IBM TRAINING

Z20

Pushing the Limits of Parallel Sysplexes:
Bigger, Smaller and Further Apart

Joan Kelley

IBM
SYSTEM z9 AND zSERIES EXPO
October 9 - 13, 2006

Orlando, FL

Trademarks

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

CICS	VTAM	RMF	DB2
MVS	IMS	z9	MQSeries
RACF	Parallel Sysplex	zSeries	GDPS

* 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.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
UNIX is a registered trademark of The Open Group in the United States and other countries.
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.

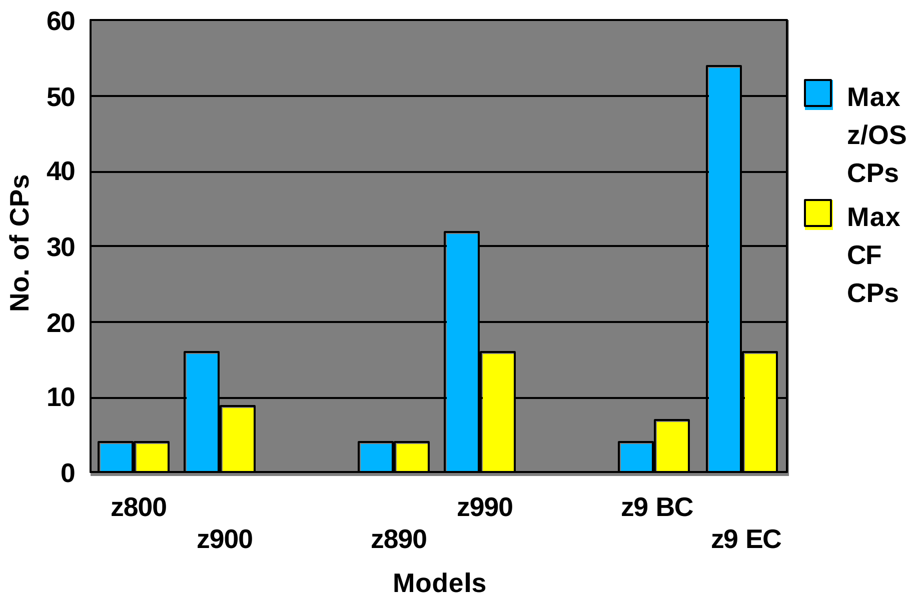
Getting bigger

Some installations are growing rapidly

Consolidated images and growing workloads->

- More processors
  - Faster technology, more CPs
- More images
  - Specialized CPs
- Larger Coupling Facilities
  - More CPs, more structures
  - Faster link technology
- Continuous availability
  - Test configurations, concurrent updates

More CEC and CF processors



CPENABLE recommendations

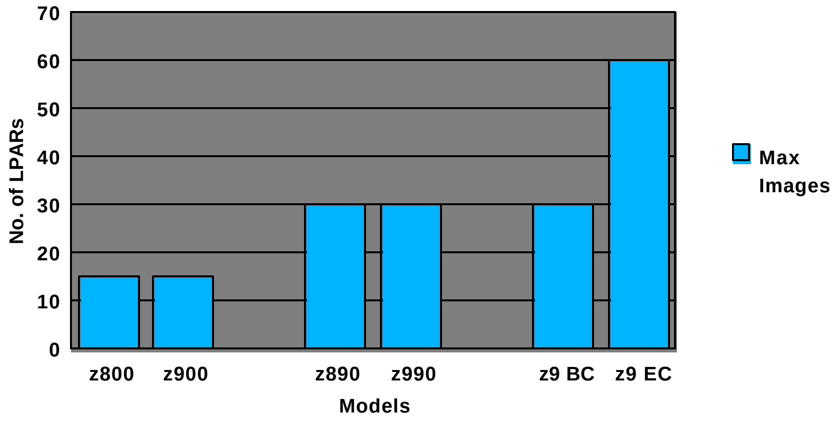
Recommended CPENABLE settings

- To maximum thruput (ITR) at
- Very slight cost to response time
- Biggest impact with high N-way or high I/O rates

Processor Family	Basic Mode	LPAR Dedicated	LPAR Shared
IBM System z9	N/A	10,30	10,30
zSeries 990 (2084) zSeries 890 (2086)	N/A	10,30	10,30
zSeries 900 (2064) zSeries 800 (2066)	10,30	10,30	0,0

Latest change (z990 recommendation) due to large number of engines -> more overhead when all enabled for I/O

More LPAR images supported



R.O.T - Logical to physical ratio - 2 or 3 at most

To reduce high LPAR overhead with many images install OA12416

Specialized CPs

Larger processors mean larger software licensing fees.
This can be reduced defining specialized CPs:

- Defined (LICed) before IML
- Excluded from model number, so do not factor into calculation of software licensing fees

MVS operating system does not run on specialized CPs.

1. ICFs - Exclusively CF microcode
2. IFLs - Exclusively Linux operating systems
3. zAAPs - Java workloads can be offloaded to zAAPs
   - Introduced on z890s and z990s
4. zIIPs - Certain DB2 processing intensive work can be offloaded to zIIPs..
   - Introduced on z9s

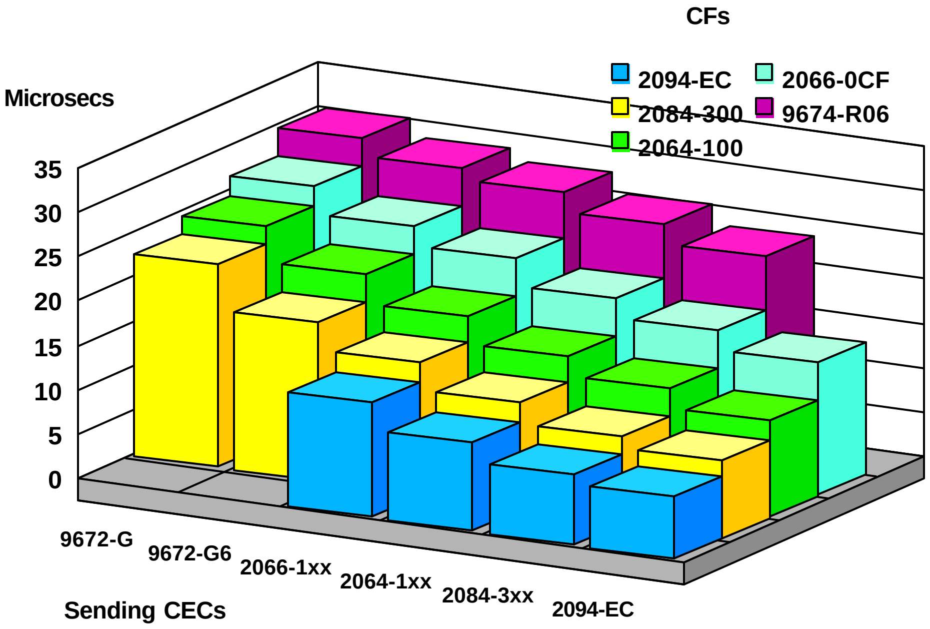
Upgrading all technologies

In Parallel sysplex environment, it's not enough to improve the speed of the sending processor

For synchronous requests, the sending CEC spins, using up cycles, waiting for request completion.

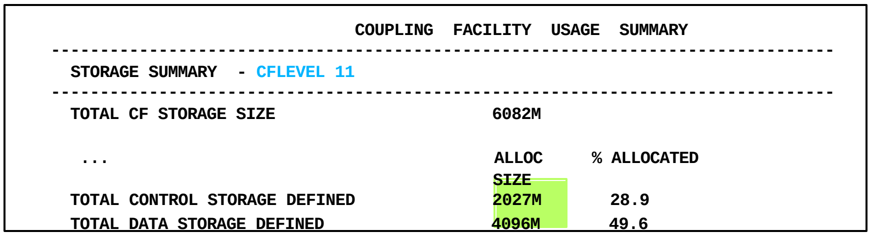
- Faster processors use more cycles in the same elpased time
- So have to improve speed of the CF and speed of the CF links to stay at the same overhead.

Improvements in CEC/CF technology

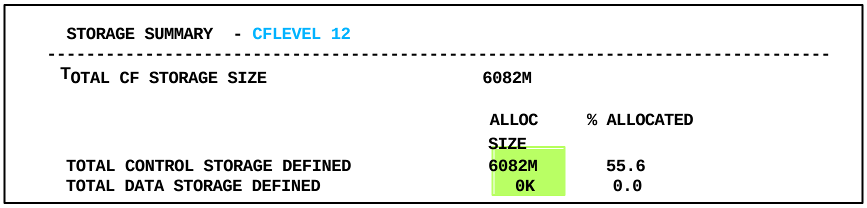


Changes to CFCC to support growth

- CFLevel 12
  - 64 bit Support, removal of the 2G line



```
                              COUPLING  FACILITY  USAGE  SUMMARY
 --------------------------------------------------------------------------------
  STORAGE SUMMARY  - CFLEVEL 11
 --------------------------------------------------------------------------------
  TOTAL CF STORAGE SIZE                         6082M

   ...                                          ALLOC     % ALLOCATED
                                                SIZE
  TOTAL CONTROL STORAGE DEFINED                 2027M        28.9
  TOTAL DATA STORAGE DEFINED                    4096M        49.6
```



```
  STORAGE SUMMARY  - CFLEVEL 12
 --------------------------------------------------------------------------------
 TOTAL CF STORAGE SIZE                         6082M

                                                ALLOC     % ALLOCATED
                                                SIZE
  TOTAL CONTROL STORAGE DEFINED                 6082M        55.6
  TOTAL DATA STORAGE DEFINED                       0K         0.0
```

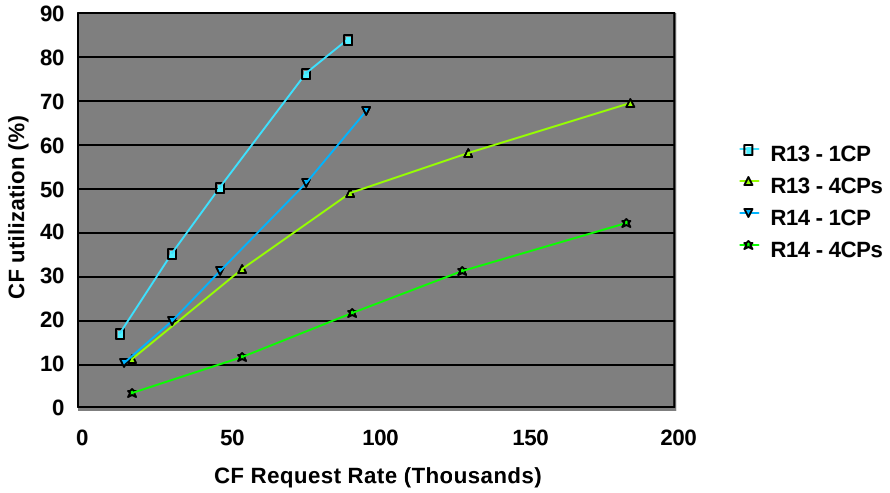
  - 48 Concurrent Tasks
  - Support for >15 LPARS

Changes to CFCC, cont.

- CFLevel 13
  - Castout performance improvements for large DB2 group bufferpools
- CFLevel 14
  - Improvement in efficiency of dispatching CF requests
  - Refinement of CF Utilization calculation to eliminate time spent searching for work (as opposed to executing work)
  - Reduction of "CF to CF" service times in duplexed environment
  - Elimination of special handling for structure likely to get RC=19 conditions.

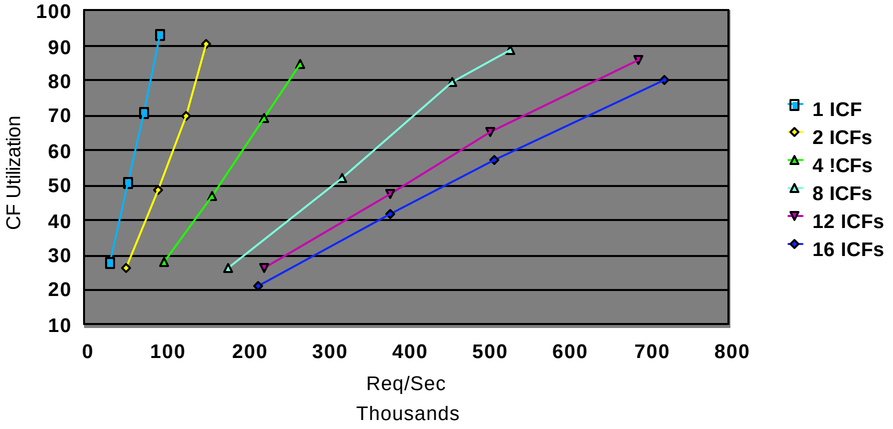
Impact of CFLevel 14 - Simplex requests

Elimination of waiting times from CF Utilization provides a more accurate view of CF capacity, especially as number of CF CPs increases



Maximum CF request rates

Typical mix of list, lock and cache reqs

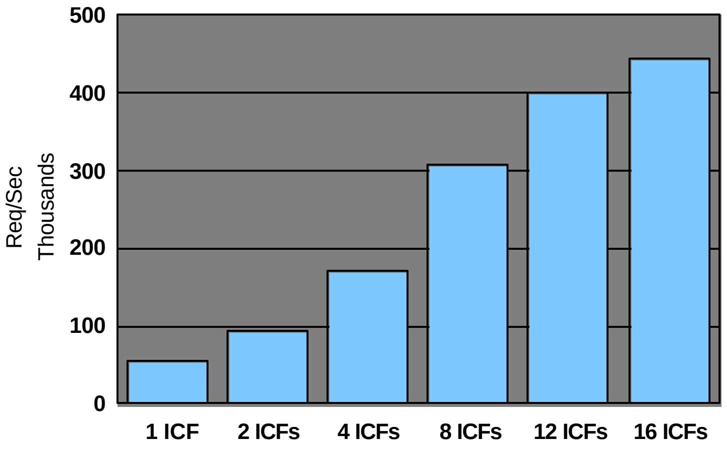


Maximum request rate for certain number of ICFs may not give

- Enough spare capacity needed for CF failover
- Acceptable response time

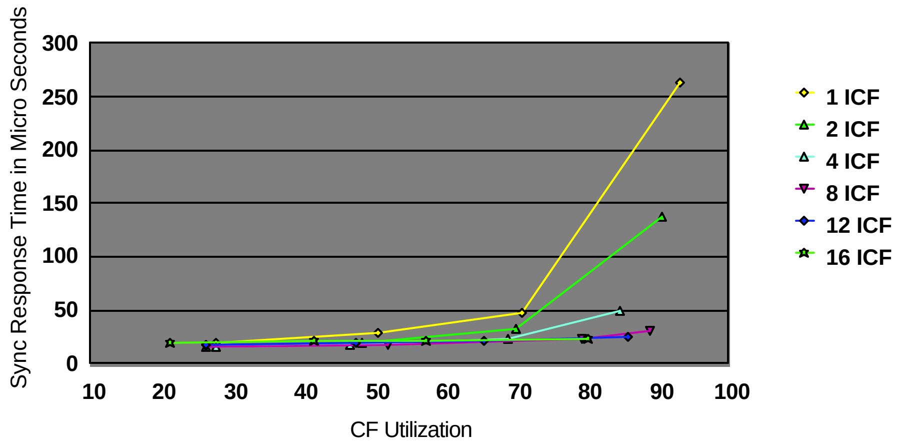
Maximum req rate for 50% CF utilization

If require spare capacity for CF failover, what are the maximum rates for a typical mix of requests?



Response time at higher utilizations

If white space is not a consideration or rate is temporary. can run higher request rates at decent response times with more ICFs

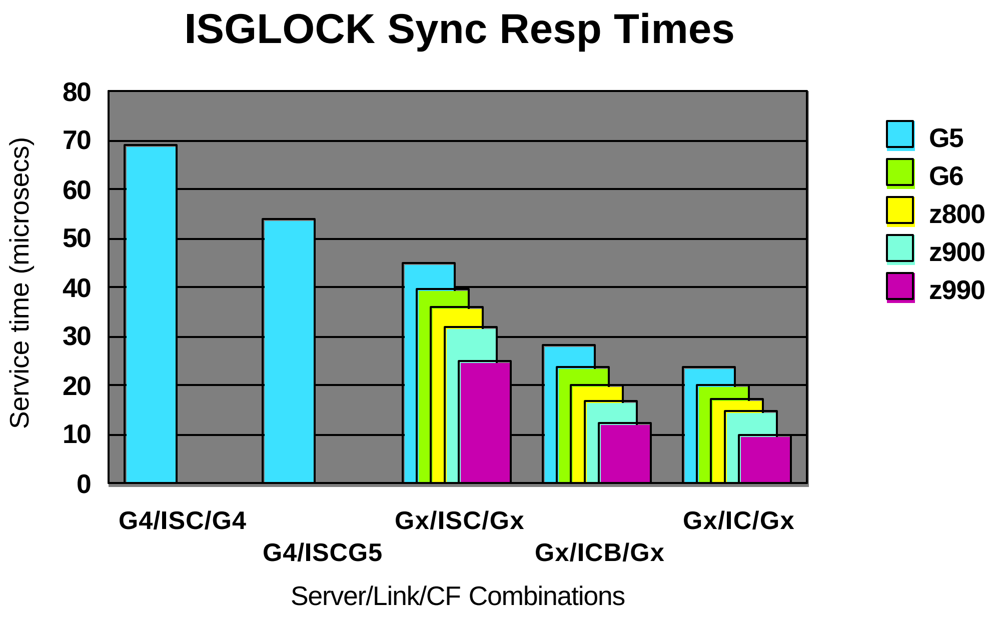


Improvement in CF link speeds

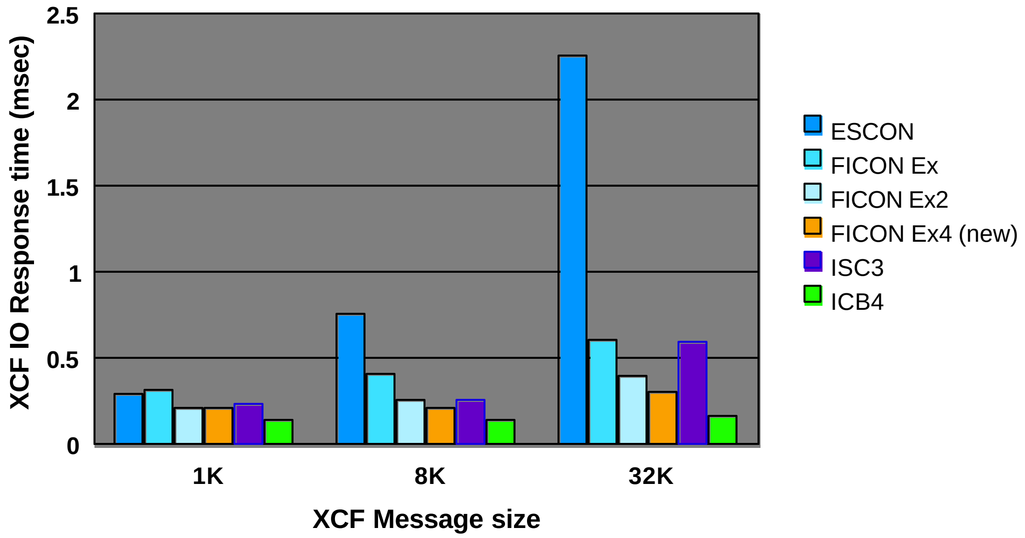
Model	ISC	ISC-3	ICB-2	ICB-3	ICB-4	IC
9672 G5/G6	100 MB/sec	-	250 MB/sec	-	-	700 MB/sec
z800	-	200 * MB/sec	-	500 MB/sec	-	1125 MB/sec
z900	-	same	250 MB/sec	500 MB/sec	-	1400 MB/sec
z890	-	same	250 MB/sec	500 MB/sec	1500 MB/sec	3500 MB/sec
z990	-	same	250 MB/sec	500 MB/sec	1500 MB/sec	3500 MB/sec
z9	-	same	-	500 MB/sec	1500 MB/sec	5000 MB/sec

* some exceptions...see later chart

Combined effect of technology improvements



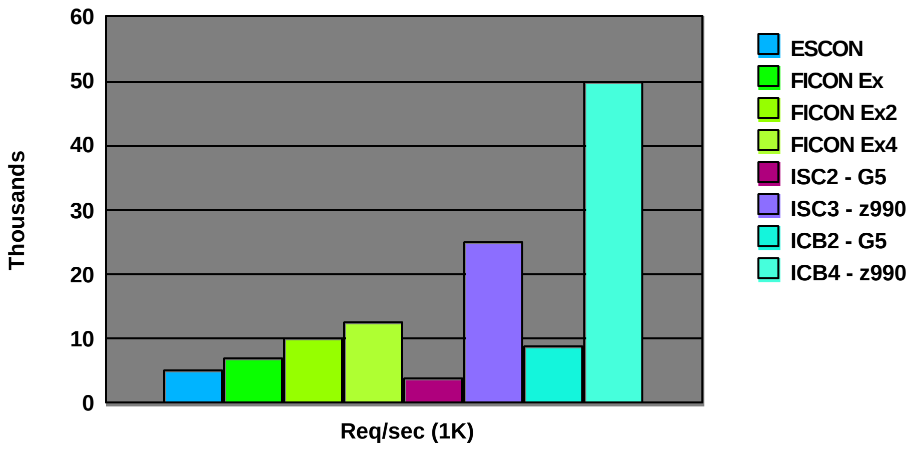
XCF - Response Time with Varying Message Size



Capacity of XCF paths

Measured when "best" response time doubles

- Depends on size of message, how many paths are defined, and other users of path (ex, VTAM)



Increasing number of structures

Number of structures growing

New exploitation of coupling - new structured

System managed Duplexing (1 structure -> 2 structure)

Number of structure allowed has increased 256 ->512, now

1023 Structures / CF

1024 Structures allowed in / policy

But this means longer recovery/rebuild times

Some performance improvements during system failure recovery and cleanup were introduced in APAR OW48624

- Only one system initiates cleanup
- Confirmation process more efficient
- CFRM I/O processing reduced for user sync point (IXLUSYNC) event processing.

z/OS I.8 introduces more CRFM enhancements

CFRM Enhancement in zOS 1.8

Reduces recovery time during Structure Rebuild, Duplexing Failover, and Sysplex Partitioning

The CFRM couple data set (CDS)

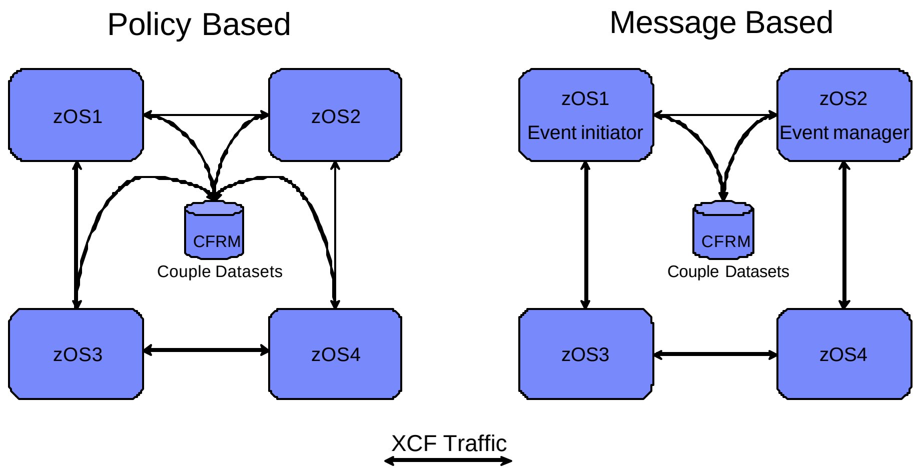
- Centralized “control point”
- Records events like
  - Connects/Disconnects to CF Structures
  - Systems joining or leaving the sysplex

Previously, every system in sysplex updated CFRM CDS.

z/OS 1.8 provides an option (Message Based Protocol)

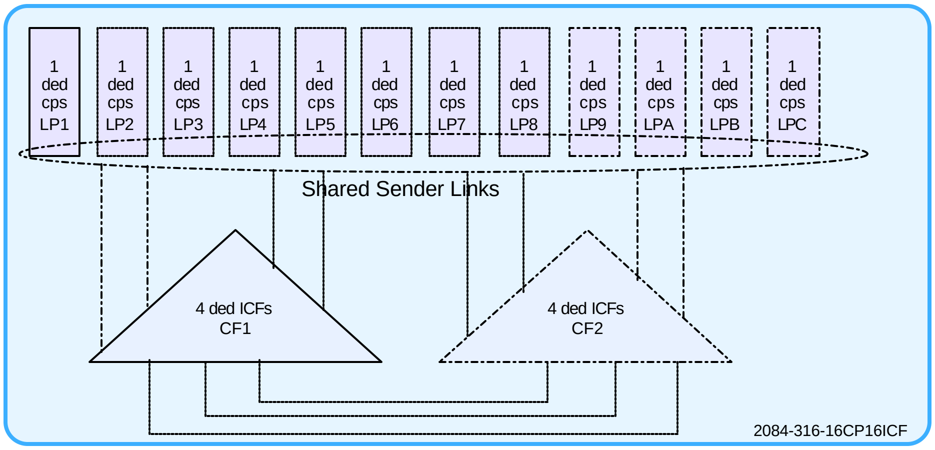
- Reduces accesses to CFRM CDS
- Requires new version of CFRM CDS
- Change with command: SETXCF STOP,MSGBASED

Policy Based Vs Message Based



- Message Based Protocol
  - Only event initiator and manager system do CFRM CDS I/O's
  - XCF signaling is used for communication between manager system and participant systems.

CFRM Test Environment



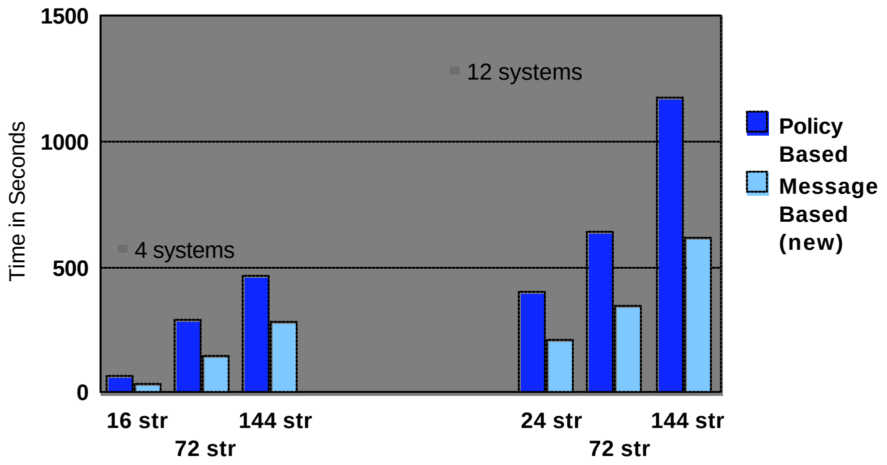
- Structure Rebuild occurs when one CF is taken out of sysplex and all structures rebuilt to other CF

IC links type ICP

ISC3 links type CFP

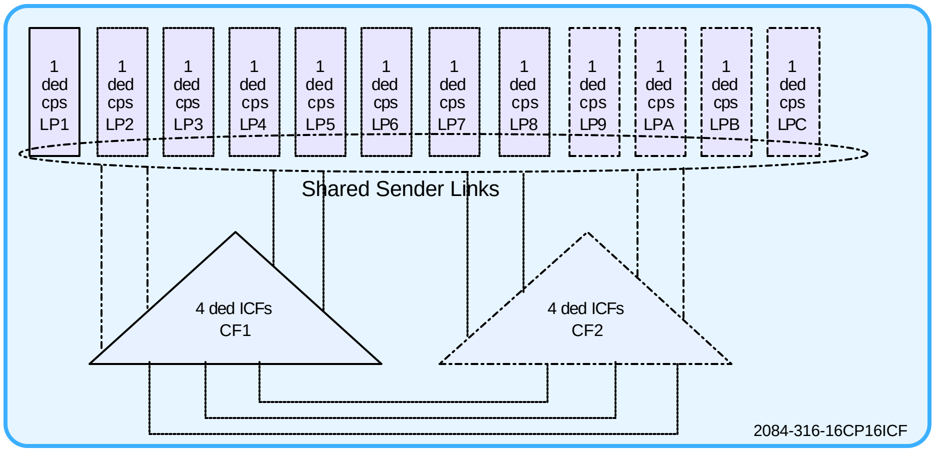
ICB4 links type CBP

Structure Rebuild Improvements



- Rebuild times still increase with number of structures and number of systems but are substantially reduced.

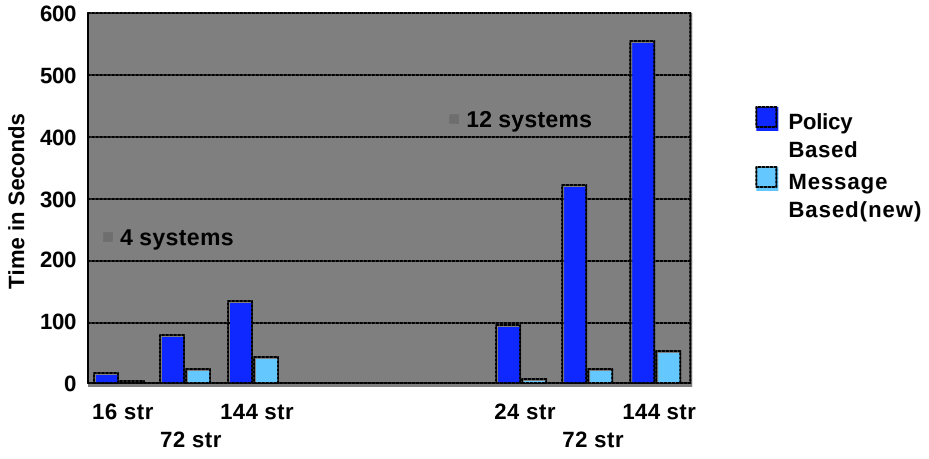
CFRM Test Environment



- Duplexing failover occurs when structure is duplexed and CF fails

IC links type ICP
ISC3 links type CFP
ICB4 links type CBP

Duplexing Failover Improvements



- The failover time is greatly improved and independent of the number of systems connected.

Smaller, more complex configurations

Larger, more expensive processors -

- Consolidation of images
  - A CEC with a single image is very rare today
  - Most have multiple images sharing resources
- Production Sysplexes share resources with
  - Test sysplexes
  - Development sysplexes
  - Recovery sysplexes
- Specialized engines becoming more popular
  - ICFs
  - IFLs
  - zAAPs
  - zIIPs

Functions supporting consolidation

- System managed duplexing
  - Allows CF to reside on same CEC as MVS image
- Separate pools for specialized CPs
  - z9 manages each type of specialize CP in its own pool
  - RMF reports each pool separately, Mon III report added
- IRD
  - Dynamically adjusts CP and IO resources based on importance of work.
  - Used by 40-50% of customers
- Concurrent Configuration changes
  - Concurrent patch apply - allows test CF to have different CFLevel (at least until IML)

CF Configuration Options

Many combinations

1. Standalone CF (ex. 2066 - 0CF, 2084 - 300)
   - Dedicated CPs - best choice for production
   - **Shared CPs**
2. Internal CF (ex. 2064 - 108)
   - Dedicated CPs (expensive - added into S/W license costs)
   - **Dedicated ICFs -**
     - **Not included in S/W license fees**
     - **Can use IC links (fastest)**
     - **Good choice for production, but one potential problem**

MVS using CF on same CEC

IF MVS is actively using structures in a CF on the same CEC, only certain structures should reside in this CF to avoid rebuild problems. See W98029 for full explanation.

- Resource Management structures are good candidates
  - IEFAUTOS
  - XCF
  - GRSSTAR
  - JES CKPT
  - Logger
  - ISTGENERIC
- Data Sharing structures -
  - Some may require time to rebuild - OW33615 improves
    - IMS Cache
    - DB2 Group Buffer pools
    - CICS TEMP STOR
  - Some will cause a sysplex wide subsystem outage
    - DB2 SCA
    - IRLM (IMS & DB2)
    - VSAM RLS lock
    - Logger (CICS & RRS)

Solution - CF duplexing

CF Duplexing establishes two copies of a given structure - changed data is written to both.

This provides:

1. An 'easy-to-implement' recovery mechanism for structures with no recovery
2. Faster recovery from CF failures
3. Failure isolation for internal CFs (ICFs)

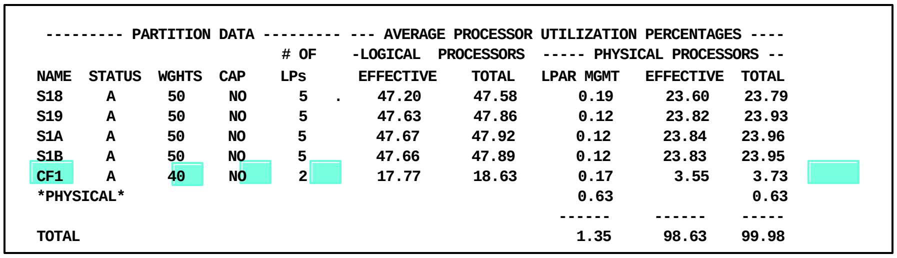
For more information, see

- System-Managed CF Structure Duplexing Implementation Summary
- System-Managed CF Structure Duplexing Implementation

Don't want to assign a whole CP to test CF?

Dynamic CF Dispatching - allows tradeoff between CF response time and CP Utilization

- **At low utilization, CFCC suspended for short periods**
  - **More CP resource for other partitions, but CF requests delayed**
- **As utilization increases, less CFCC suspension**
  - **Less CP resource for other partitions, but faster CF requests**

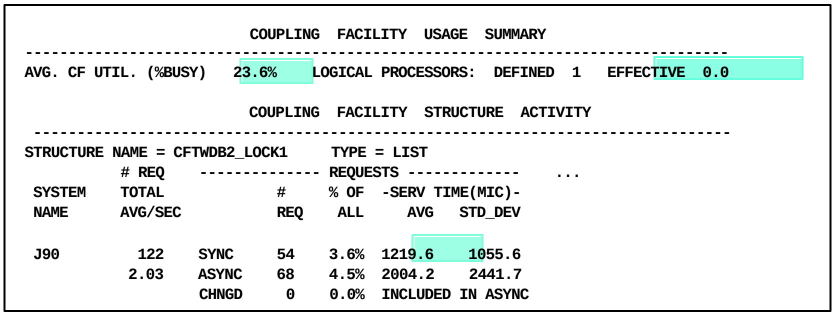


```
 --------- PARTITION DATA --------- --- AVERAGE PROCESSOR UTILIZATION PERCENTAGES ----
                               # OF   -LOGICAL  PROCESSORS  ----- PHYSICAL PROCESSORS --
NAME  STATUS  WGHTS  CAP      LPs      EFFECTIVE    TOTAL   LPAR MGMT   EFFECTIVE  TOTAL
S18     A      50     NO       5   .    47.20       47.58      0.19       23.60     23.79
S19     A      50     NO       5        47.63       47.86      0.12       23.82     23.93
S1A     A      50     NO       5        47.67       47.92      0.12       23.84     23.96
S1B     A      50     NO       5        47.66       47.89      0.12       23.83     23.95
CF1     A      40     NO       2        17.77       18.63      0.17        3.55      3.73
*PHYSICAL*                                                     0.63                  0.63
                                                             ------      ------     -----
TOTAL                                                          1.35       98.63     99.98
```

At low utilization, less CPU resource used but...

Dynamic CF Dispatching

But CF response time increases....

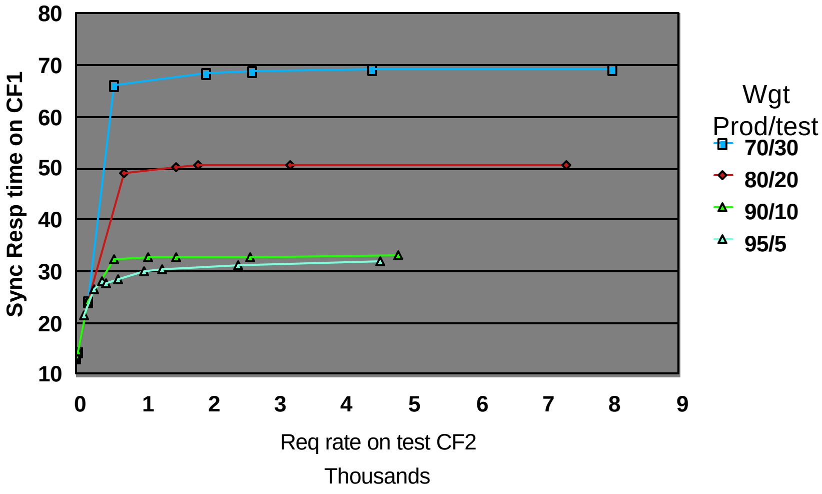


```
                             COUPLING  FACILITY  USAGE  SUMMARY
----------------------------------------------------------------------------------
AVG. CF UTIL. (%BUSY)    23.6%     LOGICAL PROCESSORS:  DEFINED  1   EFFECTIVE  0.0

                             COUPLING  FACILITY  STRUCTURE  ACTIVITY
 ---------------------------------------------------------------------------------
STRUCTURE NAME = CFTWDB2_LOCK1     TYPE = LIST
           # REQ     -------------- REQUESTS -------------    ...
 SYSTEM    TOTAL              #     % OF  -SERV TIME(MIC)-
 NAME      AVG/SEC            REQ    ALL     AVG    STD_DEV

 J90         122     SYNC     54    3.6%  1219.6     1055.6
            2.03     ASYNC    68    4.5%  2004.2     2441.7
                     CHNGD     0    0.0%  INCLUDED IN ASYNC
```

As activity in the test CF partition increases, more CPU resource is used and CF response time improves.

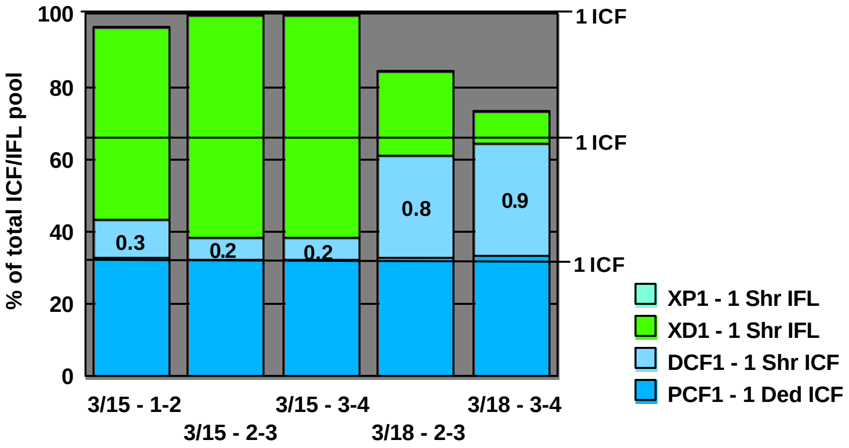
Production CF1 and test CF2 - 1 CP

Dynamic CF Disp is OFF for CF1, ON for CF2
Workload on Production CF1 is constant - 5,500 req/sec

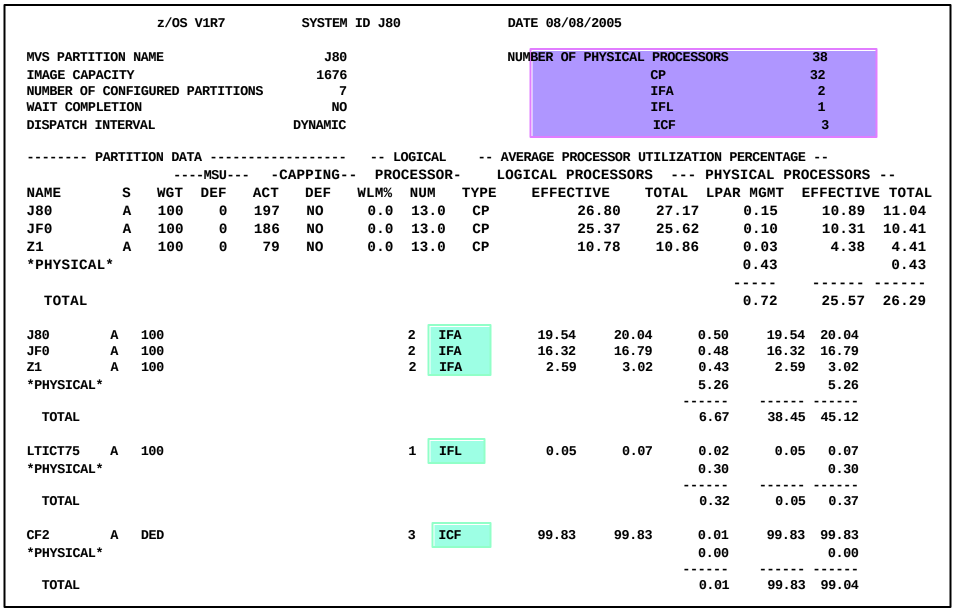


Shared ICFs and IFLs

Potential problem can occur when Specialized Shared CPs are managed as a single pool



Specialized Processors on z9



```
                    z/OS V1R7          SYSTEM ID J80              DATE 08/08/2005

MVS PARTITION NAME                        J80                NUMBER OF PHYSICAL PROCESSORS           38
IMAGE CAPACITY                           1676                                CP                     32
NUMBER OF CONFIGURED PARTITIONS             7                                IFA                     2
WAIT COMPLETION                            NO                                IFL                     1
DISPATCH INTERVAL                     DYNAMIC                                ICF                     3

-------- PARTITION DATA -----------------   -- LOGICAL    -- AVERAGE PROCESSOR UTILIZATION PERCENTAGE --
                   ----MSU---   -CAPPING--  PROCESSOR-    LOGICAL PROCESSORS   --- PHYSICAL PROCESSORS --
NAME         S   WGT  DEF   ACT   DEF   WLM%  NUM   TYPE    EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE TOTAL
J80          A   100    0   197   NO     0.0  13.0   CP         26.80    27.17     0.15      10.89  11.04
JF0          A   100    0   186   NO     0.0  13.0   CP         25.37    25.62     0.10      10.31  10.41
Z1           A   100    0    79   NO     0.0  13.0   CP         10.78    10.86     0.03       4.38   4.41
*PHYSICAL*                                                                         0.43              0.43
                                                                                  -----     ------ ------
  TOTAL                                                                            0.72      25.57  26.29

J80         A   100                              2  IFA      19.54     20.04     0.50     19.54  20.04
JF0         A   100                              2  IFA      16.32     16.79     0.48     16.32  16.79
Z1          A   100                              2  IFA       2.59      3.02     0.43      2.59   3.02
*PHYSICAL*                                                                       5.26             5.26
                                                                               ------    ------ ------
  TOTAL                                                                          6.67     38.45  45.12

LTICT75     A   100                              1  IFL       0.05      0.07     0.02      0.05   0.07
*PHYSICAL*                                                                       0.30             0.30
                                                                               ------    ------ ------
  TOTAL                                                                          0.32      0.05   0.37

CF2         A   DED                              3  ICF      99.83     99.83     0.01     99.83  99.83
*PHYSICAL*                                                                       0.00             0.00
                                                                               ------    ------ ------
  TOTAL                                                                          0.01     99.83  99.04
```

Benefits of IRD

At very low cost, LPAR clustering improves systems management by managing:

1. CPU resources
   - A. **Dynamic** distribution of capacity within LPAR cluster while protecting capacity outside LPAR cluster
   - B. Improves efficiency
   - C. Uses upgraded capacity immediately

2. IO resources
   - A. Prioritizes work when I/O is constrained (CSSQ)
   - B. Improves channel configuration efficiency (DCM)

Most useful when multiple images are consolidated on one CEC and/or workloads change dynamically.

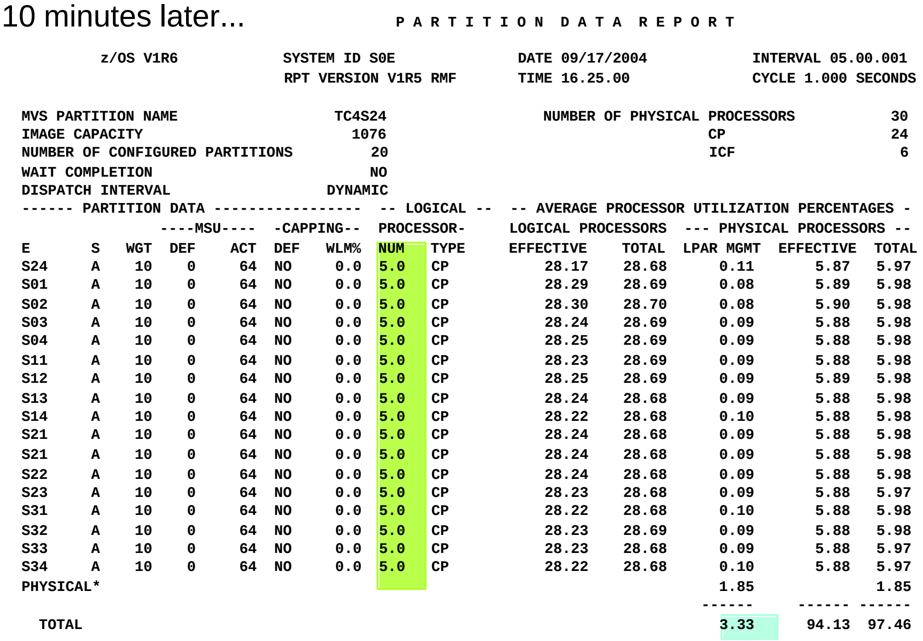
Example - IRD adjusting Logical CPs

Early in IPL...

```
                                  P A R T I T I O N   D A T A   R E P O R T

         z/OS V1R6              SYSTEM ID S0E          DATE 09/17/2004          INTERVAL 04.32.931
                                RPT VERSION V1R5 RMF   TIME 16.15.27            CYCLE 1.000 SECONDS

MVS PARTITION NAME                 TC4S24                 NUMBER OF PHYSICAL PROCESSORS           30
IMAGE CAPACITY                       1076                                CP                       24
NUMBER OF CONFIGURED PARTITIONS        20                                ICF                       6
WAIT COMPLETION                        NO
DISPATCH INTERVAL                 DYNAMIC
------ PARTITION DATA -----------------  -- LOGICAL P  -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES --
                ----MSU----  -CAPPING--  PROCESSOR-    LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS ---
NAME       S   WGT  DEF    ACT  DEF   WLM%  NUM   TYPE    EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL
TDCS24     A    10    0     63  NO     0.0 11.0   CP          12.25    12.70       0.21       5.61   5.82
TDCS01     A    10    0     62  NO     0.0 11.8   CP          11.34    11.75       0.20       5.57   5.77
...12 similar images
TDCS33     A    10    0     62  NO     0.0 11.7   CP          11.46    11.89       0.21       5.59   5.80
TDCS34     A    10    0     63  NO     0.0 11.7   CP          11.46    11.93       0.23       5.58   5.81
  PHYSICAL*                                                                        5.13              5.13
                                                                                 ------     ------ ------
  TOTAL                                                                            8.46      89.41  97.87
                                    L P A R   C L U S T E R   R E P O R T

                            ------ WEIGHTING STATISTICS ------   ---- PROCESSOR STATISTICS ----
                           --- DEFINED ---  ---- ACTUAL -----   ---- NUMBER ---   -- TOTAL% --
CLUSTER   PARTITION  SYSTEM    INIT  MIN  MAX  AVG  MIN %  MAX %   DEFINED  ACTUAL    LBUSY  PBUSY
ENGTEST1  TC4S01     S00         10    0    0   10      -      -        24    11.8    11.75   5.77
          TC4S02     S04         10    0    0   10      -      -        24    11.4    12.14   5.78
          ... 12 similar images
          TC4S33     S0B         10    0    0   10      -      -        24    11.7    11.89   5.80
          TC4S34     S0F         10    0    0   10      -      -        24    11.7    11.93   5.81
---------------------------   ---------------------------------   ------------------------------
                      TOTAL     160                                    384            193.3  92.74
```

IRD example - Cont.

10 minutes later...



```
                                   P A R T I T I O N   D A T A   R E P O R T

          z/OS V1R6            SYSTEM ID S0E            DATE 09/17/2004           INTERVAL 05.00.001
                               RPT VERSION V1R5 RMF     TIME 16.25.00             CYCLE 1.000 SECONDS

MVS PARTITION NAME                   TC4S24             NUMBER OF PHYSICAL PROCESSORS            30
IMAGE CAPACITY                         1076                                         CP           24
NUMBER OF CONFIGURED PARTITIONS          20                                         ICF           6
WAIT COMPLETION                          NO
DISPATCH INTERVAL                   DYNAMIC
------ PARTITION DATA -----------------  -- LOGICAL --  -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES -
                ----MSU----  -CAPPING--  PROCESSOR-     LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS --
E       S   WGT  DEF     ACT  DEF   WLM%  NUM    TYPE    EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL
S24     A    10    0      64  NO     0.0  5.0    CP          28.17    28.68       0.11       5.87   5.97
S01     A    10    0      64  NO     0.0  5.0    CP          28.29    28.69       0.08       5.89   5.98
S02     A    10    0      64  NO     0.0  5.0    CP          28.30    28.70       0.08       5.90   5.98
S03     A    10    0      64  NO     0.0  5.0    CP          28.24    28.69       0.09       5.88   5.98
S04     A    10    0      64  NO     0.0  5.0    CP          28.25    28.69       0.09       5.88   5.98
S11     A    10    0      64  NO     0.0  5.0    CP          28.23    28.69       0.09       5.88   5.98
S12     A    10    0      64  NO     0.0  5.0    CP          28.25    28.69       0.09       5.89   5.98
S13     A    10    0      64  NO     0.0  5.0    CP          28.24    28.68       0.09       5.88   5.98
S14     A    10    0      64  NO     0.0  5.0    CP          28.22    28.68       0.10       5.88   5.98
S21     A    10    0      64  NO     0.0  5.0    CP          28.24    28.68       0.09       5.88   5.98
S21     A    10    0      64  NO     0.0  5.0    CP          28.24    28.68       0.09       5.88   5.98
S22     A    10    0      64  NO     0.0  5.0    CP          28.24    28.68       0.09       5.88   5.98
S23     A    10    0      64  NO     0.0  5.0    CP          28.23    28.68       0.09       5.88   5.97
S31     A    10    0      64  NO     0.0  5.0    CP          28.22    28.68       0.10       5.88   5.98
S32     A    10    0      64  NO     0.0  5.0    CP          28.23    28.69       0.09       5.88   5.98
S33     A    10    0      64  NO     0.0  5.0    CP          28.23    28.68       0.09       5.88   5.97
S34     A    10    0      64  NO     0.0  5.0    CP          28.22    28.68       0.10       5.88   5.97
PHYSICAL*                                                                     1.85              1.85
                                                                            ------     ------ ------
  TOTAL                                                                       3.33      94.13  97.46
```

CFCC Concurrent Patch Apply

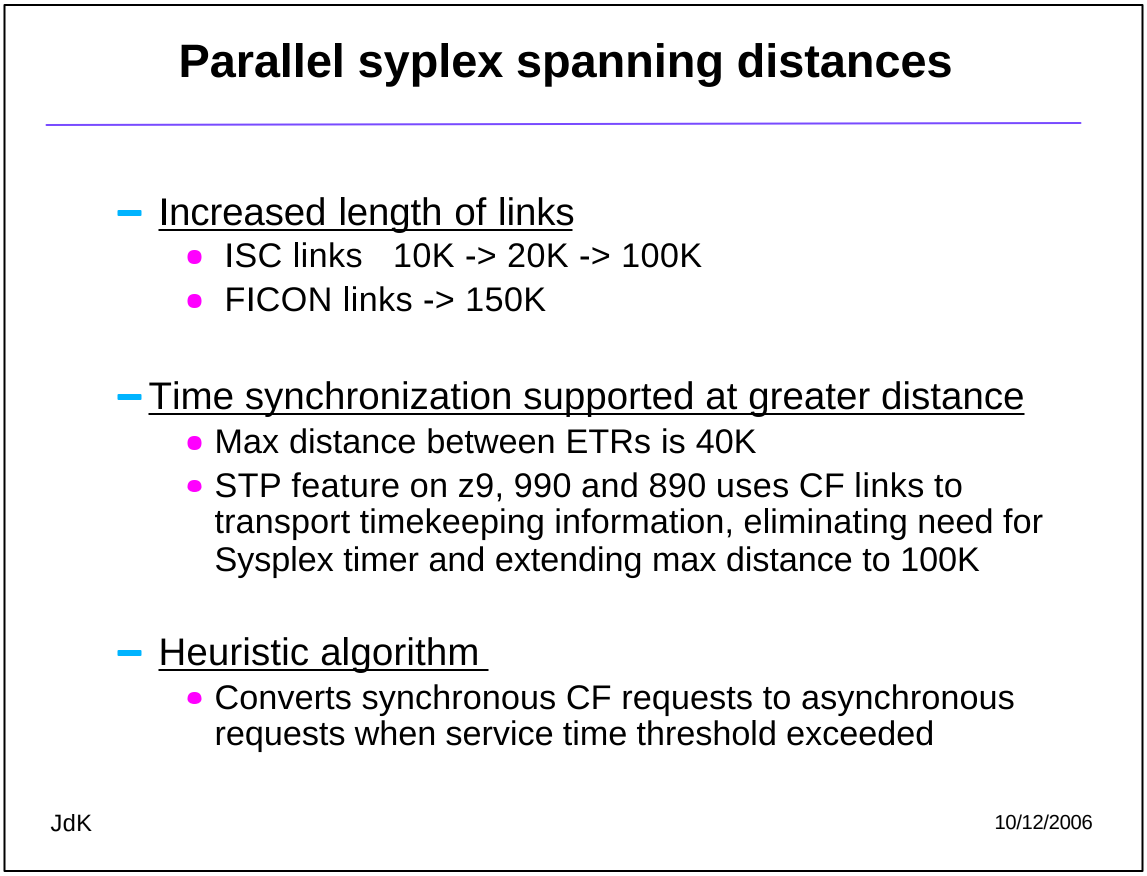
CFCC Enhanced Patch Apply

- available on z890, z990 and z9
- allows disruptive install of new CFCC code (no POR) on a test CF without changing production CF image.
  - Allows different CFLevels on the same CEC
  - Note: Any activation or reactivation of a CF image will pick up newest version of CFCC.

Reminder: If CF images are sharing CPs/ICFs

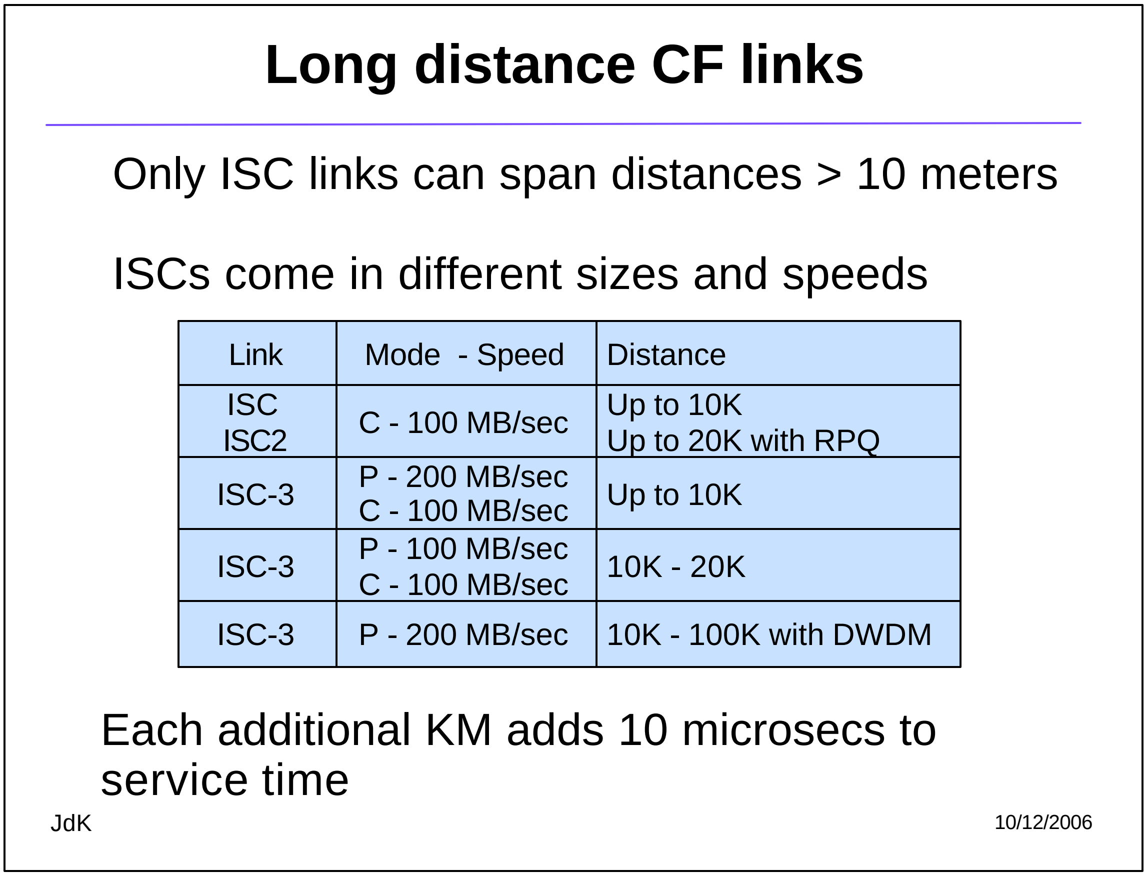
Test CF Image - DYNDISP=ON

Production CF Image - DYNDISP=OFF



Parallel syplex spanning distances

- Increased length of links
  - ISC links 10K -> 20K -> 100K
  - FICON links -> 150K
- Time synchronization supported at greater distance
  - Max distance between ETRs is 40K
  - STP feature on z9, 990 and 890 uses CF links to transport timekeeping information, eliminating need for Sysplex timer and extending max distance to 100K
- Heuristic algorithm
  - Converts synchronous CF requests to asynchronous requests when service time threshold exceeded



Long distance CF links

Only ISC links can span distances > 10 meters

ISCs come in different sizes and speeds

Link	Mode - Speed	Distance
ISC ISC2	C - 100 MB/sec	Up to 10K Up to 20K with RPQ
ISC-3	P - 200 MB/sec C - 100 MB/sec	Up to 10K
ISC-3	P - 100 MB/sec C - 100 MB/sec	10K - 20K
ISC-3	P - 200 MB/sec	10K - 100K with DWDM

Each additional KM adds 10 microsecs to service time

DWDM

Dense Wave Division Multiplexer

- Uses optical multiplexing technique to increase the carrying capacity of a fiber network beyond what can currently be accomplished by time division multiplexing (TDM) techniques.
- Different wavelengths of light are used to transmit multiple streams of information along a single fiber
- One pair can handle all connectivity - DASD and CF

Estimating Additional subchannels

Subchannel utilization can be calculated as

$$\frac{(\text{Sync Rate}*\text{Sync serv.time}) + (\text{Async Rate}*\text{Async Serv time})}{\#\text{Subchannels}}\ 100$$

```
                   # REQ                                   ----------- REQUESTS --
SYSTEM    TOTAL    -- CF LINKS --   PTH                    #    -SERVICE TIME(MIC)-
NAME      AVG/SEC TYPE   GEN   USE   BUSY                  REQ      AVG     STD_DEV

FA         16161K CFP      2     2      0   SYNC       5481K    35.3        28.1
           53869  SUBCH   28    14          ASYNC     10696K   150.6       133.0
```

$$\text{Sync rate} = \frac{5{,}481{,}000}{300} = 18{,}270 \qquad \text{Async rate} = \frac{10{,}696{,}000}{300} = 35{,}653$$

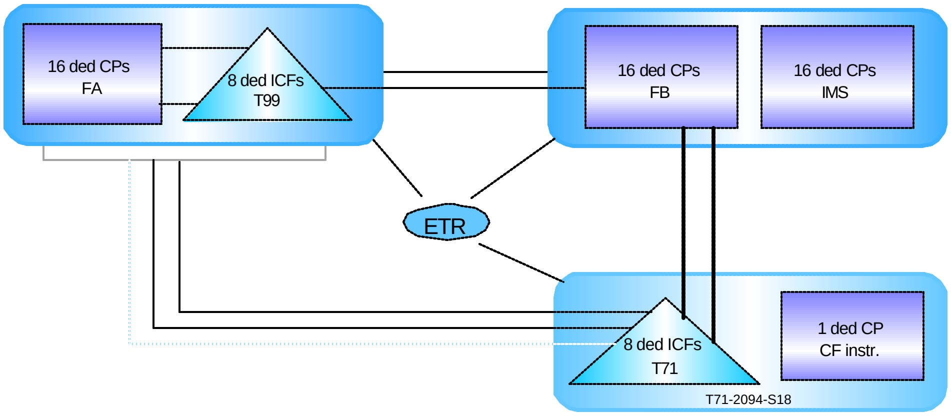
$$\text{Util} = \frac{(18{,}270 * .000035 + 35{,}653 * .000151)}{14} * 100 = 43\%$$

With 100K links, service times increase by 1000 usecs

$$\text{Util} = \frac{(18{,}270 * .001035 + 35{,}653 * .001151)}{14} * 100 = 350\%$$

To keep 43% subchannel utilization, would need 20 links

Test config - ETR - short



All 100K Links use one common DWDM pair

ISC (Long) links type CFP

ICB4 links type CBP

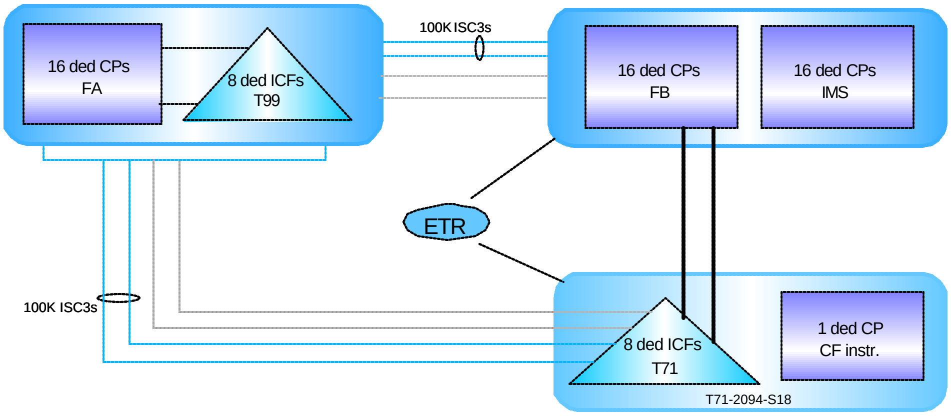
ISC links type CFP

IC Links

JdK

10/12/2006

Test config - Mixed - short and long



All 100K Links use one common DWDM pair

ISC (Long) links type CFP

ICB4 links type CBP

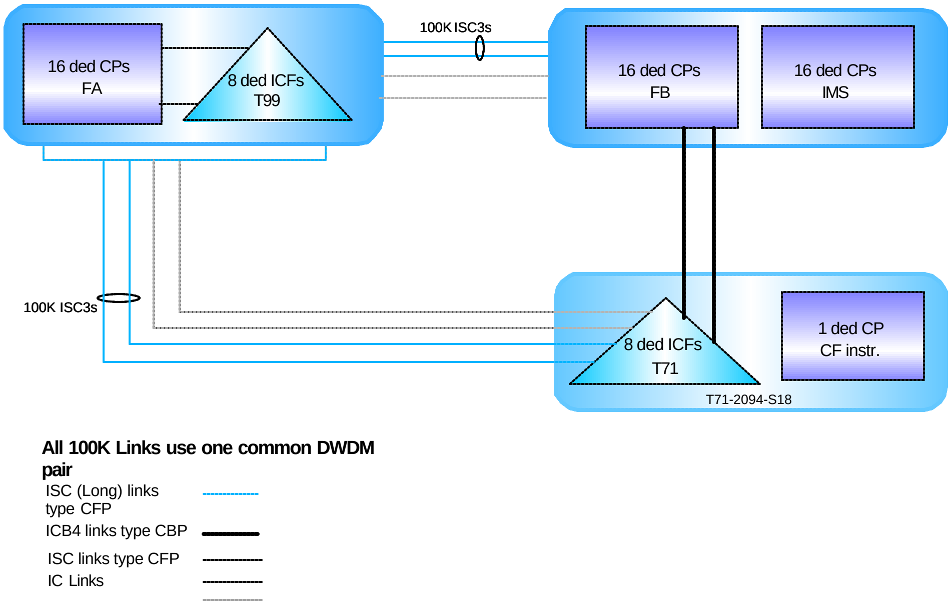
ISC links type CFP

IC Links

JdK

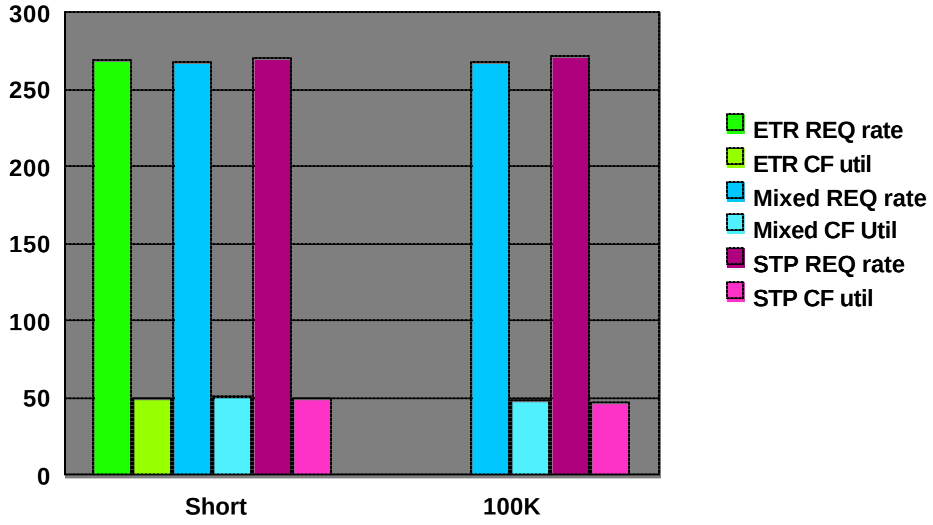
10/12/2006

Test config - STP only



All 100K Links use one common DWDM pair

ETR - Mixed - STP Comparison



Method of synchronization has unobservable effect on throughput or CF utilization

Simplex Comparison - Mixed

- Service times increase as expected

  100K * 10usec/K = 1000 usecs

T71	FA - 2 ISC			FB - 2 ICB		
	Serv Time	# Sync	# Async	Serv Time	# Sync	# Async
Short	38	23%	77%	18	95%	5%
100K	1046	1%	99%	17	95%	5%

FA - 2 IC			T99	FB- 2 ISC		
Serv Time	# Sync	# Async		Serv Time	# Sync	# Async
12	99%	1%	Short	42	14%	86%
11	98%	2%	100K	1047	1%	99%

- More SYNCs converted to ASYNCs
- Total request rate drops/less requests to distant CF

Heuristic Algorithm

Long running SYNC CF requests use more CPU on sender.

Prior to z/OS1.2, XES changed some LIST/CACHE SYNC requests to ASYNC based on preset rules. Factors included

1. Request type
2. Sender and receiver processor type
3. Amount of data being sent

In z/OS 1.2, CF response time for SYNC requests is monitored for every request type (LIST/LOCK/CACHE) and compared to threshold so all/only long requests (for whatever reason) are converted.

- Thresholds are based on SYNC and ASYNC pathlenghs for various requests types - LIST, LOCK, CACHE, Simplex, Duplex.
- When SYNC pathlength plus cycles spent waiting for a response is greater than ASYNC pathlengh, request is converted to ASYNC

Heuristic Algorithm, cont.

Requests which are changed from SYNC to ASYNC based on the Heuristic Algorithm are counted as ASYNC

- not included in the CHNGD counts

Thresholds are normalized by processor type - (cycles spent waiting for response varies with speed of processor)

Thresholds are not externally adjustable

- ► OW51813 for the latest threshold adjustment

The decision to convert SYNC to ASYNC is continuously reevaluated by allowing every nth SYNC request to be issued unchanged and comparing it with the thresholds.

Value of Sync => Async heuristic

Heuristic algorithm tries to limit the impact of

- ► DISTANCE
- ► Technology mismatch
- ► High CF utilization

- Benchmark results
  - CICS/DB2 data sharing workload
  - z900 host and CF technology

Distance between CFs	Cost of d.s. pre z/OS 1.2	Cost of d.s. z/OS 1.2
5 m	10%	10%
10 km	20%	14%

Simplex Comparison - Path Busy

Observed a few PTH BUSY conditions on long links

T71 # ,	FA - 2ISCs	Shr w. CF		FB - 2 ICB		
	ETR	Mixed	STP	ETR	Mixed	STP
Short	.00%	.00%	.00%	.00%	.00%	.00%
100K	-	.05%	.00%	.00%	.00%	.00%

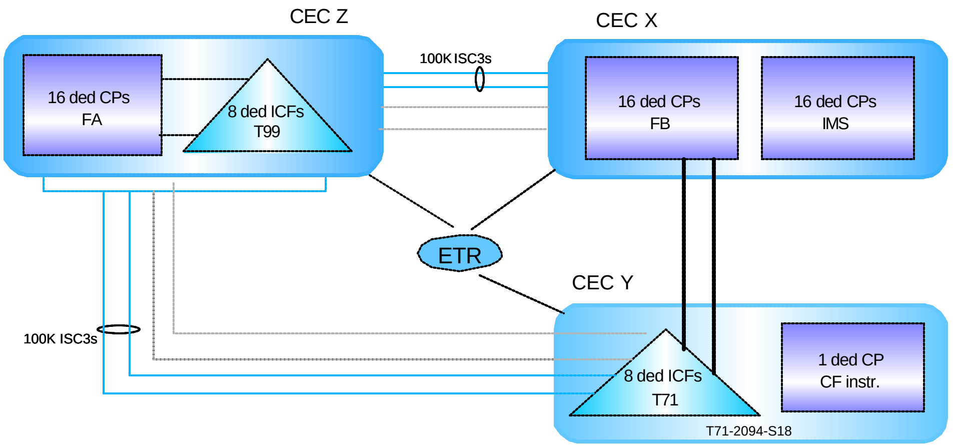
FA - 2 IC			T99	FB- 2 ISC	Ded	
ETR	Mixed	STP		ETR	Mixed	STP
.00%	.00%	.00%	Short	.00%	.00%	.00%
.00%	.00%	.00%	100K	-	.14%	.21%

- ETR - CF sends "health signals" when connected to CF
- STP - Timing signals sent every 64 msecs could be using the link. Occupy link longer at greater distances.

Additional Information

- Websites www.ibm.com/servers/eserver/zseries
  - Parallel sysplex (CF sizer, CFLevel description) .. /pso
    - System-Managed CF Structure Duplexing Implementation Summary
    - System-Managed CF Structure Duplexing
  - FICON - http://www.ibm.com/servers/eserver/zseries/connectivity/
  - DWDMs - http://www.redbooks.ibm.com/abstracts/tips0058.html?Open
- WSC FLASHs www.ibm.com/support/techdocs
  - Flash10159 New Heuristic Algorithm for CF Request Conversion
  - Flash10337 z/OS CPENABLE Settings IBM 9672 / zSeries Processor
  - WP100743 Parallel Sysplex Performance: XCF Performance V3.1
- Publications
  - Setting up a Sysplex (SA22-7625-06)
  - z/Series 900 System Overview (SA22-1027-03b)
  - z/Series 990 System Overview (SA22-1032-00a)
  - System z9 Enterprise Class Overview (SA22- 6833-02a)
  - Processor Resource/System Manager Planning Guide (SB10-7033-05)

Test config - ETR, Mixed, STP - short/ long



All 100K Links use one common DWDM pair

ISC (Long) links type CFP

ICB4 links type CBP

ISC links type CFP

IC Links

Mode 1: CEC X, Y and Z stepping up to ETR

Mode 2: CEC X and Y to ETR
CEC Z in STP getting time from X and Y

Mode 3: CEC X, Y and Z in STP mode