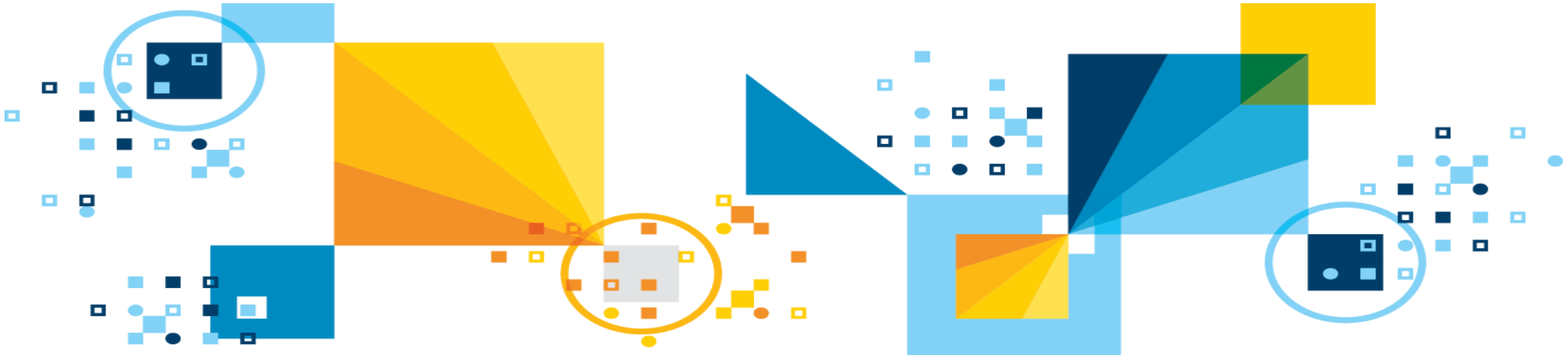
Db2 12 for z/OS and Asynchronous Lock Structure Duplexing: Update

- **Tridex**
- Mark Rader, Db2 for z/OS
- IBM Washington Systems Center
- *September 19, 2019*



Session Objectives

- Review CF structure duplexing
- System managed duplexing of Db2 lock structure (LOCK1)
- Asynchronous system managed duplexing of LOCK1
- Customer experiences
- Summary and request...

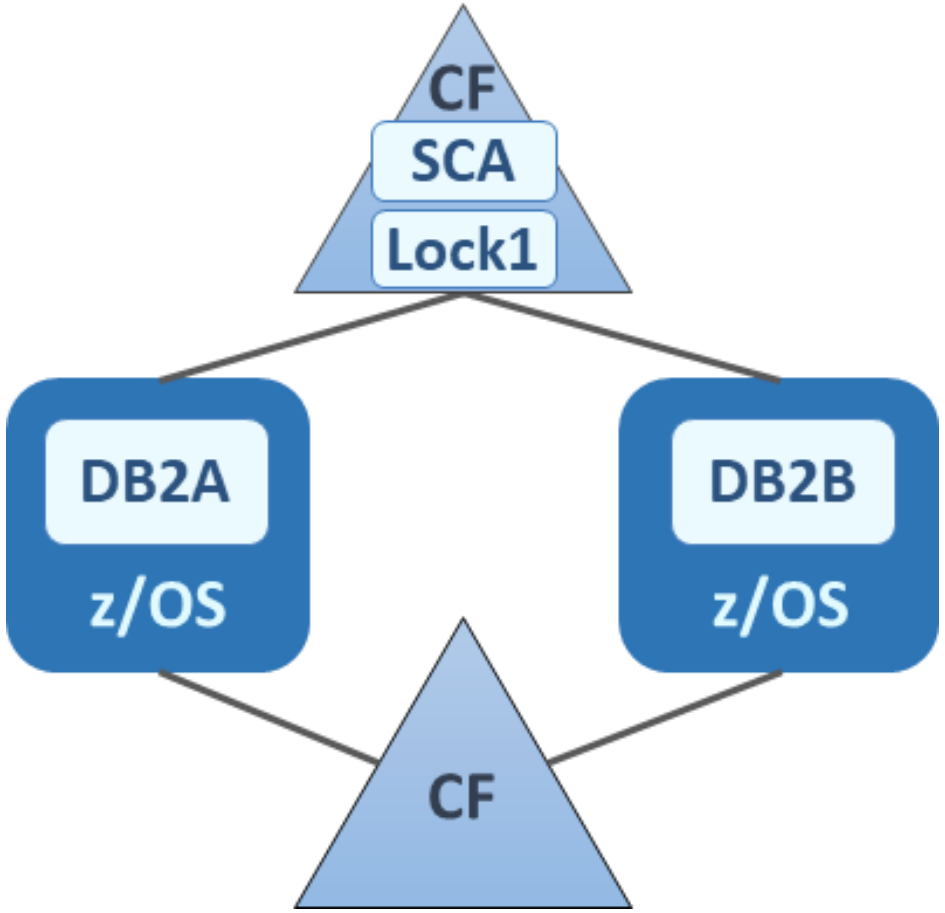
Parallel Sysplex basics for Db2 for z/OS

- Multiple Db2 members in a Db2 data sharing group
- Coupling facility (CF) structures used for high speed sharing of information about locks, status, data (tables, indexes)
- *ssnm*IRLM allocates lock structure: *dsngrpnm*_LOCK1 (LOCK1)
- *ssnm*MSTR allocates shared communication area: *dsngrpnm*_SCA (SCA)
- *ssnm*DBM1 allocates group buffer pool for each local BP with shared data
  - *dsngrmnm*_GBPn (GBP0, GBP1, GBP2, ...GBP8K0, ..GBP16K0, ...GBP32K)

- LOCK1 and SCA are required structures
  - If Db2 cannot allocate them, Db2 fails

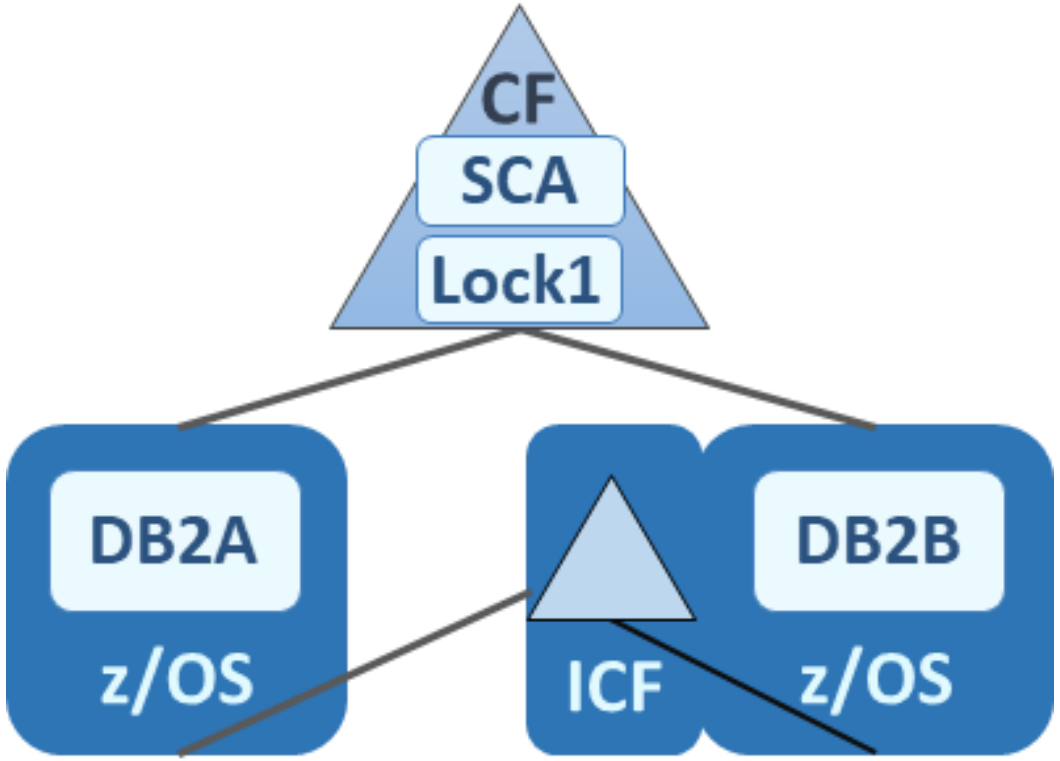
‘Original’ Parallel Sysplex configuration

- 2 external Coupling Facilities (CFs)
- SCA and LOCK1 isolated from MSTR and IRLM
- GBPs spread across CFs
- Failure of any single CEC tolerated
  - Structures rebuilt to other CF by connectors
  - Db2 members restarted on other LPAR to release retained locks



Second PSX configuration

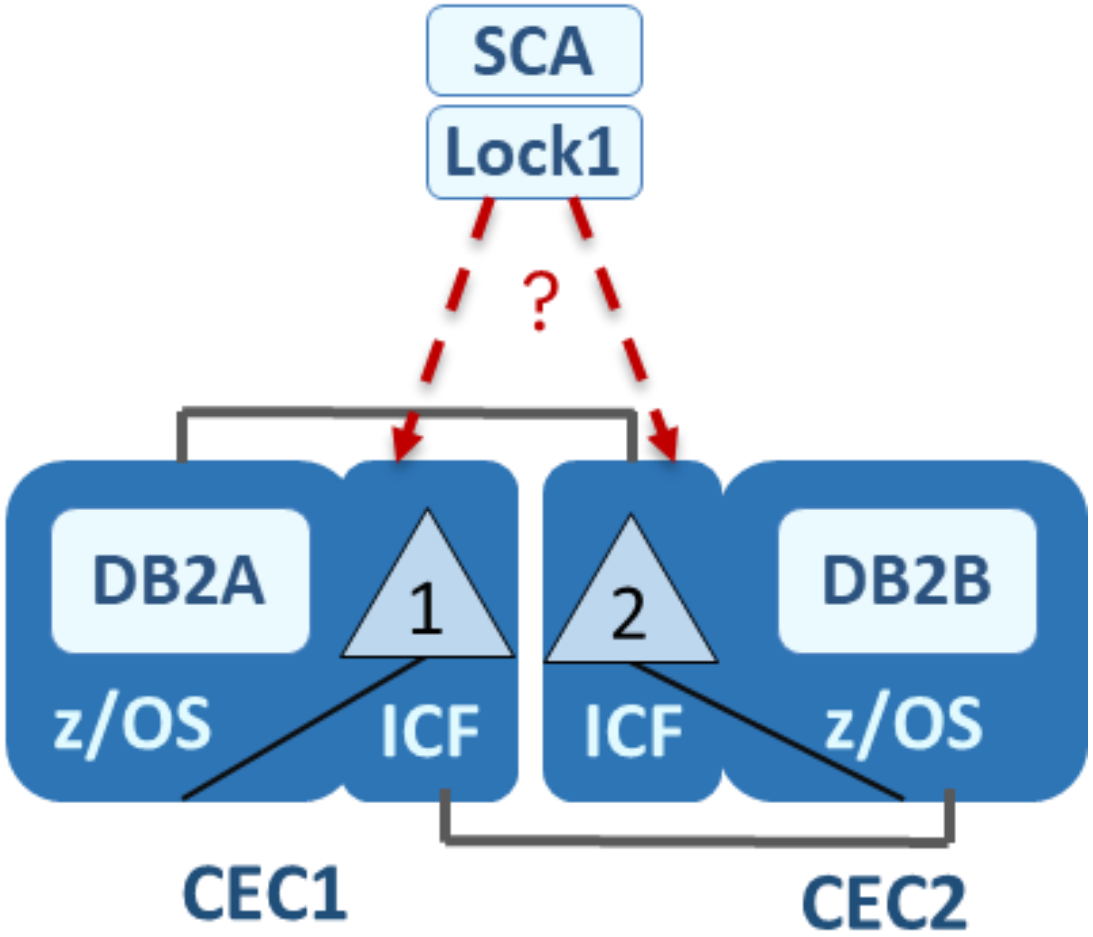
- 1 external CF
- 1 integrated CF (ICF)
- 3 CECs
- SCA and LOCK1 isolated from MSTR and IRLM
- GBPs spread across CFs
  - Duplexed GBPs
  - Primary GBPs spread across CFs
    - Secondary GBP on other CF



Third PSX configuration

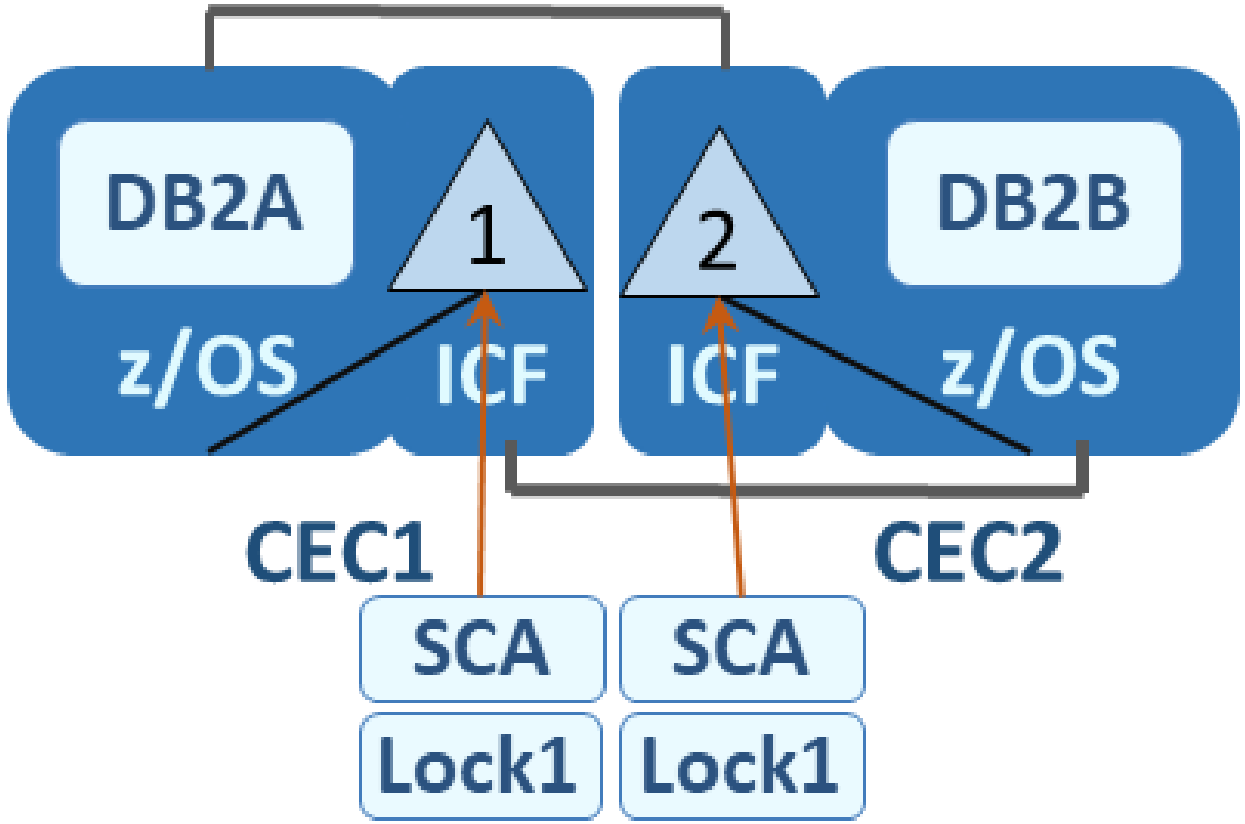
- 2 integrated CFs, 2 CECs
- SCA and LOCK1 not isolated from MSTR and IRLM
- Duplexed GBPs spread across CFs

- What if one CEC is lost?
- It depends…
- Could be loss if 1 Db2 member
- Could be loss of whole Db2 data sharing group
  - Because loss of SCA or LOCK1 ***and*** Db2 means the SCA or LOCK1 cannot be rebuilt



Third PSX configuration: recommended

- System Managed Duplexing
  - 2 integrated CFs, 2 CECs
  - SCA and LOCK1 duplexed (synchronous)
- User Managed Duplexing
  - Duplexed GBPs spread across CFs
- If either CEC fails
  - Db2 on other CEC available
  - Restart failed Db2 on other LPAR
    - Release retained locks
- Increased cost of lock requests
  - Performance and CPU costs
  - ***New with Db2 12: Async CF duplexing mitigates cost and performance***
    - Focus of the rest of this material

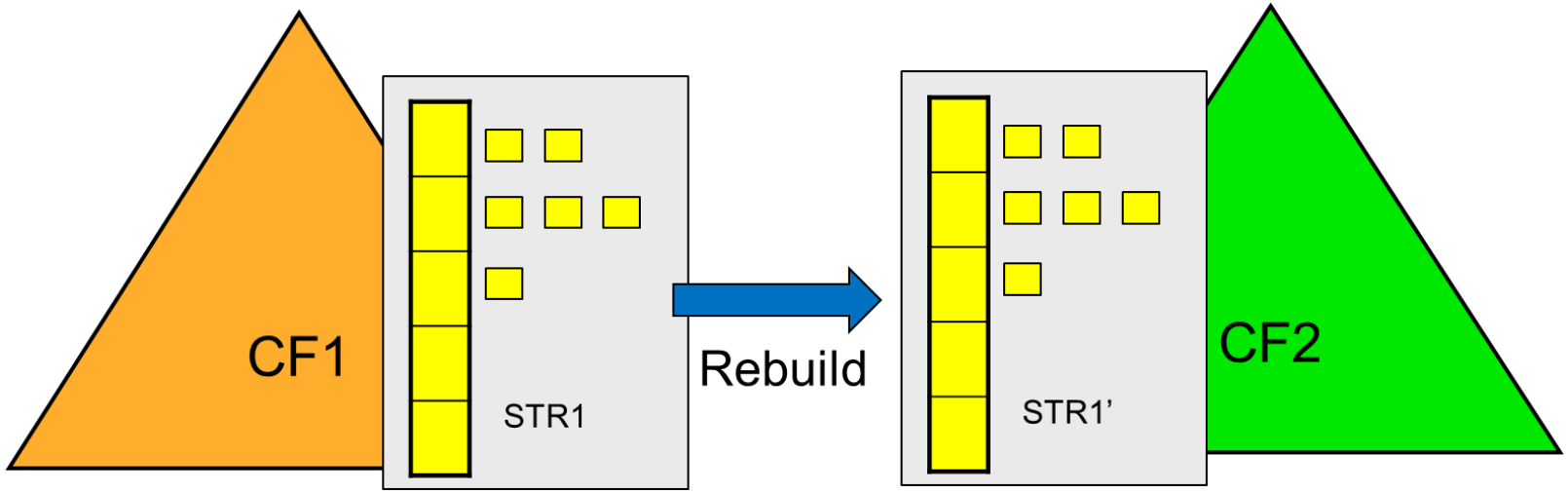


CF structure rebuild

— **Rebuild**
  - Process by which sysplex allocates new instance of CF structure, populates the structure with data, and proceeds to use the new structure instance

— Simplex rebuild
  - Discards the old instance, uses the new

— Duplex rebuild
  - Uses both instances, synchronizing updates so the structure contents remain the same



Two ways to accomplish the rebuild

User-managed

- Can be tailored to application
- Significant programming effort
  - Exploiter must coordinate process and propagate data
- Must be at least one active connector to do the work
- Variety of ways to propagate data to new structure
  - Copy from old structure
  - Reconstruct from in-storage data
  - Reconstruct from DASD and logs
  - Start fresh with empty structure

System-managed

- General purpose
- Virtually no development cost
  - System coordinates process and propagates data
- System does the work; can rebuild even if no connectors
- To propagate data to new structure, the system must have access to the old structure instance

Structure rebuild use cases

User-managed rebuild

- Planned reconfiguration
- Failure recovery

System-managed rebuild

- Planned reconfiguration
- *Limited/no use for failure recovery (must copy old structure)*

User-managed duplex

- Only for cache structures
  - DB2 group buffer pools
- Improved availability

System-managed duplex

- Robust failure recovery

Cost factor estimates for synchronous System Managed duplexing

- Percentage of requests that get duplexed
  - Cache 20% is typical (ranges 1% to 100%)
  - List 100% (or close to it)
  - Lock 100%

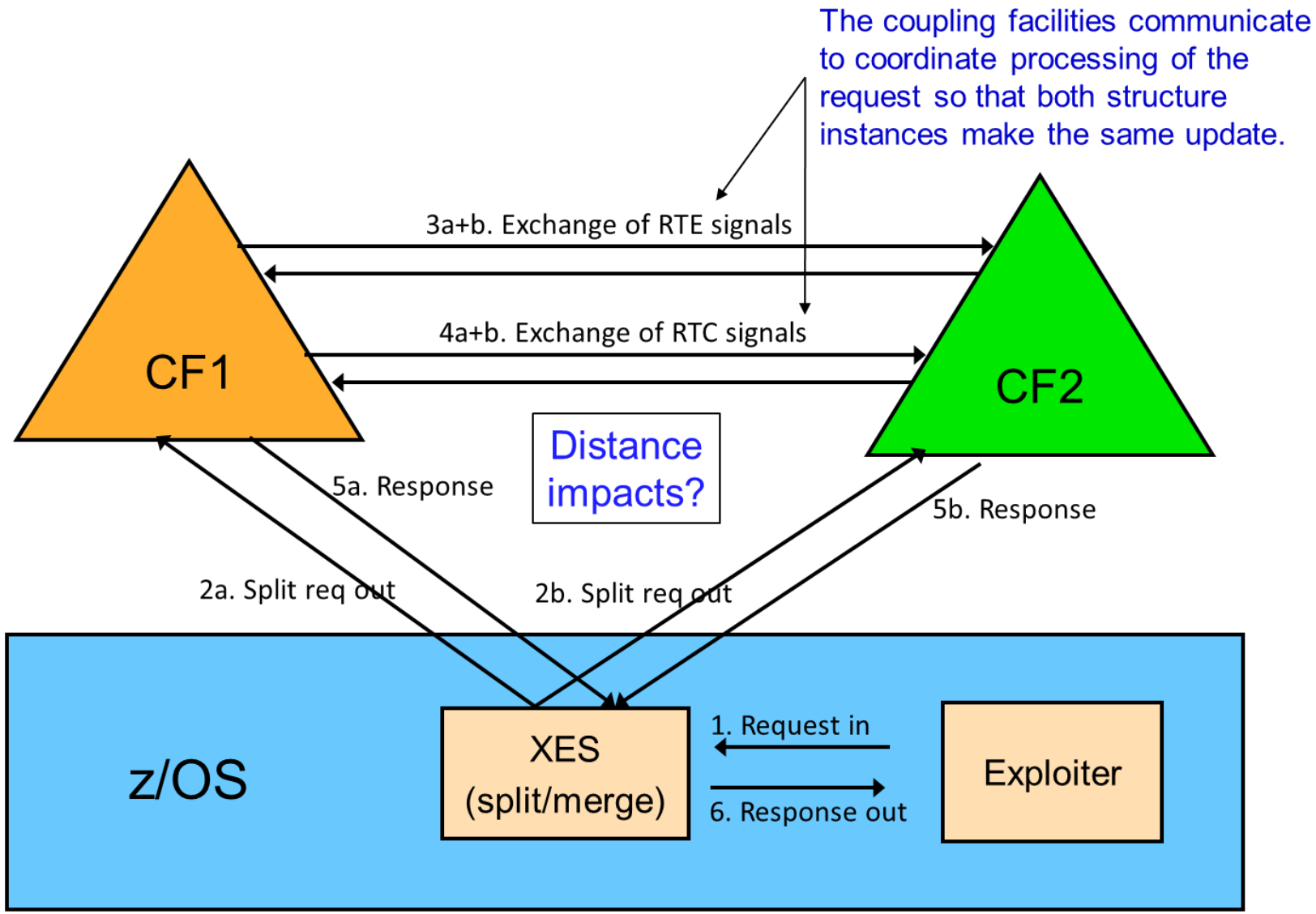
Examples:
[not Db2 GBPs]
Db2 SCA
Db2 LOCK1

- Cost of duplexed request vs. simplex request
  - z/OS CPU = 3x to 4x
  - CF CPU = 4x to 5x
  - CF Link = 6x to 8x

Highly visible

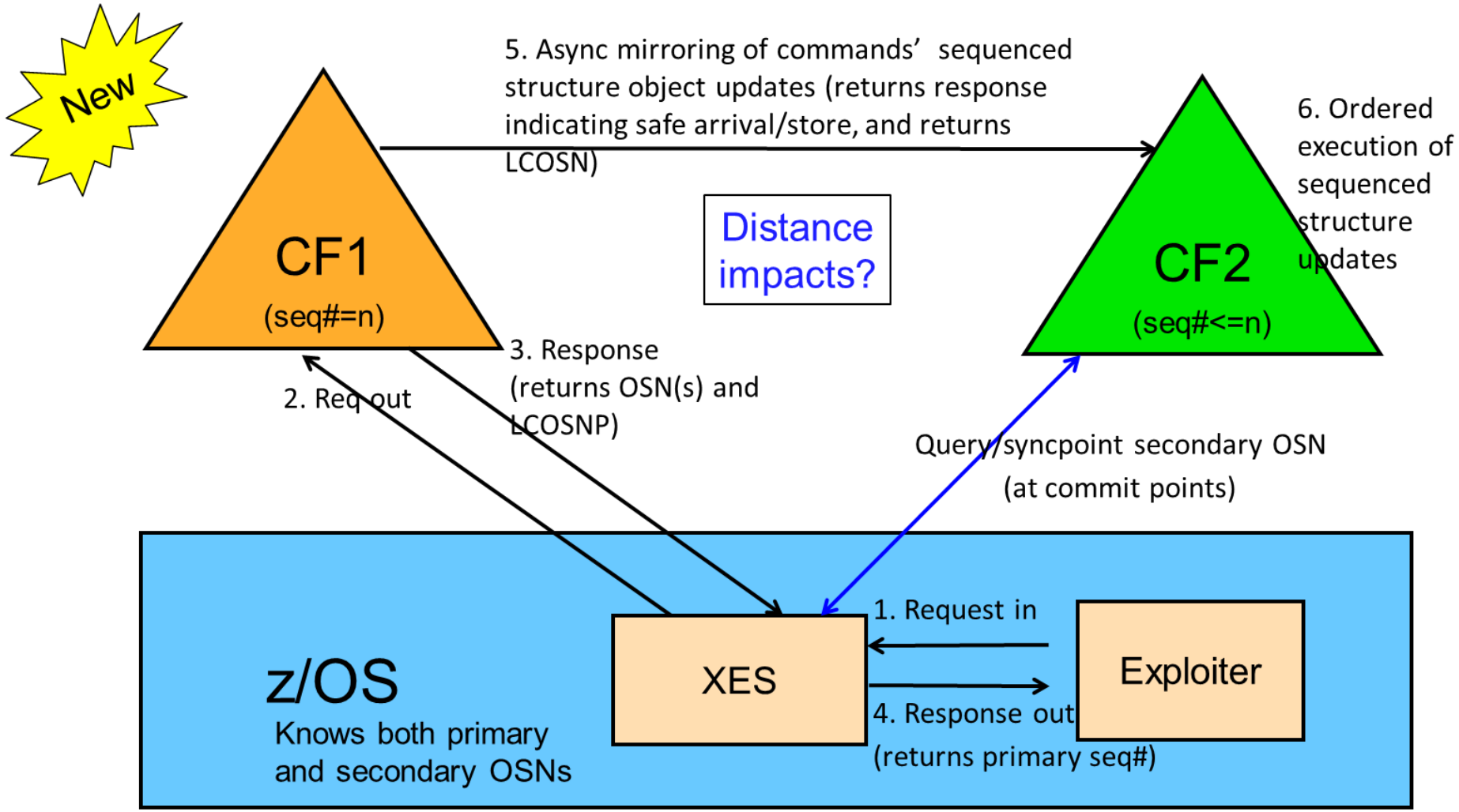
- Total impact to system would depend on particular structures, request frequency, read/write ratios
- Typically significant increase in request response time relative to simplex (further magnified by distance)
  - 2nd order effects on workload difficult to predict
    - Queuing
    - Contention
    - Timeouts

Synchronous System Managed CF structure duplexing



RTE – ready to execute
RTC – ready to complete

Asynchronous System Managed CF structure duplexing



Values known to each system are likely different.

OSN = Operation Sequence Number
LCOSN = Last OSN completed by secondary
LCOSNP = LCOSN known to primary

Asynchronously duplexed structure may require “sync up” since secondary instance normally lags the primary

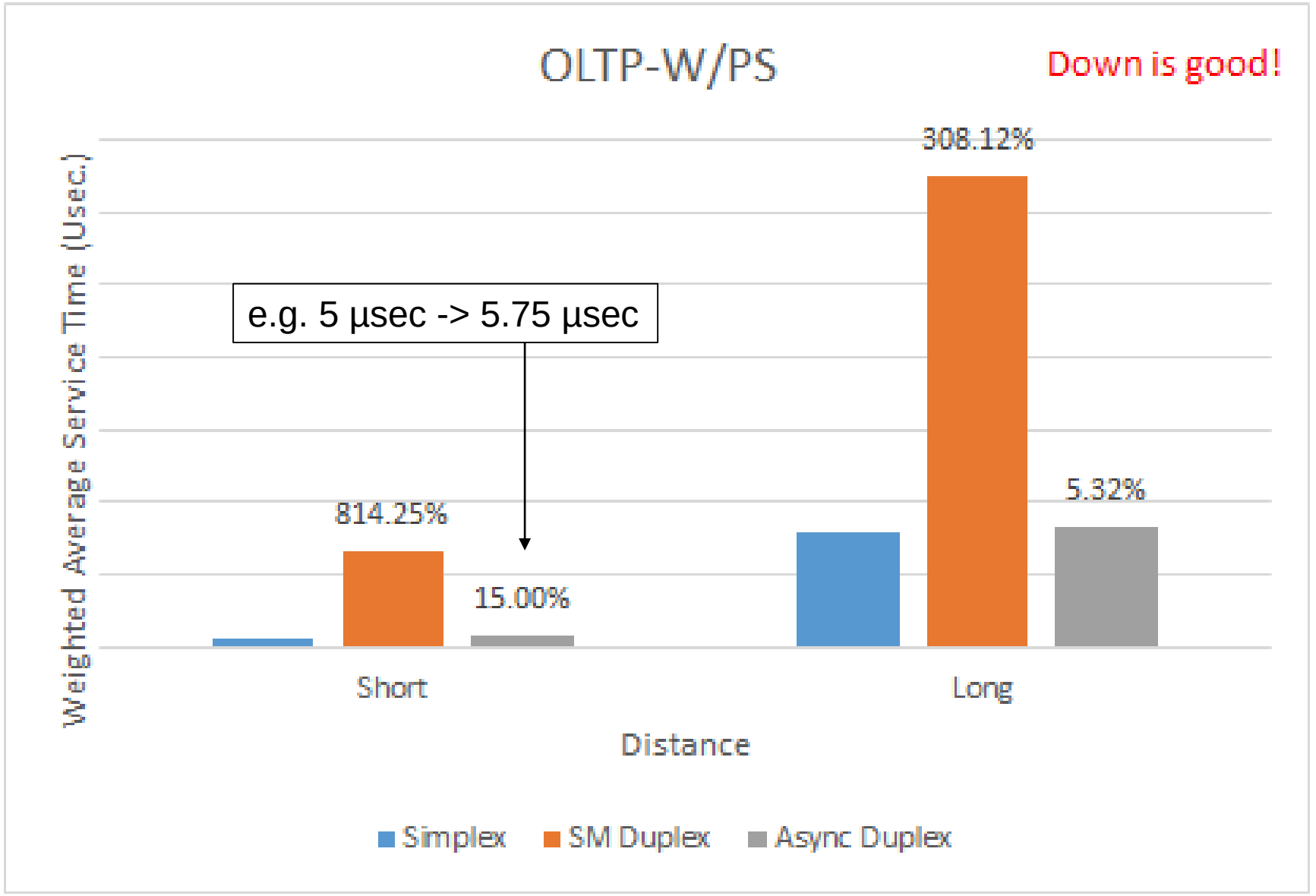
- z/OS may need to make sure the secondary instance gets caught up before it can allow traditional failure processing to proceed.
- So each system maintains a Secondary Update Recovery Table (SURT) to log local in-flight updates not yet known to have been hardened in the secondary structure instance.
- If sync up is needed, a sysplex wide coordinator is nominated to:
  - Gather the logs (SURTs) and use them to…
  - Reconstruct the final result of uncommitted in-flight requests, and
  - Update secondary instance to match the reconstructed results
- Connectors:
  - Might have requests held/delayed until sync up is completed
  - Might need to back out uncommitted transactions related to requests with ADupReqSeqNum values higher than the highest hardened request

Failover for async duplex might be a little slower than for sync duplex.

Applies when “lose” SURT for some connector

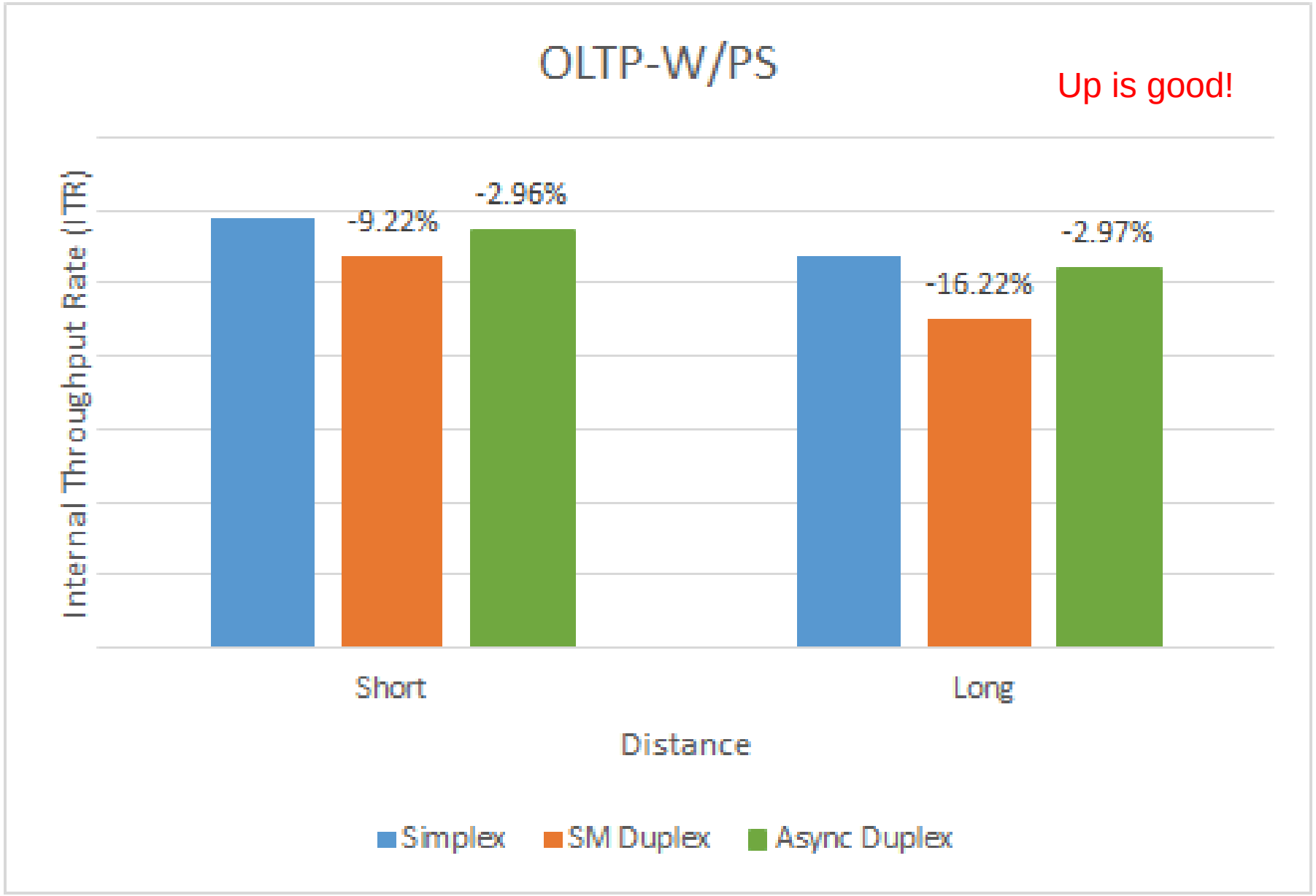
Is it really worth all this effort? …

Near simplex service times!



Short = 150m Long=10km

With less overhead!



Short = 150m Long=10km

Lock structure duplexing is now practical in more situations

- Performance is very similar to simplex operation
  - Even at distance !

- With all the benefits of robust duplex failure recovery
  - Though failover is a bit more work since secondary is not identical copy
  - And so recovery process is not necessarily transparent to exploiter

- Exploiter participation?
  - Not needed for lock structures without record data if connectors support system managed processes (XES can do it all)
  - Is needed for structures with record data
    - Exploiter must be prepared to roll back a transaction initiated by a failed connector that could not be committed because updates are missing in the surviving secondary
    - Db2 LOCK1 has record data

Requirements for Asynchronous Duplexing of CF lock structures

- At least two peer connected coupling facilities
  - CFLEVEL=21 minimum service level 02.16, or CFLEVEL=22 or higher
  - z13 GA2+, z13s, z14, z14 ZR1

- z/OS V2.2 with APARs:
  - OA47796, OA49148, OA51945, OA52015, OA52618
- z/OS V2.3 with APARs:
  - OA52618

- Db2® 12 with enabling APAR PI66689
- IRLM 2.3 with APAR PI68378

- All systems in the sysplex must be capable of doing async duplex protocols

  - Note: additional APARs may be required

RMF: CF Usage Summary

C O U P L I N G   F A C I L I T Y   A C T I V I T Y

PAGE 1

z/OS V2R2

SYSPLEX THEPLEX
RPT VERSION V2R2 RMF

START 01/05/2017-10.29.00
END 01/05/2017-11.29.00

INTERVAL 001.00.00
CYCLE 01.000 SECONDS

COUPLING FACILITY NAME = CFN7
TOTAL SAMPLES(AVG) = 1800 (MAX) = 1800 (MIN) = 1800

COUPLING FACILITY USAGE SUMMARY

GENERAL STRUCTURE SUMMARY

TYPE	STRUCTURE NAME	STATUS CHG	ALLOC SIZE	% OF CF STOR	# REQ	% OF ALL REQ	% OF CF UTIL	AVG REQ/ SEC	LST/DIR ENTRIES TOT/CUR	DATA ELEMENTS TOT/CUR	LOCK ENTRIES TOT/CUR	DIR REC/ DIR REC XI'S
LOCK	EXAMPLE_LOCK1	ACTIVE SEC A	250M	1.3	5555K	100	100	1543.0	531K 620	0 0	34M 1024	N/A N/A
	STRUCTURE TOTALS		250M	1.3	5555K	100	100	1543.0				

STATUS – Appending ‘A’ indicates that the structure is async duplexed (for this case, secondary structure instance)

RMF: Report of Coupling Facility Structure Activity

```
STRUCTURE NAME = EXAMPLE_LOCK1      TYPE = LOCK    STATUS = ACTIVE  PRIMARY ASYNC
          # REQ     -------------- REQUESTS --------------   -------------- DELAYED REQUESTS -------------
SYSTEM    TOTAL                #     % OF  -SERV TIME(MIC)-    REASON     #    % OF  ---- AVG TIME(MIC) -----  EXTERNAL REQUEST
NAME      AVG/SEC              REQ    ALL     AVG   STD_DEV              REQ    REQ   /DEL     STD_DEV   /ALL   CONTENTIONS

SYS1        300M   SYNC     294M    52.6     4.6     4.5     NO SCH     1     0.0  140.0        0.0     0.0   REQ TOTAL       395M
          83299    ASYNC   5649K     1.0    64.6    21.8                                                       REQ DEFERRED   2054K
                   CHNGD       0     0.0  INCLUDED IN ASYNC                                                    -CONT          1897K
                                                                                                               -FALSE CONT     267K

SYS2        259M   SYNC     254M    45.5     4.6     4.1     NO SCH     1     0.0  146.0        0.0     0.0   REQ TOTAL       345M
          72049    ASYNC   5134K     0.9    64.8    21.8                                                       REQ DEFERRED   2003K
                   CHNGD       0     0.0  INCLUDED IN ASYNC                                                    -CONT          1922K
                                                                                                               -FALSE CONT     233K

-------------------------------------------------------------------------------------------------------------------------------
TOTAL       559M   SYNC     548M    98.1     4.6     4.3     NO SCH     2     0.0  143.0        4.2     0.0   REQ TOTAL       740M
          155.3K   ASYNC     11M     1.9    64.7    21.8                                                       REQ DEFERRED   4057K
                   CHNGD       0     0.0                                                                      -CONT          3819K
                                                                                                               -FALSE CONT     500K
```

```
STRUCTURE NAME = EXAMPLE_LOCK1      TYPE = LOCK    STATUS = ACTIVE  SECONDARY ASYNC
          # REQ     -------------- REQUESTS --------------   -------------- DELAYED REQUESTS -------------
SYSTEM    TOTAL                #     % OF  -SERV TIME(MIC)-    REASON     #    % OF  ---- AVG TIME(MIC) -----  EXTERNAL REQUEST
NAME      AVG/SEC              REQ    ALL     AVG   STD_DEV              REQ    REQ   /DEL     STD_DEV   /ALL   CONTENTIONS

SYS1       2797K   SYNC    2797K    50.4    17.0     3.5     NO SCH     0     0.0    0.0        0.0     0.0   REQ TOTAL       395M
          777.1    ASYNC       0     0.0     0.0     0.0                                                       REQ DEFERRED   2054K
                   CHNGD       0     0.0  INCLUDED IN ASYNC                                                    -CONT          1897K
                                                                                                               -FALSE CONT     267K

SYS2       2757K   SYNC    2757K    49.6    15.6     3.6     NO SCH     0     0.0    0.0        0.0     0.0   REQ TOTAL       345M
          766.0    ASYNC       0     0.0     0.0     0.0                                                       REQ DEFERRED   2003K
                   CHNGD       0     0.0  INCLUDED IN ASYNC                                                    -CONT          1922K
                                                                                                               -FALSE CONT     233K

-------------------------------------------------------------------------------------------------------------------------------
TOTAL      5555K   SYNC    5555K     100    16.3     3.6     NO SCH     0     0.0    0.0        0.0     0.0   REQ TOTAL       740M
          1543     ASYNC       0     0.0     0.0     0.0                                                       REQ DEFERRED   4057K
                   CHNGD       0     0.0                                                                      -CONT          3819K
                                                                                                               -FALSE CONT     500K
```

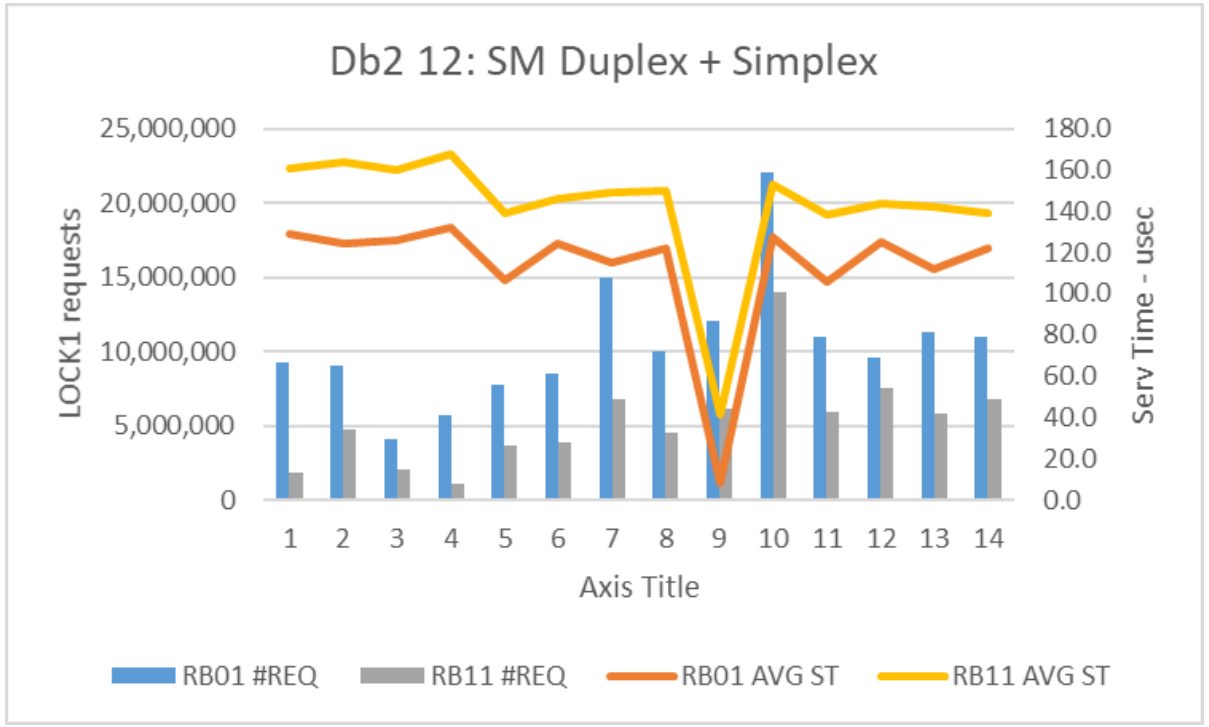
RMF: Report of Coupling Facility Structure Activity (continued)

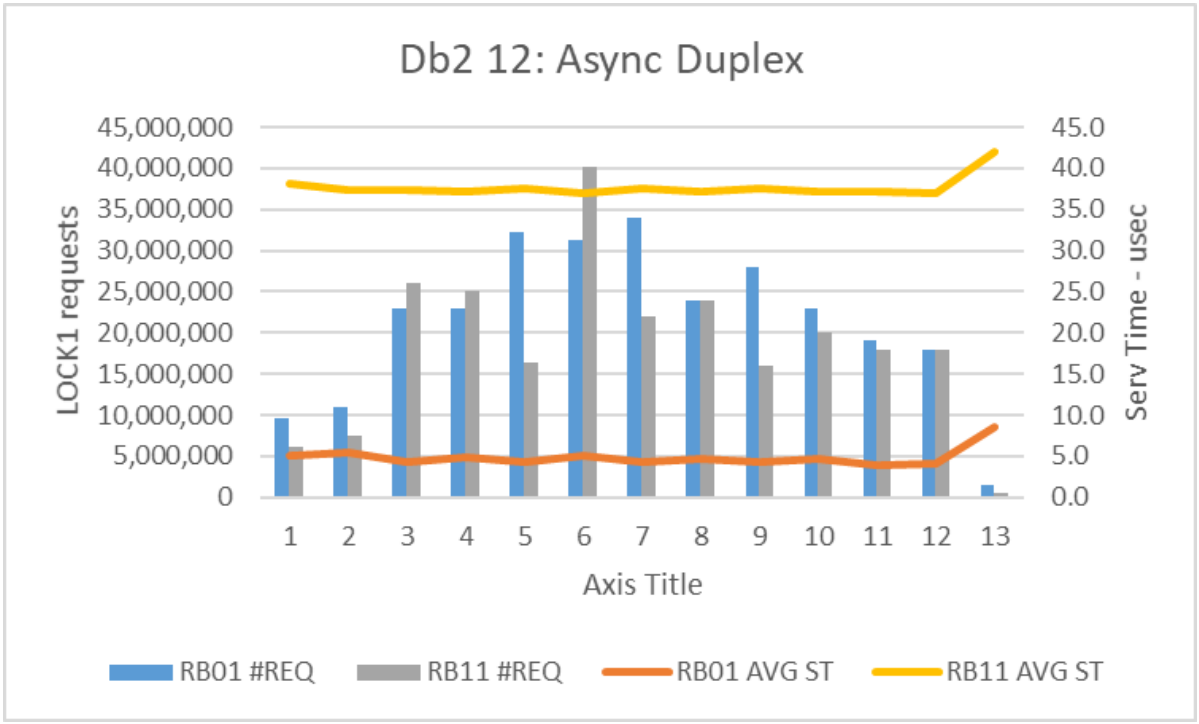
Things to note in the reports (prior slide):

- Big difference in number of requests to each lock structure instance
  - Primary instance is target for the actual lock requests
  - Secondary is target for “inquiries” as to progress
- Response times are different
  - Locks vs “inquiries”
  - The requests have very different natures

Customer experience

— European bank
- 2 sites, 2-3 km apart
- Had deployed System Managed duplexing except on first workday of each month
  - Too expensive then
- Now use async CF duplexing on all days of the month

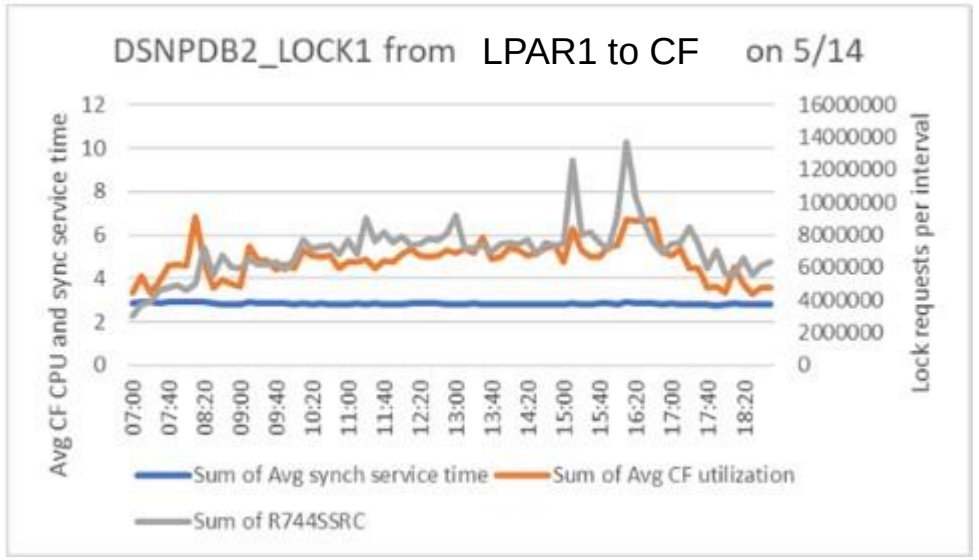


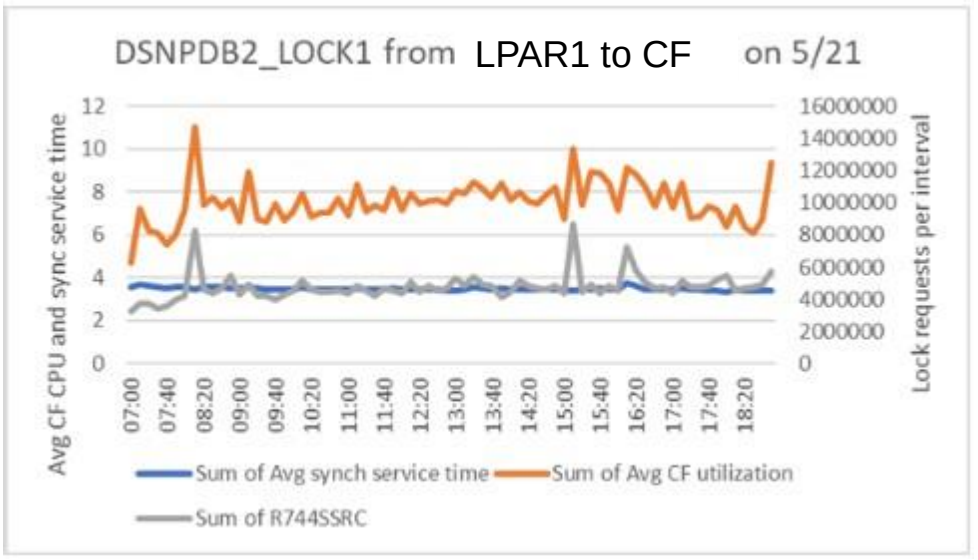


Customer experience

— North American bank (1|2)

- Had 2-CEC and 2-ICF configuration, and was exposed to ‘double failure’ situation; chose not to use synchronous SMD (although did test it)
- Implemented asynchronous SMD earlier this (2019)

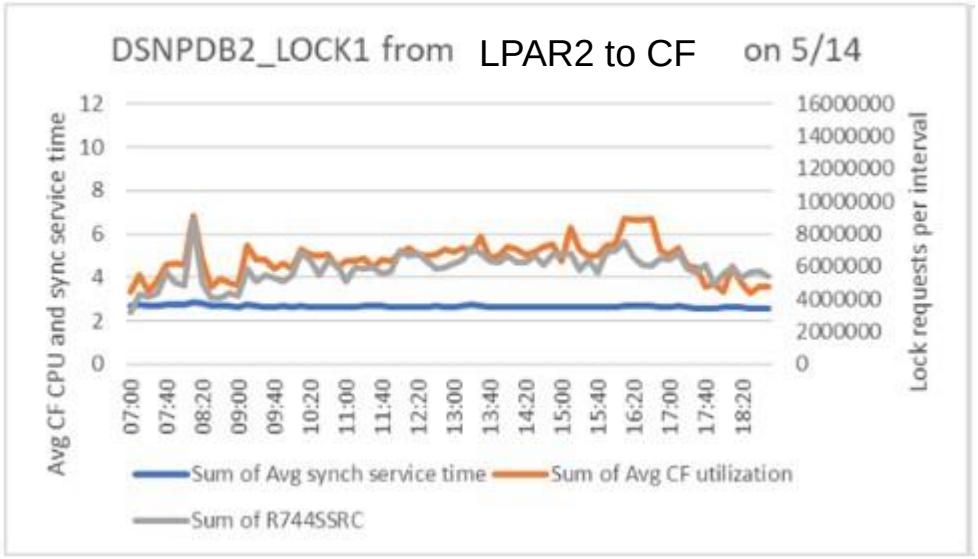


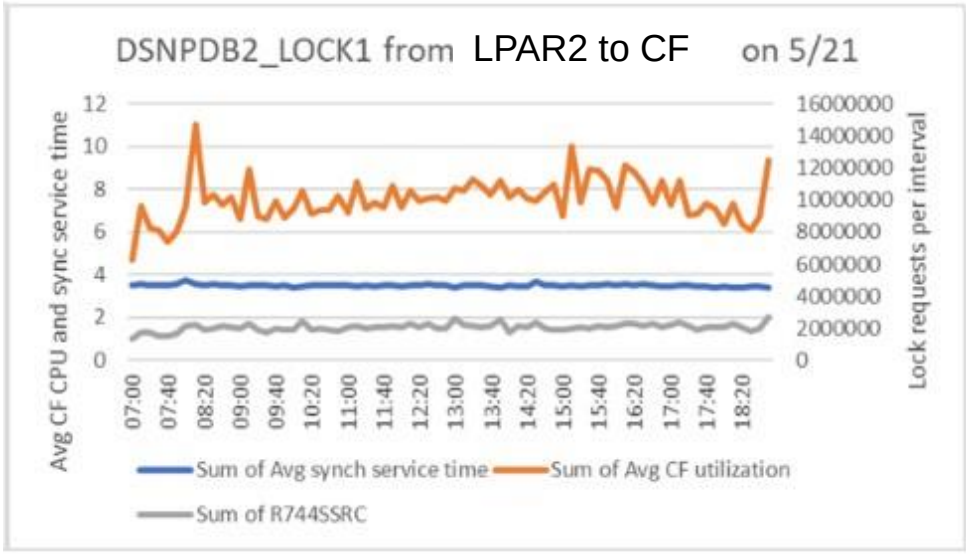


Customer experience

— North American bank (2|2)

- Had 2-CEC and 2-ICF configuration, and was exposed to ‘double failure’ situation; chose not to use synchronous SMD (although did test it)
- Implemented asynchronous SMD earlier 2019





Are you a candidate for async CF duplexing?

— Are you System Managed duplexing your LOCK1 structure?

- Should you be?
- Are you exposed to the availability risk of not duplexing your LOCK1 and SCA?

— If you are a candidate for asynchronous CF duplexing, please let me know.

- This technology can be very helpful, and there are a number of shops that should be doing this.

Questions?

— ? ? ?

Session summary

— Asynchronous CF duplexing for lock structures provides

- Robust failure recovery
- Simplex-like response time
- Even at distance

— Request: are you interested in deploying async systems managed duplexing?

- I am interested in working with you.
- Please consider sharing your data; I will anonymize it…

Notices and disclaimers

- © 2019 International Business Machines Corporation.
  No part of this document may be reproduced or transmitted in any form without written permission from IBM.
- **U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**
- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.**
  IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
- IBM products are manufactured from new parts or new and used parts.
  In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**
- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those
- customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
- It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers continued

- Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of 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. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**
- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.
- IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml