMSystems Magazine

In contrast to a multiprocessor, sysplex scaling is near linear. Adding another system to the sysplex may give you more effective capacity than adding another CP to an existing system.

Installations effectively exploiting sysplex today for availability will be well positioned to exploit sysplex to increase capacity.

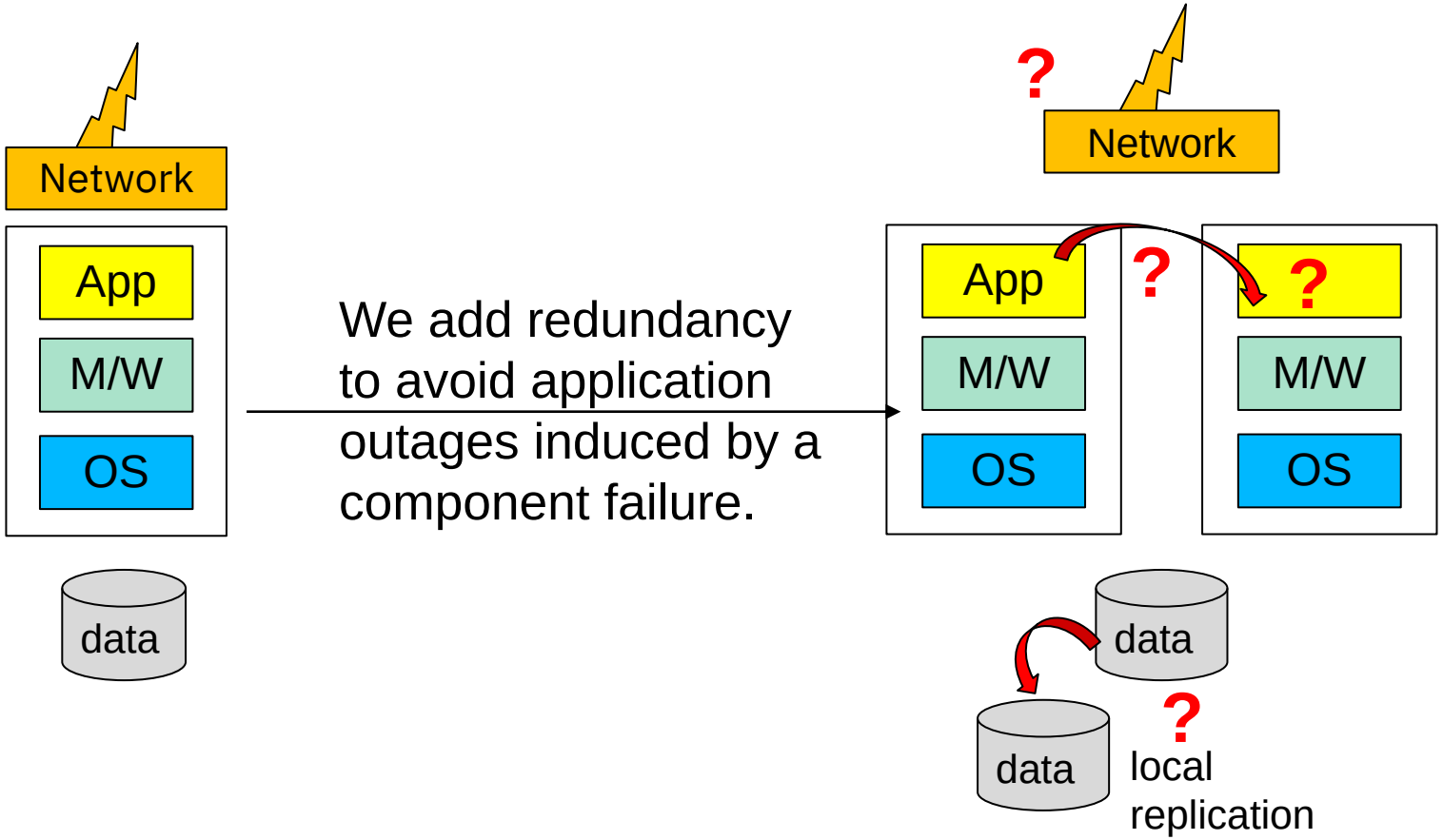
The Sysplex Way

Basic terminology and concepts

Components of a sysplex

How those components are used to provide business value

Thinking about availability



Some issues:

- Recovery model?
- Failover? Both active?
- Failure detection?
- Coordination?
- Data replication?
- Network connections?
- Reconnect? Reroute?
- **Serialization?**
- **Scaling?**

App	Application (the business service you care about)
M/W	Middleware (Db2, CICS, MQ, WebSphere, …)
OS	Operating System (Linux, z/OS, …)

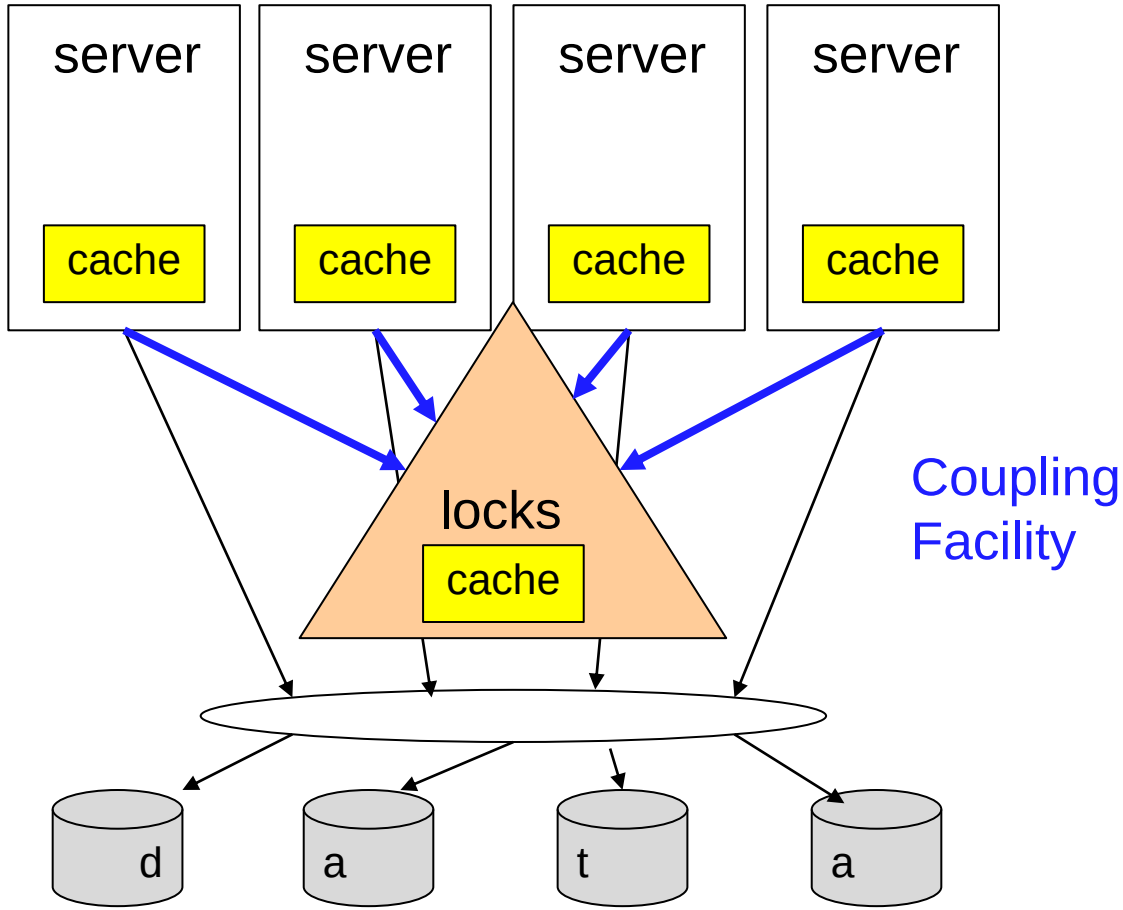
Sysplex provides the foundation for dealing with these issues.

Parallel Sysplex shared data model

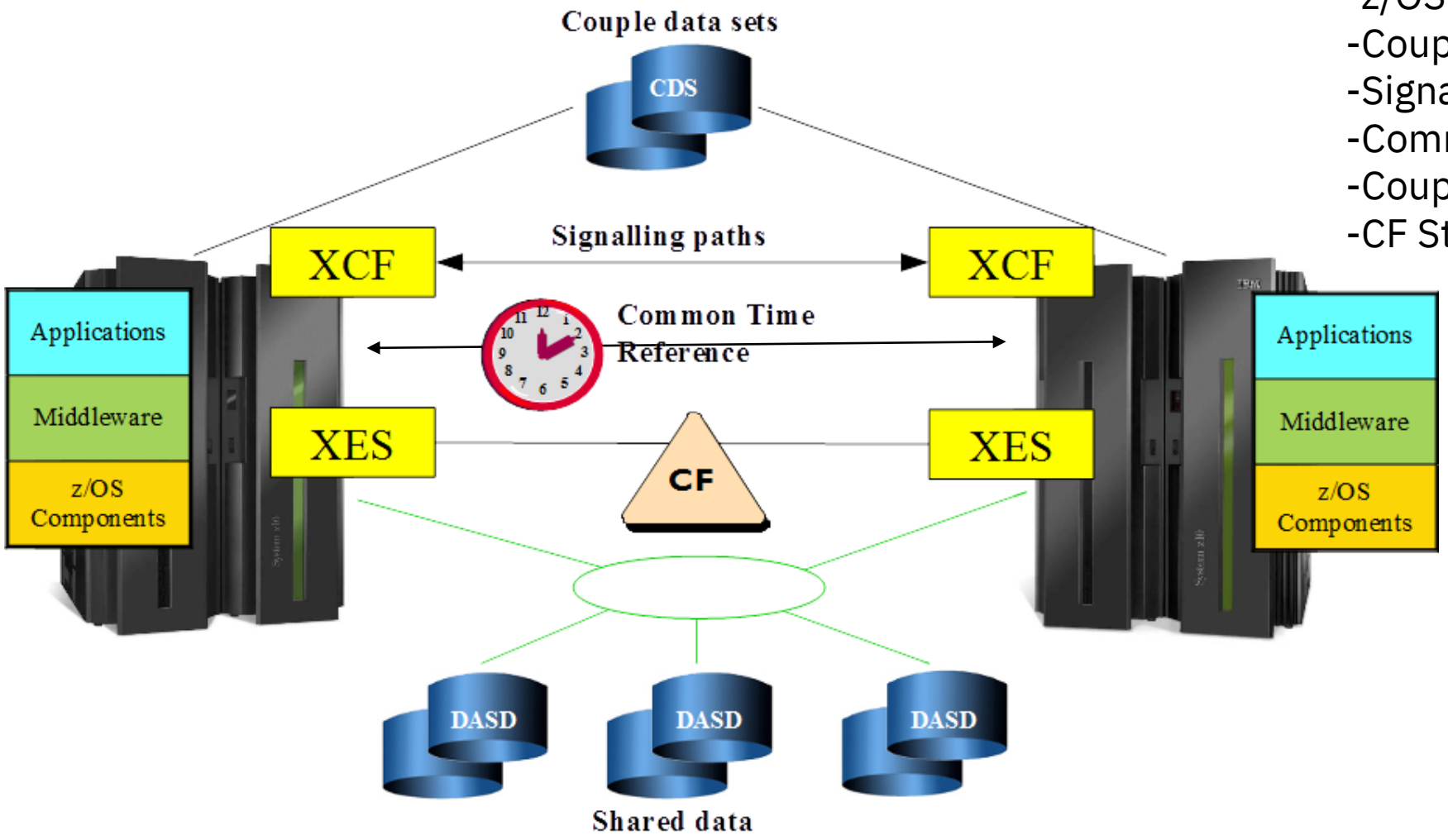
Centralized, shared data model

- Every server can access entire DB
  - But serialization is needed
  - **High speed locking with CF**
  - **High speed caching with coherency**
- Flexible workload distribution
- Tuning? Utilization? Growth?
  - **CF performance is critical**

Scales linearly!



Sysplex componentry



Terminology

-Sysplex (1 to 32 z/OS images)
-z/OS components, XCF and XES
-Couple Data Sets (CDS)
-Signal paths for message passing
-Common time reference (STP)
-Coupling Facility (CF)
-CF Structures (list, lock, cache)

Single system vs multisystem sysplex

- XCF-Local Mode
  - No couple data sets (CDS), no coupling facilities
  - Often used to do set up or to fix things (such as a parmlib member)

COUPLE=**
COUPLE=COUPLE00 (as supplied by IBM)

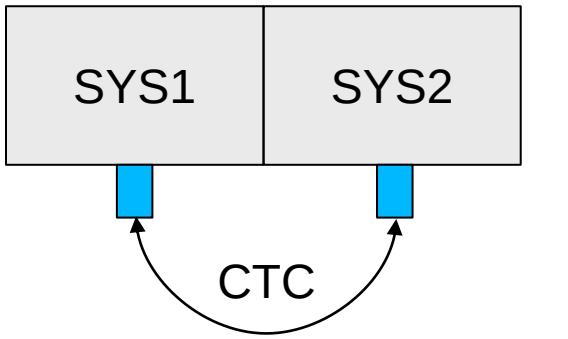
- Monoplex
  - Has couple data sets, may or may not have coupling facilities
  - Determined by PLEXCFG=MONOPLEX (IEASYSxx parmlib member)
    - First system into sysplex updates the sysplex couple data set to indicate that no other system is permitted to join the sysplex.
    - Must re-IPL the system to change
- Multisystem sysplex
  - Has couple data sets, may or may not have coupling facilities
  - A set of two or more z/OS instances with the same sysplex name that:
    - Use the same sysplex couple data sets
    - Have XCF signal connectivity with each other
    - Have the same common time reference

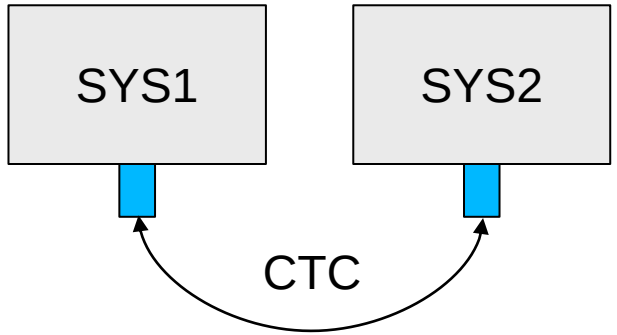
Sysplex configurations

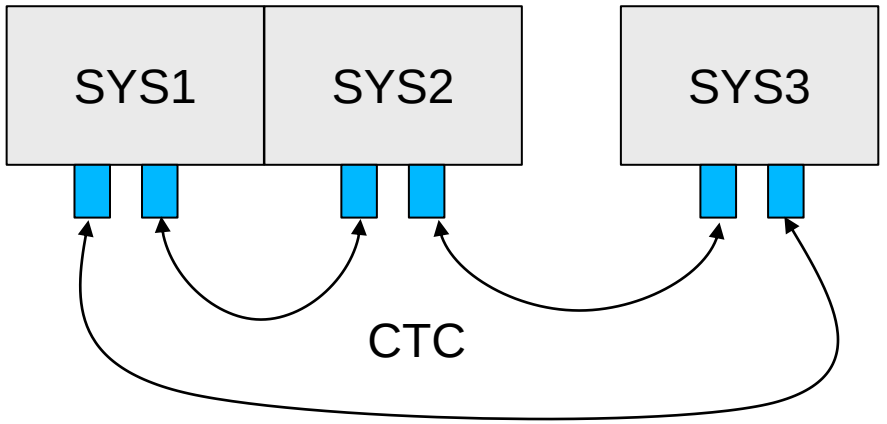
- Base Sysplex
  - Does not have a coupling facility
  - Typically used for resource sharing
- Parallel Sysplex
  - Has a coupling facility (CF)
  - Used for resource sharing, data sharing, or both

We tend to focus on parallel sysplex, but a base sysplex does provide value. They do exist.

Variety of base sysplex configurations

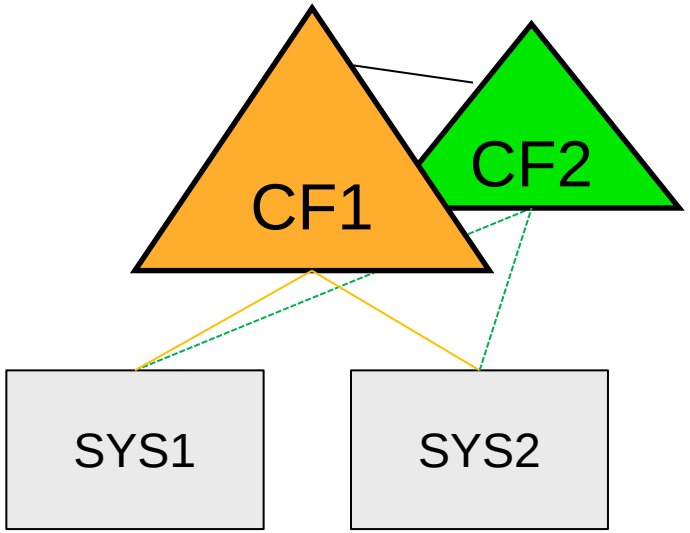




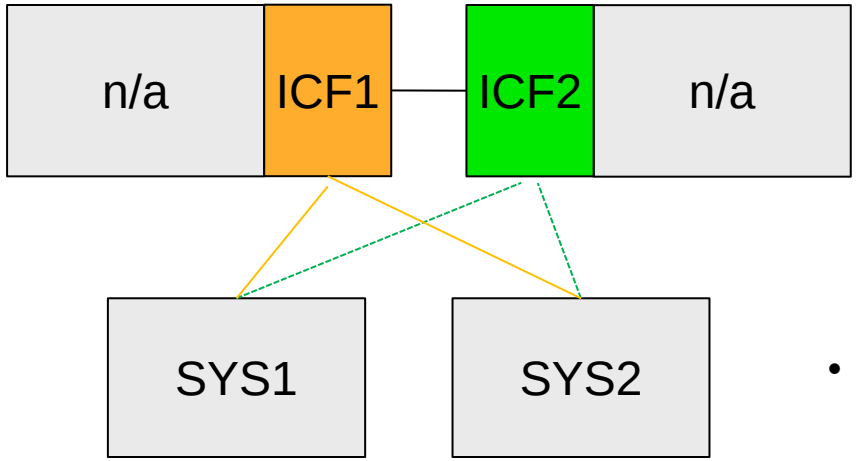


- One or more z/OS images, spread across one or more CECs.
  - CECs could be different generations (generally N-2 is supported)
  - Systems could be running different software releases
- Since no coupling facility, need CTC devices for communication.
- As number of nodes increases, CTC configuration (which is $O(n^2)$) tends to become unmanageable. GRS ENQ performance (ring mode) might also be an issue.

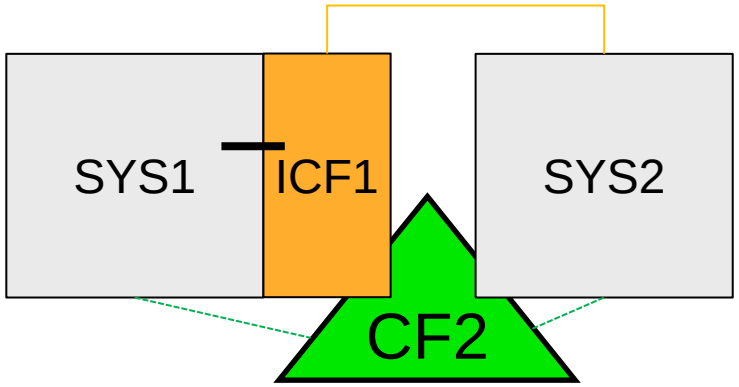
Variety of parallel sysplex configurations

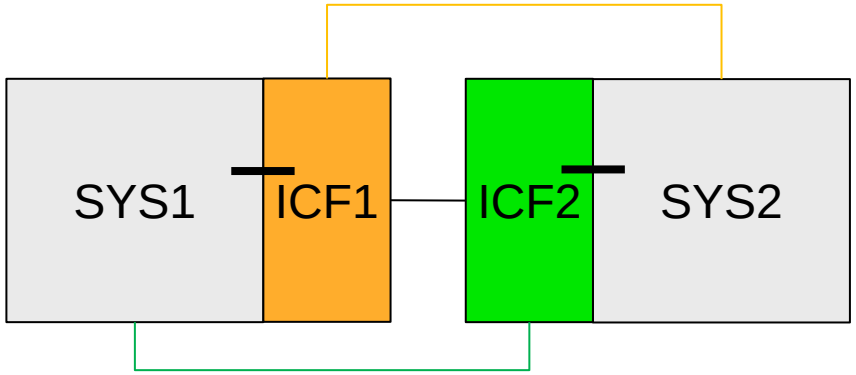


- Stand alone CF
- Internal CF
- “External to me” CF

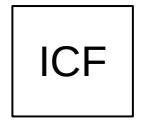


- CF placement relative to z/OS images critical for failure recovery.
- Could have CTC devices for communication as well





External CF (stand alone)



Internal CF (LPAR)

- One or more z/OS images, spread across one or more CECs.
- Use XCF Signal Structures for communication (and/or CTCs).

Node Management

Joining the sysplex

Leaving the sysplex

Sysplex partitioning

Node management

- A system running z/OS can be a “node” in the sysplex cluster
  - Sysplex cluster currently limited to at most 32 systems
- System must IPL to join the sysplex
  - A system that is a member of the sysplex is said to be “active” in the sysplex
  - It is the responsibility of the IPLing system to verify that it can participate in the sysplex
    - **Same sysplex CDS, signal connectivity, same time reference, sysplex name**
    - *What if system is down level and cannot participate with sysplex of up level systems ?*
    - *What if system is up level and there are down level systems in the sysplex ?*

No update should ever require a simultaneous sysplex-wide IPL.*

- System must wait-state to leave the sysplex
  - The surviving peer systems “partition” the inactive system to remove it from the sysplex once they are certain that it has been appropriately “isolated” from shared resources.
  - Once “sysplex partitioning” is complete, the survivors are notified and can then perform appropriate cleanup for the system that was removed from the sysplex.

** Exceptions, if any, should be rare. The installation might choose to do so for their own reasons.*

Sympathy sickness

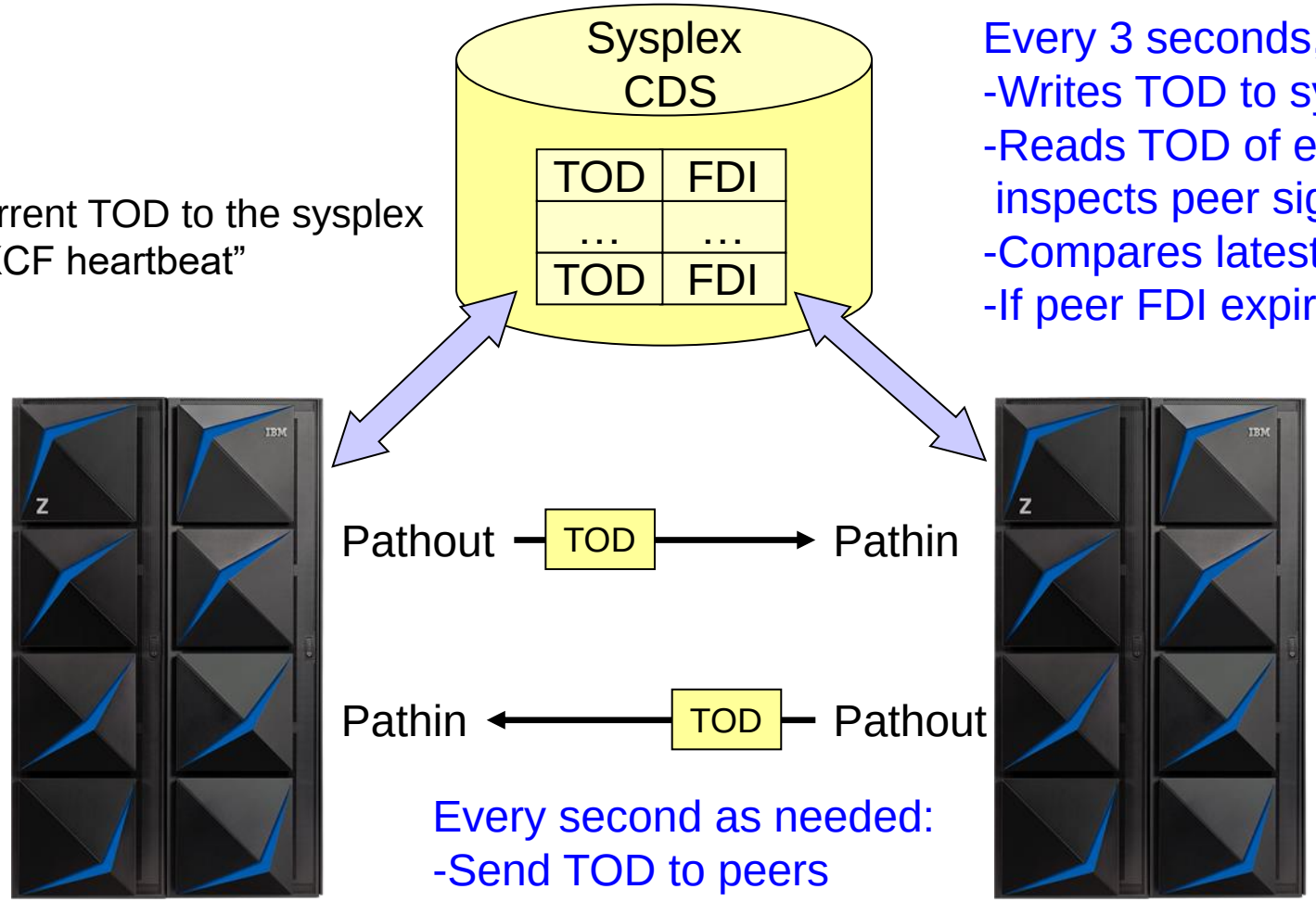
- When a system becomes unresponsive
  - It may be serializing shared resources with RESERVEs, ENQ's, Locks
  - It may stop sending responses or otherwise fail to participate in various "group" protocols
- Other work may experience delays and hangs
  - Problem compounds as the sympathy sickness spreads
  - Can quickly become difficult to distinguish culprit from victim
- Timely intervention is needed
  - One must correctly identify the culprit
  - Take corrective action

How does XCF determine that a system is unresponsive?

Using FDI to detect unresponsive system

This periodic writing of current TOD to the sysplex CDS is often called the “XCF heartbeat”

Sending system puts a TOD in nearly every signal it sends. “Peer signal TOD” is the most recent such TOD observed by the receiving system.



Every 3 seconds, each system:
-Writes TOD to sysplex CDS
-Reads TOD of each peer system and inspects peer signal TOD
-Compares latest peer TOD to peer FDI
-If peer FDI expired, take action

Every second as needed:
-Send TOD to peers

What happens if peer FDI expires …

Sysplex partitioning process (roughly)

- Update sysplex record in sysplex CDS to claim ownership
  - Could be racing with other systems. Winner will do partitioning.
- Issue message IXC101I “system being removed”
- (Maybe) Send “system going” signals to other systems (who then notify local members)
- **Isolate the system to ensure that it cannot be updating shared data**
- Update system record in sysplex CDS to indicate system no longer active
- **Cleanup and announce “system gone”**
  - Update sysplex/CFRM CDS records to indicate system’s members/connectors are gone
  - ENF signal “system gone”
  - XCF group “member gone” notifications
  - XES “disconnect failure” event notifications
  - ARM cross-system restart processing
- Update sysplex record in sysplex CDS to release ownership claim
- Issue message IXC105I “system removed”

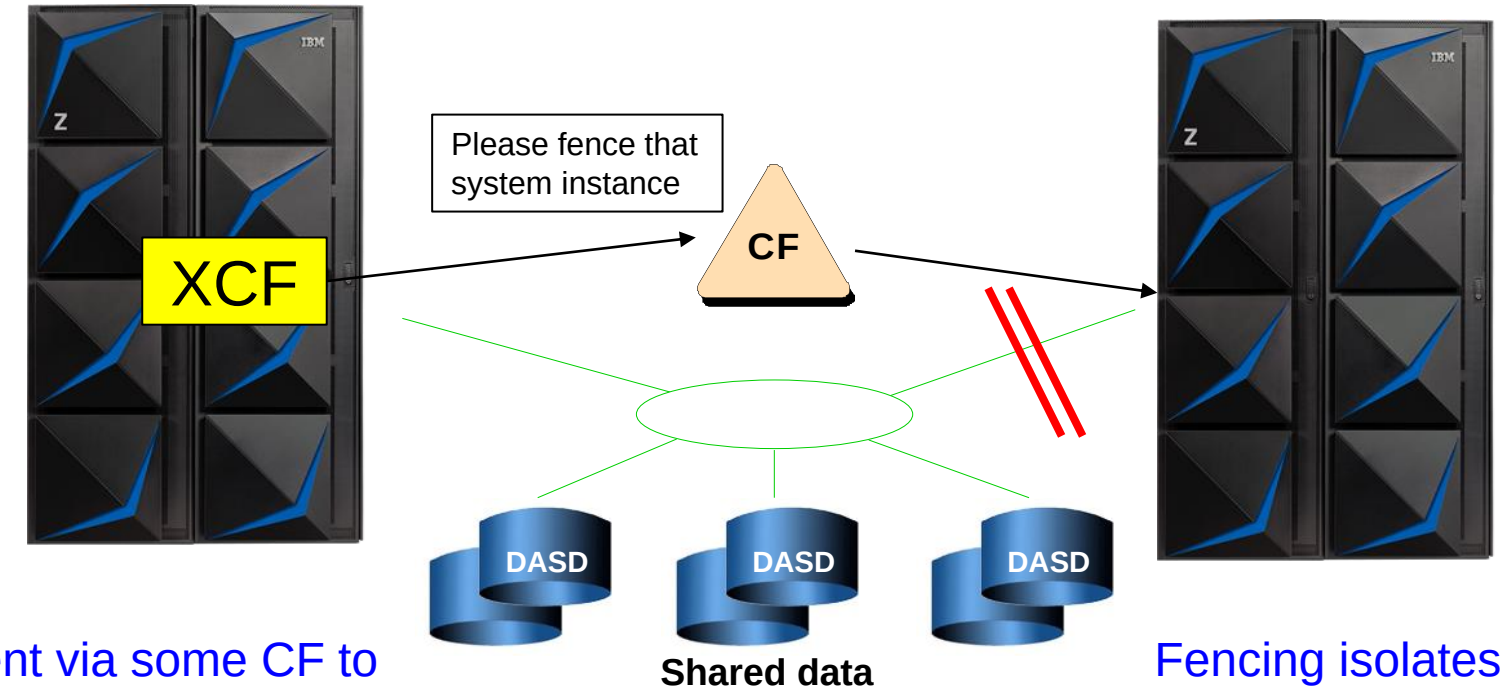
Some one system in the sysplex takes charge of coordinating the partitioning process.

Cleanup spawned by these notifications may continue beyond “system removed”.

Isolation is critical factor for safe removal of system from sysplex

- It is **critical** for the system to be “isolated” from shared resources before the survivors do any cleanup to remove it from the sysplex
- Failure to isolate the system creates the potential for **data corruption**
  - Corruption occurs when the system continues to make updates to data that are not under the control of the manager of the data
  - Even though all its CPs are in a wait-state, a system can continue to access data (ie, do I/O) !
  - A running system can certainly do so
- So XCF will NOT remove a system from the sysplex until it has been isolated
- Isolation techniques
  - Fencing (if parallel sysplex)
  - XCF use of BCPii
    - System Status Detection (SSD) protocol
  - Manual – operator resets the system and “tells” XCF that the reset was done (better be truth!)

Fencing



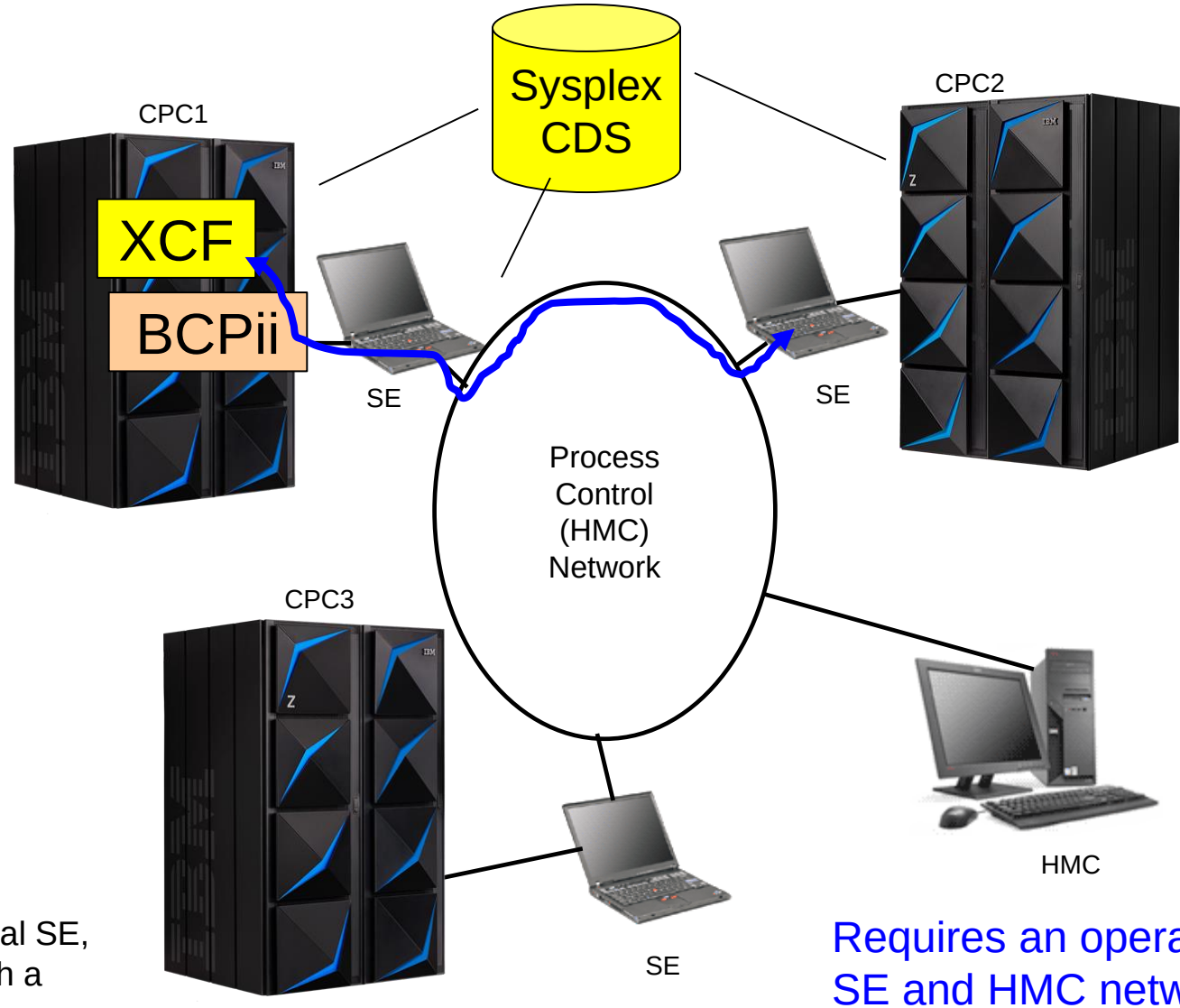
A command is sent via some CF to the target CEC to “fence” a particular system instance. The target image will not be able to initiate any new I/O and ongoing I/O will be terminated.

Fencing isolates a system so that it cannot access shared data, thus making it safe for the survivors to release serialization of the shared resources.

XCF with BCPii – System Status Detection (SSD) Partitioning Protocol

z/OS Images

(not VM guests)



XCF uses BCPii to:

- Obtain identity of z/OS image (gets stored in Sysplex CDS)
- Query status of remote CPC and z/OS image instance (is it in a wait-state?)
- Reset a z/OS image instance (if it needs to be isolated)

BCPii communicates with the local SE, which can then communicate with a remote SE via the HMC network.

Requires an operational SE and HMC network

Couple Data Sets (CDS)

Primary and Alternate

Sysplex CDS and Function CDS's

Utilities to create CDS and policies

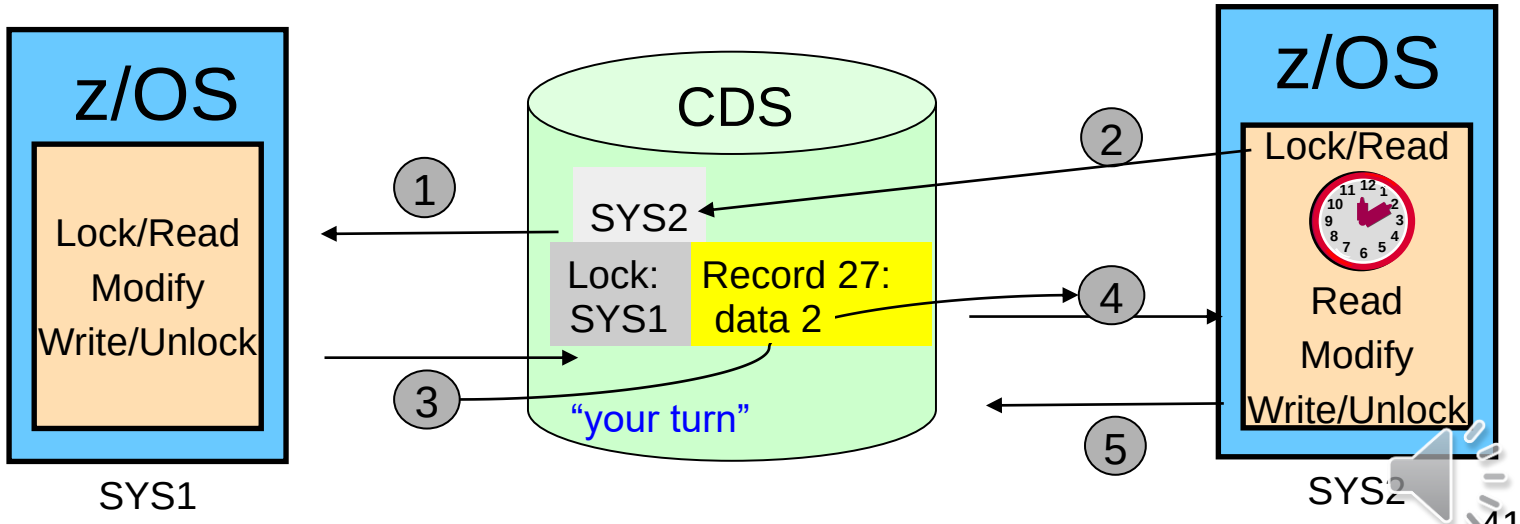
Administrative and active policies

Couple Data Sets (CDS)

- Provide means to harden data and share it between systems in the sysplex with the serialization needed to maintain its integrity
- Accessed via XCF channel programs and protocols
  - Typical usage
    - Lock record and read content into storage
    - Modify in-store copy
    - Write modified content to CDS and unlock record
  - “Lock steal protocols” to mitigate sympathy sickness

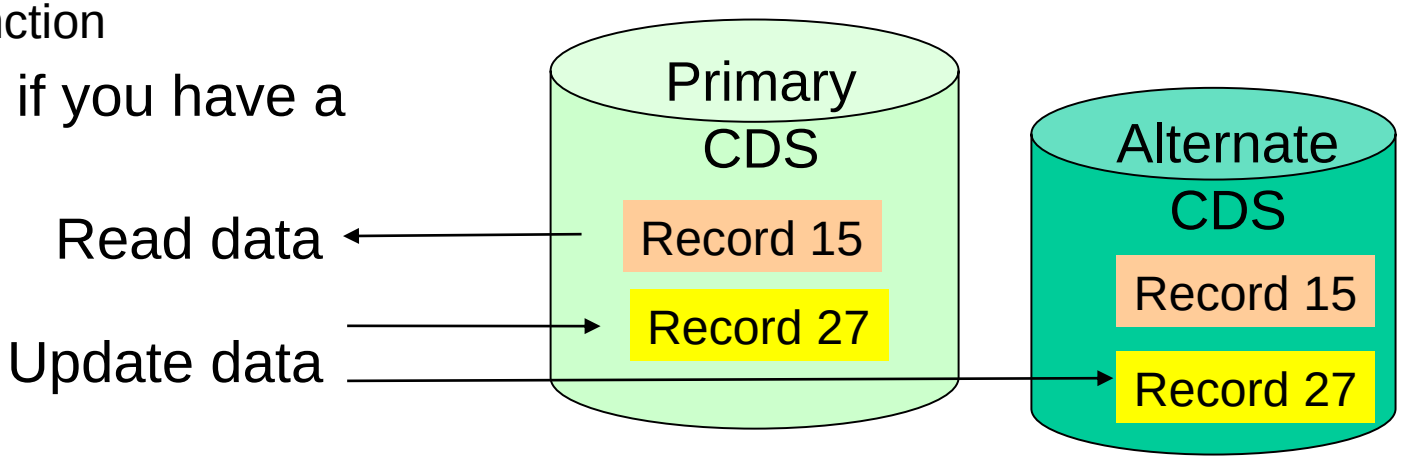
XCF does NOT use traditional access methods for CDS I/O requests.

1. SYS1 reads record, locks it.
2. SYS2 tries the same, but must wait since SYS1 has the lock.
3. SYS1 writes new data 1, sends signal to SYS2 to continue.
4. SYS2 reads (still locked) record.
5. SYS2 writes new data 2, unlocks record.

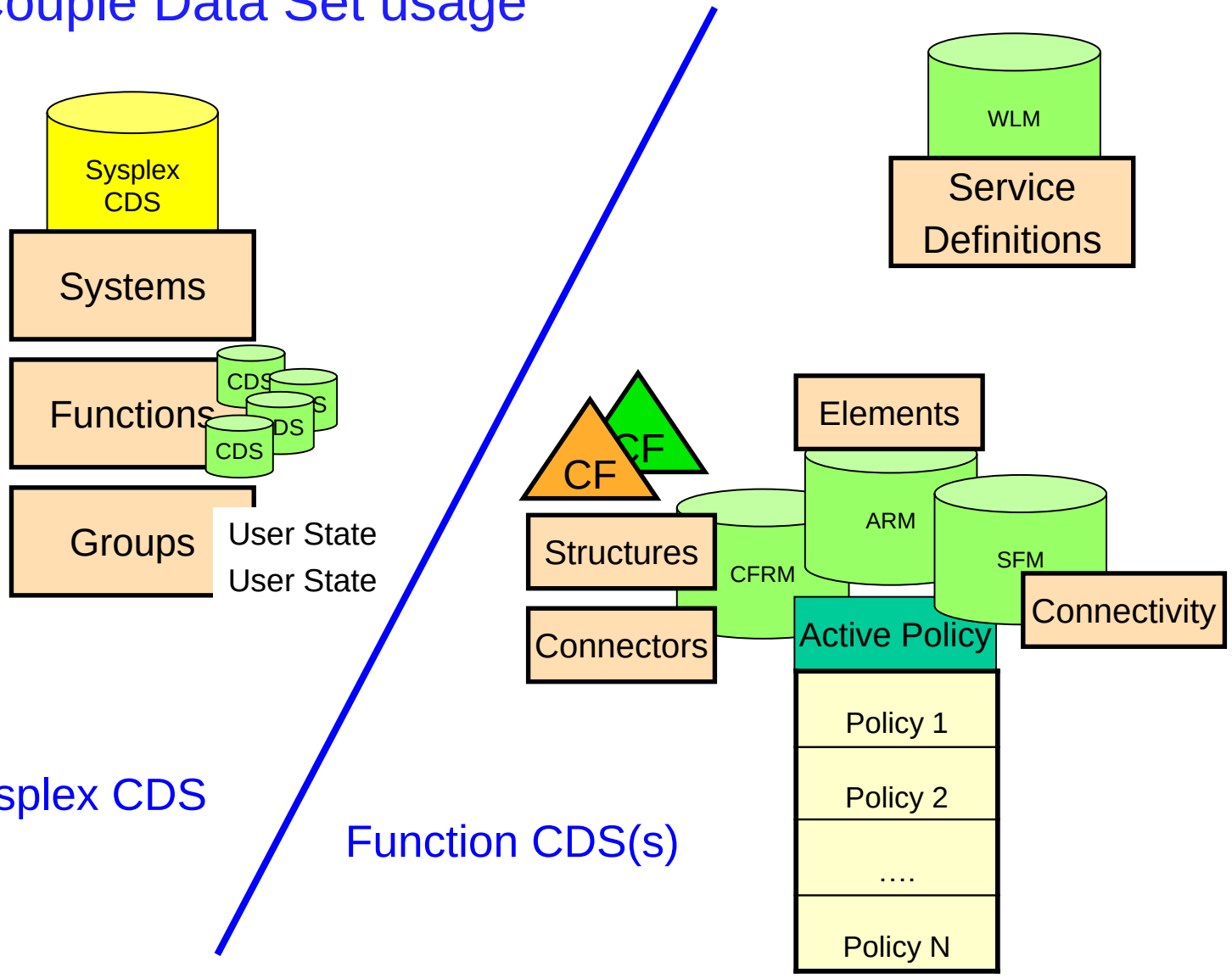


Define primary and alternate couple data sets to avoid single point of failure

- Normally run with both Primary and Alternate CDS
  - Read requests directed to primary
  - Update requests written first to primary, then to alternate
    - Both must complete for I/O request to finish
- Sysplex automatically switches to alternate if primary fails
  - Have automation in place to add a new alternate CDS if either the primary or alternate fails
- Loss of both primary and alternate can be disastrous
  - Wait-state of every system in the sysplex, or
  - Significant loss of sysplex function
- z/OS Health Checker warns if you have a single point of failure
  - XCF_CDS_SPOF



Couple Data Set usage



XCF Services

System status monitoring

Services to enable multi-system “applications”

- Groups and members
- Member status monitoring
- Communication

XCF services for multisystem applications

- XCF is central point of control for determining whether a given system can effectively participate in the sysplex
  - XCF determines who is in or out of the sysplex (or needs to be removed when “sick”)
  - Exploiters can rely on the common time reference being intact at all times
  - Exploiters can assume XCF signal connectivity exists between every pair of systems in the sysplex
- The various instances of the multisystem application are “members” of an XCF “group”
  - Each exploiter has a uniquely named group of their own choosing
  - Each member has a name of their own choosing that is unique within the group
  - Each system in the sysplex can have zero or more members of a given group
- XCF Services
  - Group Services which are used to define the application’s members to XCF (required)
  - Monitor Services to have XCF monitor member health (optional, relatively few use this)
  - Communication Services to send signals between members of the group (optional, but most do so)

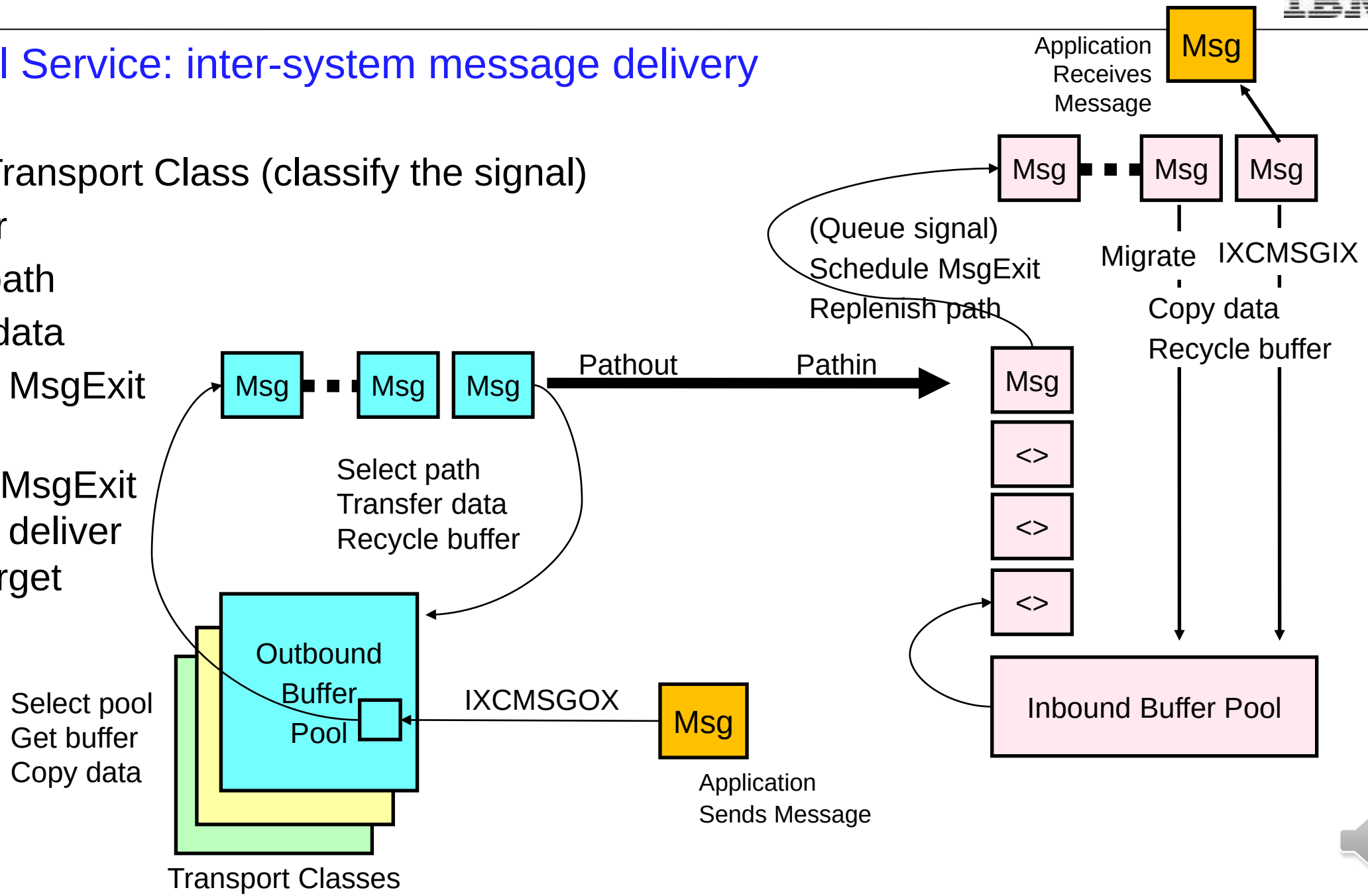
XCF Signal Paths

XCF signal service is responsible for:

- Delivering messages (signals) between members of an XCF group
- Managing the physical resources used to transport those messages within the sysplex

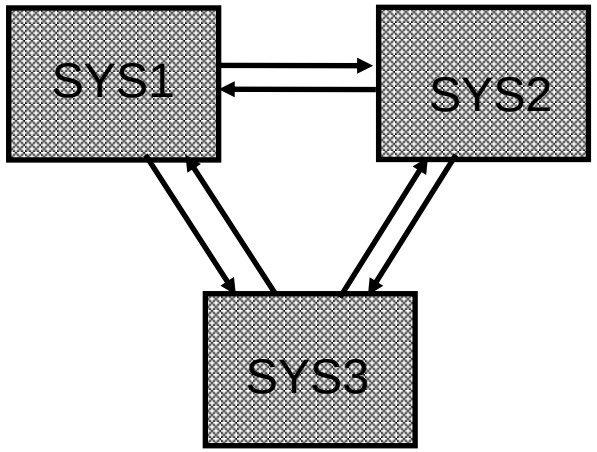
XCF Signal Service: inter-system message delivery

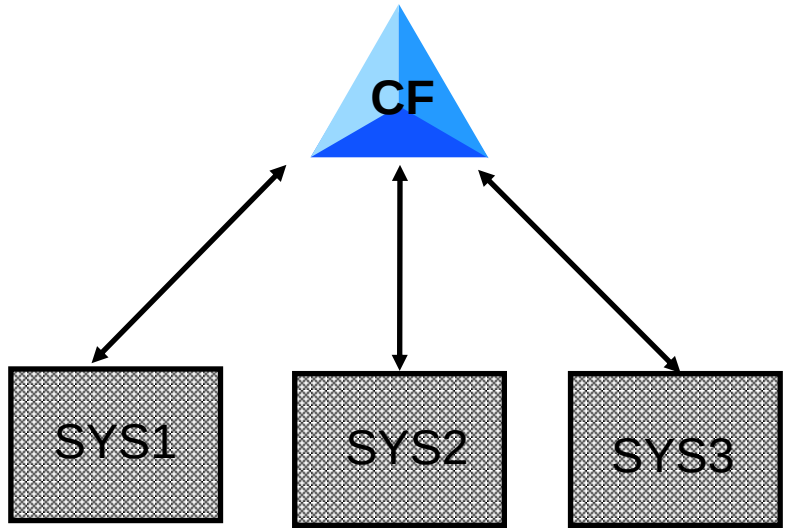
- Choose Transport Class (classify the signal)
- Get buffer
- Choose path
- Transfer data
- Schedule MsgExit SRB
- Call user MsgExit routine to deliver msg to target member



XCF signal paths

Provide point to point communication between systems for message passing between operating system components, middleware, and applications that exploit XCF sysplex services





CTC Signalling Path

- Requires a device for each direction
- You define point to point connectivity
- Message stream
- Good for small messages
- Harder to define when lots of systems O(n^2)
- Drives SAP utilization

CF List Structure Signalling Path

- Can be used in both directions
- XCF self defines list paths for full connectivity
- Mailbox approach
- Good for all message sizes
- Easier to define when lots of systems O(n)
- Increased SRB time in XCF address space
- Fastest data transfer rate

Coupling Facility

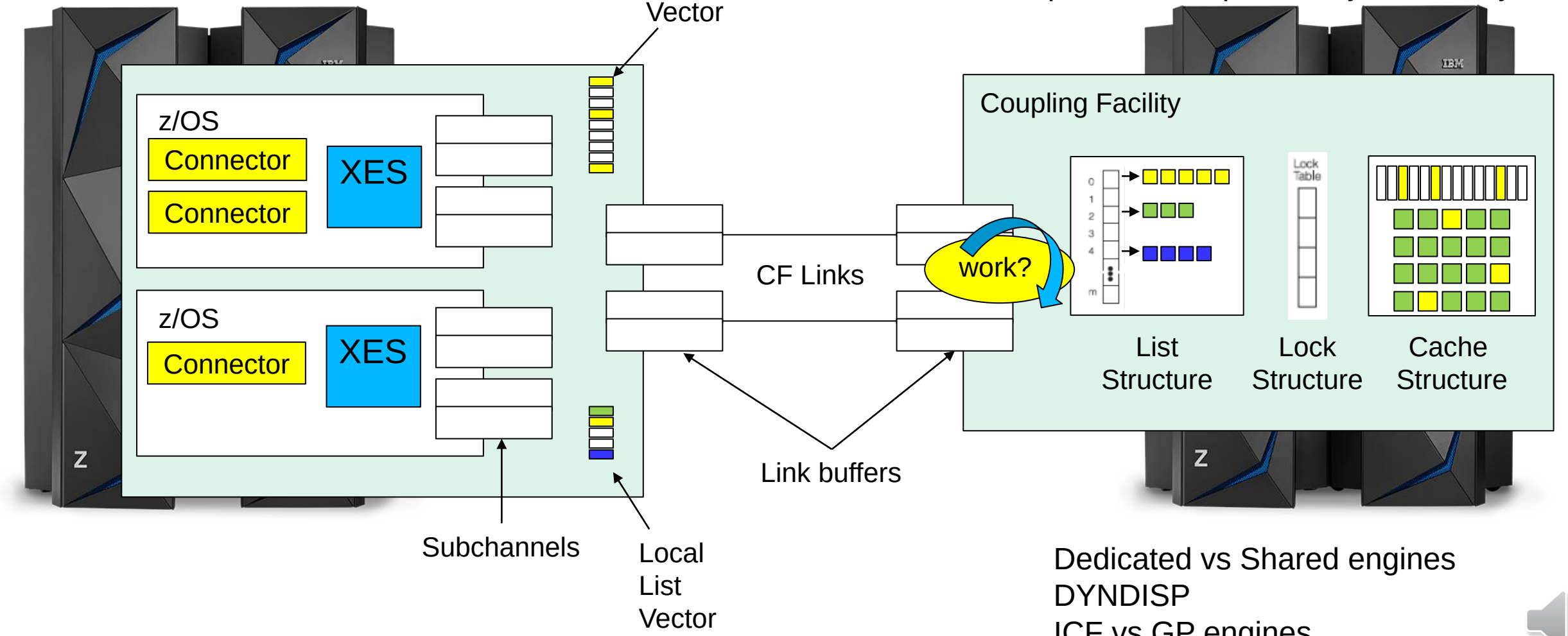
An LPAR running Coupling Facility Control Code (CFCC)

Hosts CF structures (list, lock, cache) per the active CFRM policy

Links for sending requests from z/OS to the CF

Coupling Facility overview

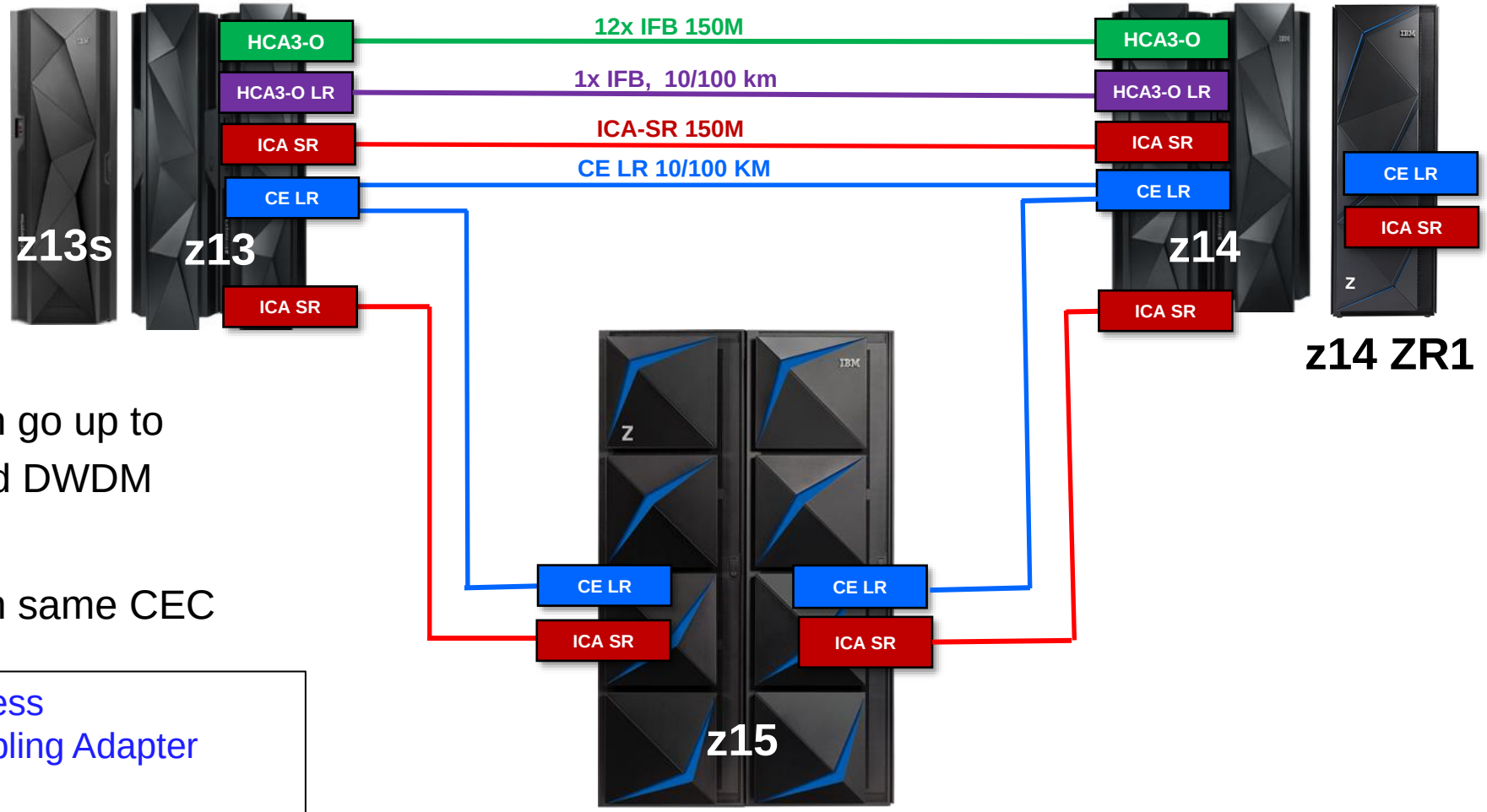
CFCC – Coupling Facility Control Code
CFLEVEL – similar to “release”
Pure memory, no paging (or I/O)
Exploiters manipulate objects, not bytes



Dedicated vs Shared engines
DYNDISP
ICF vs GP engines

CF links provide coupling connectivity

- Links tend to evolve with HW generations
- Some flavor of:
  - Short range (SR)
    - <= 150 meters
  - Long range (LR)
    - <= 10 km, though can go up to
    - 100 km with approved DWDM
- Internal Channels (IC)
  - Connect LPARs to CF on same CEC



CE - Coupling Express
ICA - Integrated Coupling Adapter
IFB - Infiniband
1x - IFB version of LR
12x - IFB version of SR
HCA3-O Host Channel Adapter-Optical

Coupling Facility request processing

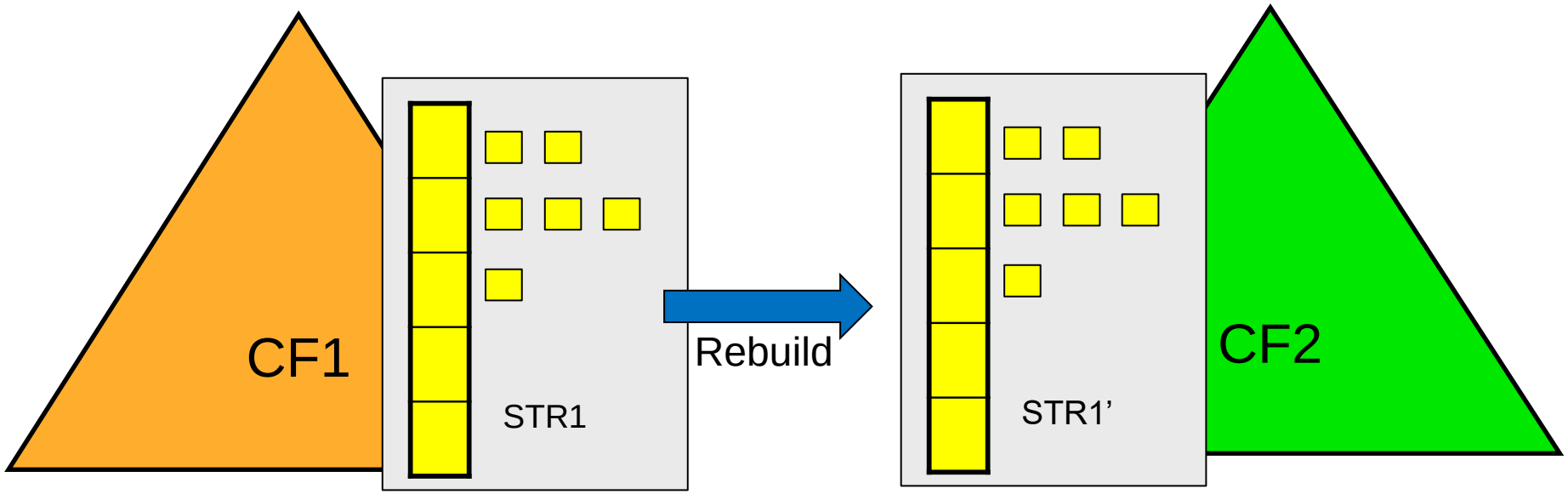
- Requests can be processed synchronously or asynchronously
  - Synchronous: send request to CF then “wait” for results to be returned
    - The concern here is opportunity cost for the z/OS image
    - XES may decide that it is more cost effective to process the request asynchronously
  - Asynchronous: send request to CF, then do other work while the CF processes the request. z/OS will eventually observe that the request has completed.
    - The concern here is application response time
    - How quickly can z/OS observe completion and get the results to the application
  - This is really a function of z/OS interaction with the CF link as opposed to the CF itself
- XES heuristic algorithm may convert a synch request to async in an attempt to maximize work on the z/OS image
  - You can choose the conversion thresholds (but most should not do so)

Structure Rebuild

A key recovery technique for parallel sysplex

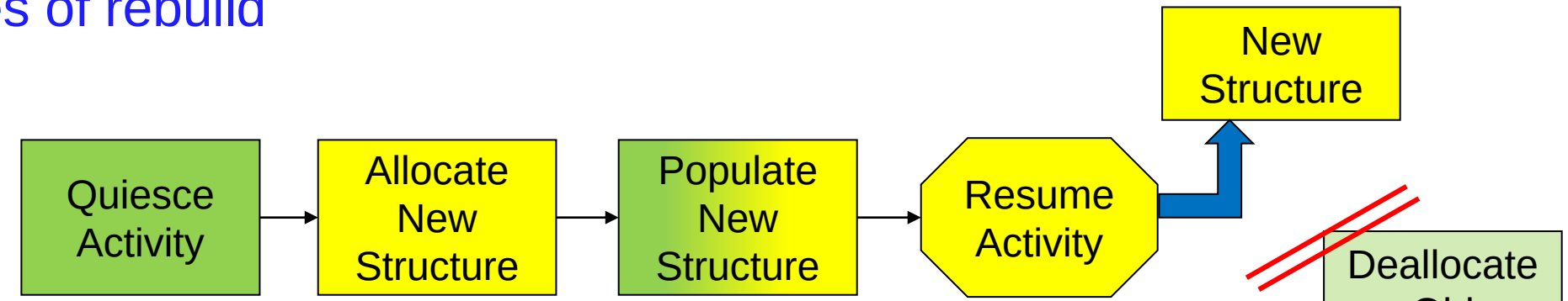
Choice of sysplex configuration matters!

CF structure rebuild



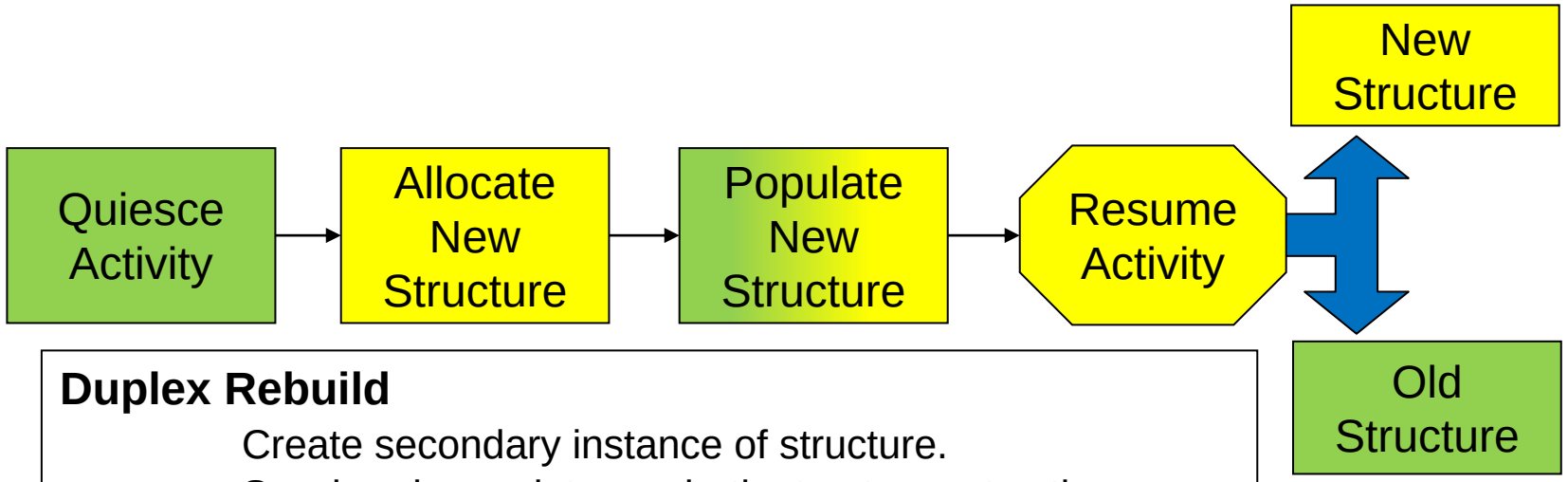
Rebuild is the process by which the sysplex allocates a new instance of a given CF structure, populates that structure with data, and proceeds to use the new structure instance.

Two types of rebuild



Simplex Rebuild

Create and use new instance of structure.
Discard the old instance.

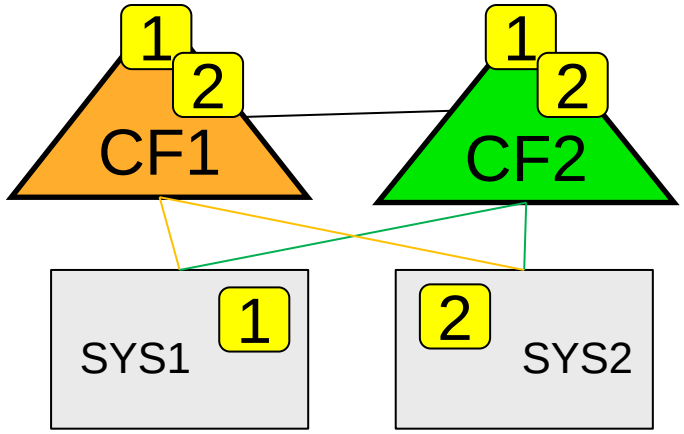


Duplex Rebuild

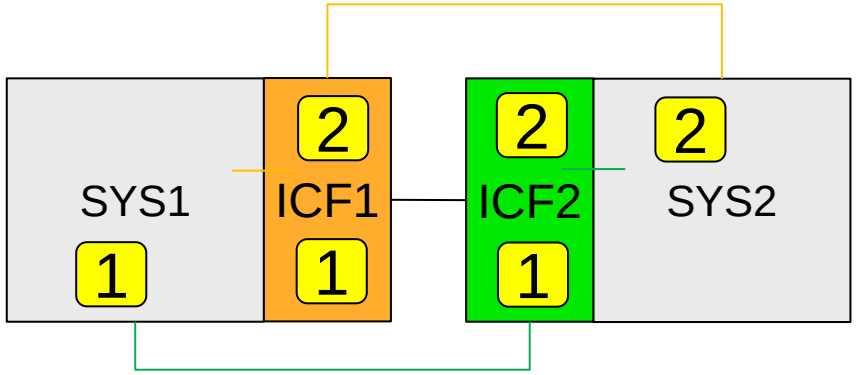
Create secondary instance of structure.
Synchronize updates so both structures stay the same.
Can quickly failover to one or the other.

Using duplexed structures for fast, robust failure recovery

- For any of these failures:
  - Failure of any one CF
  - LossConn to any one CF from any source (one or all)
  - Failure of any one CEC
- Normal operation continues with structure in simplex mode in the surviving/accessible CF



Duplexed structure recovery



Duplexing options:
- SM synchronous
- SM async for lock structures
- UM for Db2 GBP

Common Time Reference

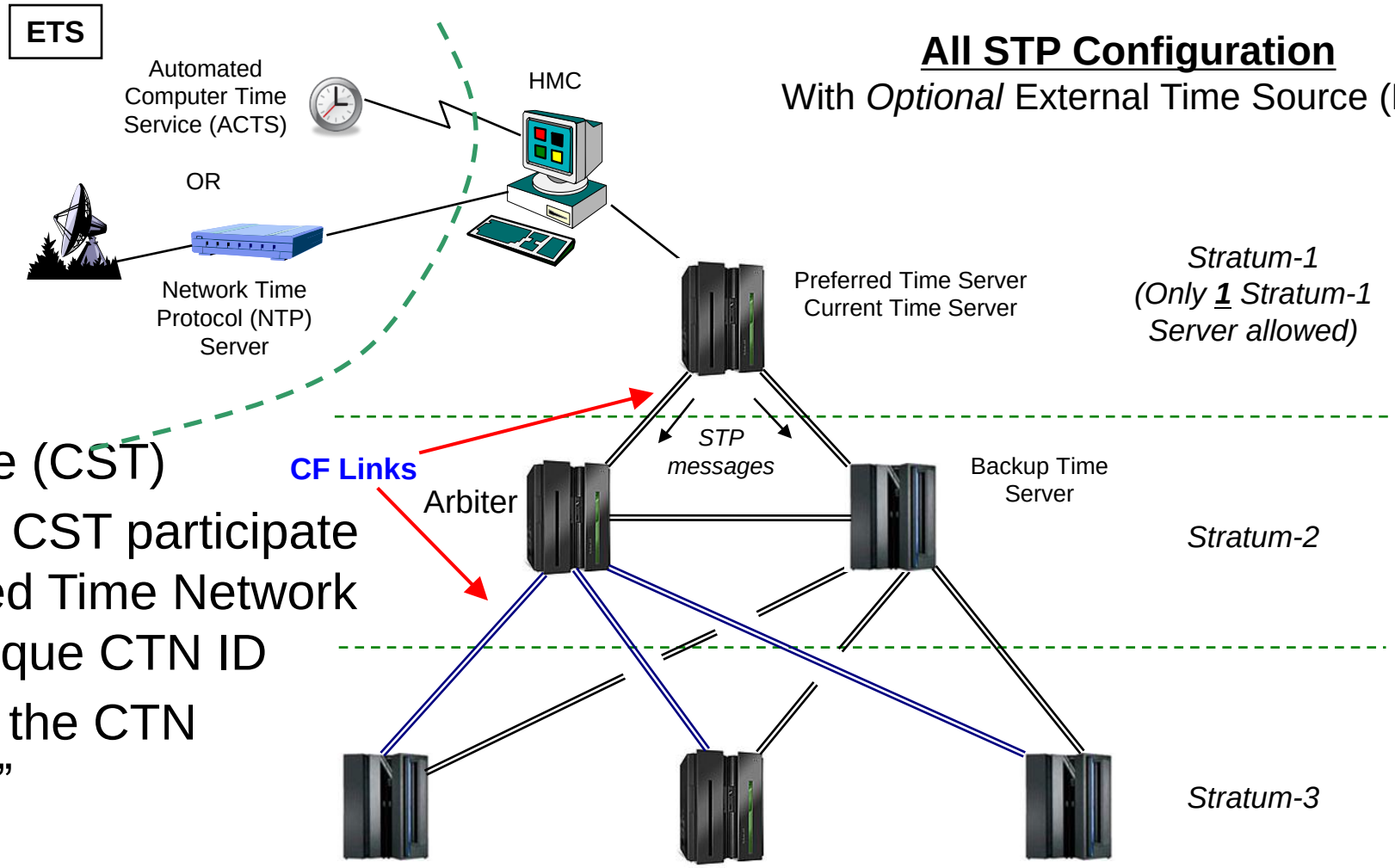
Sever Time Protocol (STP)

Time synchronization

- z/Architecture ensures that two successive time stamps are unique within a CEC
  - Critical if you want to have unambiguous ordering of events
- Through a “common time reference”, this guarantee can be extended to time stamps taken on different CECs
  - Thus there can be an unambiguous ordering of events within the sysplex
  - A tremendous simplification for applications and people
- The “common time reference” is implemented via Server Time Protocol (STP)

Common time reference - Server Time Protocol (STP)

- Timer signals distributed in layers called “stratums”
- Servers use these timer signals to synchronize their TOD clocks to Coordinated Server Time (CST)
- Servers synchronized to CST participate in a common Coordinated Time Network (CTN) identified by a unique CTN ID
- Designated servers with the CTN are assigned STP “roles”



All CPCs working with the same Coordinated Timing Network (CTN) ID
Supports multi-site Sysplex of at least 100km

High Availability with Sysplex

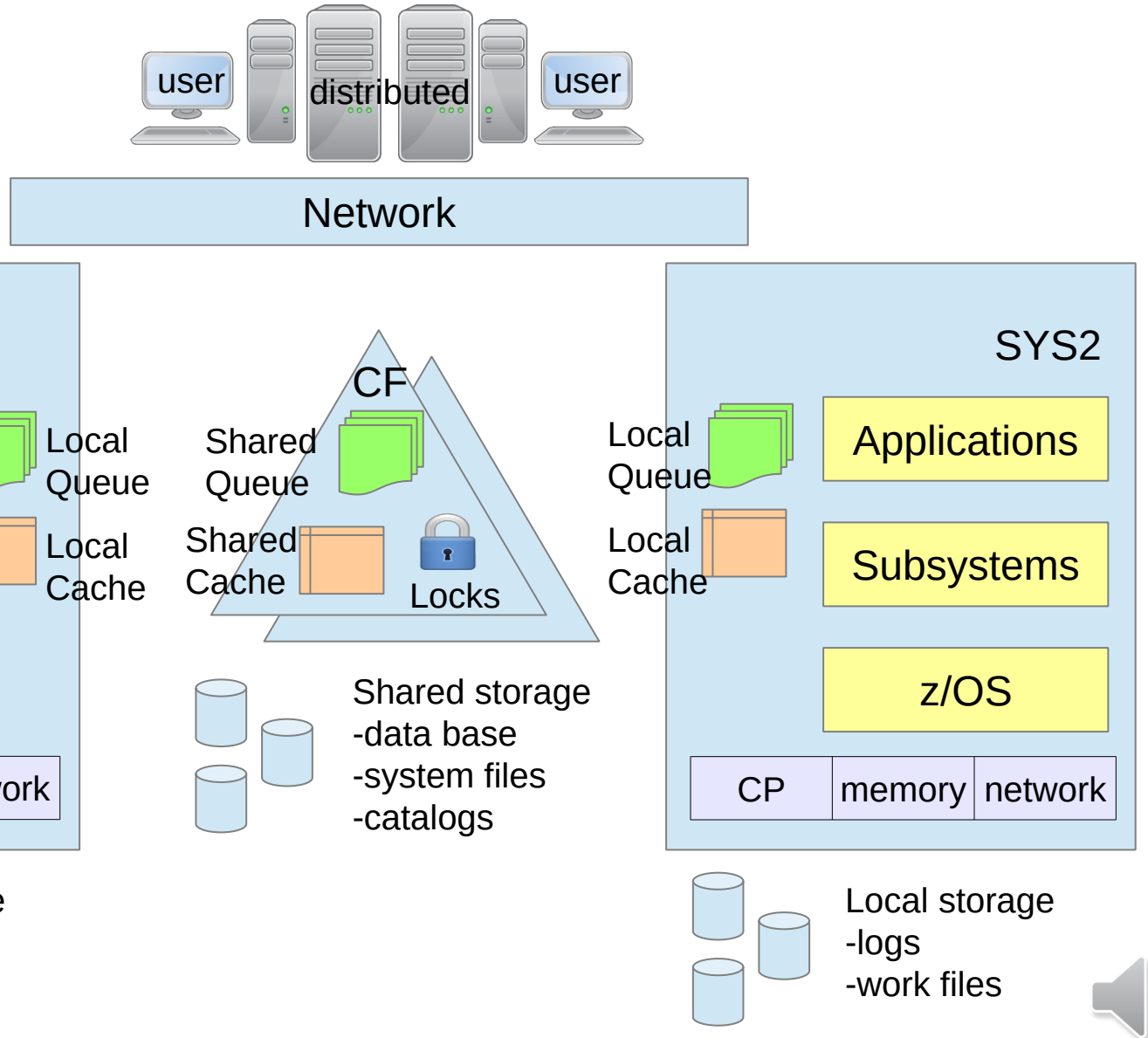
Redundancy alone is not enough

“Cloning” for high availability

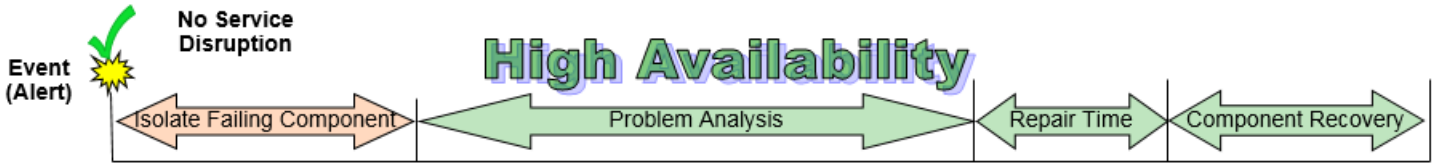
Ideally, every system is capable of processing all work.

Work automatically flows to where it gets the best performance.

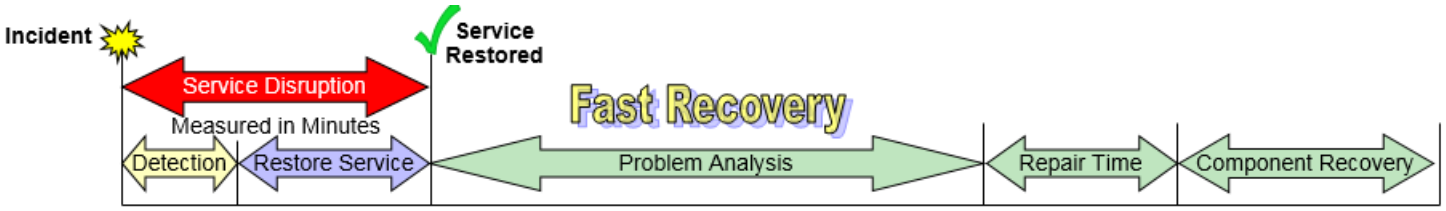
After a failure, the survivors simply continue to process the shared work.



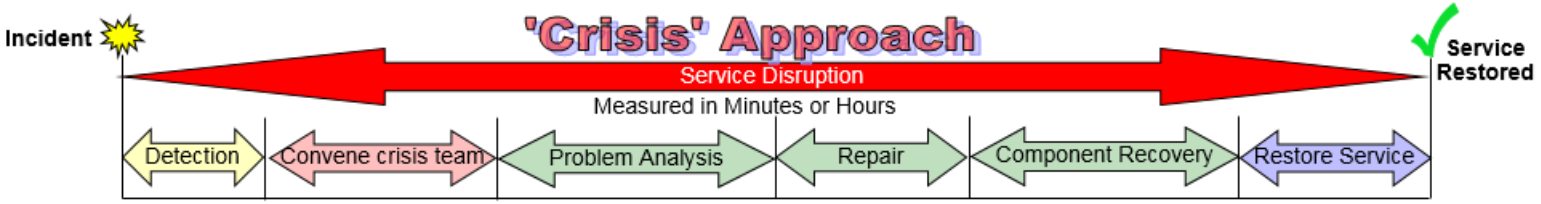
Service restoration models



Ideally, a fault tolerant architecture and infrastructure allows service to continue uninterrupted despite component failure.



Not all failures will be masked. Rapidly restoring service minimizes the business impact.



When service restoration requires human intervention, outages tend to have unpredictable (long) duration. Risks significant business impact.

Conclusion

A short recap

Summary

We covered basic sysplex componentry and terminology

- Sysplex topologies
  - Base vs parallel
  - Single vs multisystem
- Couple Data Sets
  - Sysplex vs Function
  - Format Utility
  - Administrative Data Utility
  - CFRM Policies
- XCF signal paths
  - CTC
  - Signal Structures
- Server Time Protocol
  - Synchronized time
  - Stratum levels
- Coupling Facilities
  - CF Structures
  - CFLEVEL
  - CF Links
- XCF groups and members
- XES connectors

Please submit your session feedback!

- Do it online at https://conferences.gse.org.uk/2021/feedback/4BJ
- This session is 4BJ

1. What is your conference registration number?

This is the three digit number on the bottom of your delegate badge

2. Was the length of this presention correct?

1 to 4 = "Too Short" 5 = "OK" 6-9 = "Too Long"

1 2 3 4 5 6 7 8 9

3. Did this presention meet your requirements?

1 to 4 = "No" 5 = "OK" 6-9 = "Yes"

1 2 3 4 5 6 7 8 9

4. Was the session content what you expected?

1 to 4 = "No" 5 = "OK" 6-9 = "Yes"

1 2 3 4 5 6 7 8 9

GSE GUIDE SHARE EUROPE UK REGION

GSE UK Conference 2021 Charity Raffle

- The GSE UK Region team hope that you find this presentation and others that follow useful and help to expand your knowledge of z Systems.
- Please consider showing your appreciation by kindly donating to our charities this year, Royal National Lifeboat Institution (RNLI) & Guide Dogs for the Blind. Then follow the link on your receipt to enter your receipt number & amount donated into the GSE Raffle. You will get a raffle entry for every pound donated.
- Follow the link below or scan the QR Code:

http://uk.virginmoneygiving.com/GuideShareEuropeUKRegion

Supporting

Become a member of GSE UK

- Company or individual membership available
- Benefits include:
  - GSE Annual Conference: Receive 5 free places + 2 free places for trainees
  - 20% discount on fees for IBM Technical Conferences
  - 20% on IBM Training Courses in Europe
  - 15% discount for IBM STG Technical Conferences in the USA
  - 20% discount on the fee for taking the Mainframe Technology Professional (MTP) exams
  - European events – via GSE HQ

- Contact membership@gse.org.uk for details