# Coupling Technology
# Overview and Planning

## What's the right stuff for me?

SHARE Seattle 16813

EWCP

Gary King
IBM

March 3, 2015

# Trademarks

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

AlphaBlox*
APPN*
CICS*
CICS/VSE*
Cool Blue
DB2*
DFSMS
DFSMShsm
DFSMSrmm
DirMaint
DRDA*
DS6000
DS8000
ECKD
ESCON*
FICON*
FlashCopy*

* Registered trademarks of IBM Corporation

GDPS*
HiperSockets
HyperSwap
IBM*
IBM eServer
IBM logo*
IMS
Language Environment*
Lotus*
Large System Performance Reference™ (LSPR™)
Multiprise*
MVS
OMEGAMON*
Parallel Sysplex*
Performance Toolkit for VM
PowerPC*
PR/SM
Processor Resource/Systems Manager

RACF*
Redbooks*
Resource Link
RETAIN*
REXX
RMF
S/390*
Scalable Architecture for Financial Reporting
Sysplex Timer*
Systems Director Active Energy Manager
System/370
System p*
System Storage
System x*
System z
System z9*
System z10

Tivoli*
Tivoli Storage Manager
TotalStorage*
VSE/ESA
VTAM*
WebSphere*
zEnterprise
xSeries*
z9*
z10
z10 BC
z10 EC
z/Architecture*
z/OS*
z/VM*
z/VSE
zSeries*

**The following are trademarks or registered trademarks of other companies.**

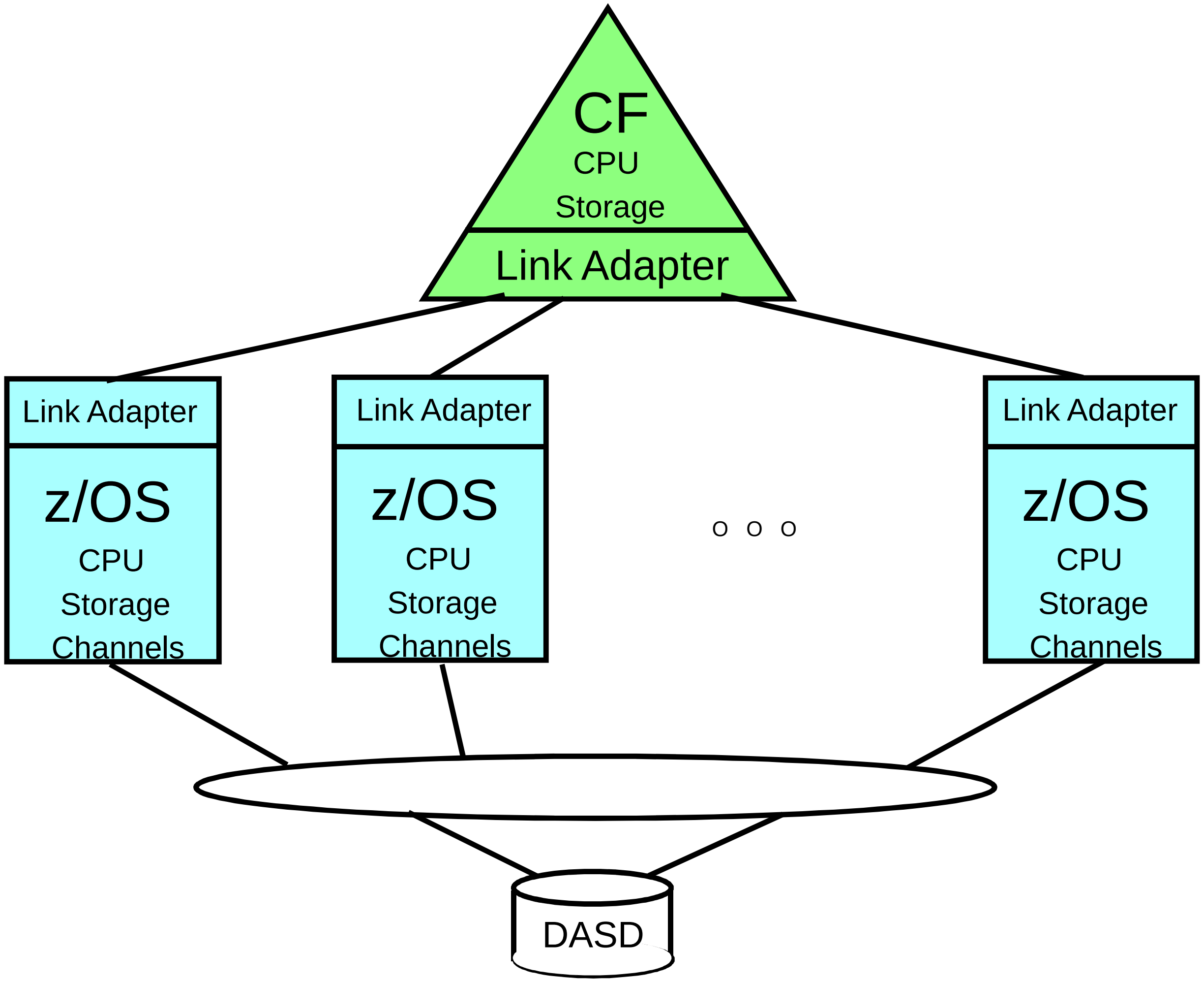
Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* 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.

# Parallel Sysplex
# Overview and hottest players

- ►Resource Sharing
  - –XCF
  - –GRS Star
  - –logs
- ►Data sharing
  - –locking
  - –global buffer pool
- ►Workload management
  - –VTAM
  - –CICS TS
  - –IMS SMQ
  - –MQ shared queues

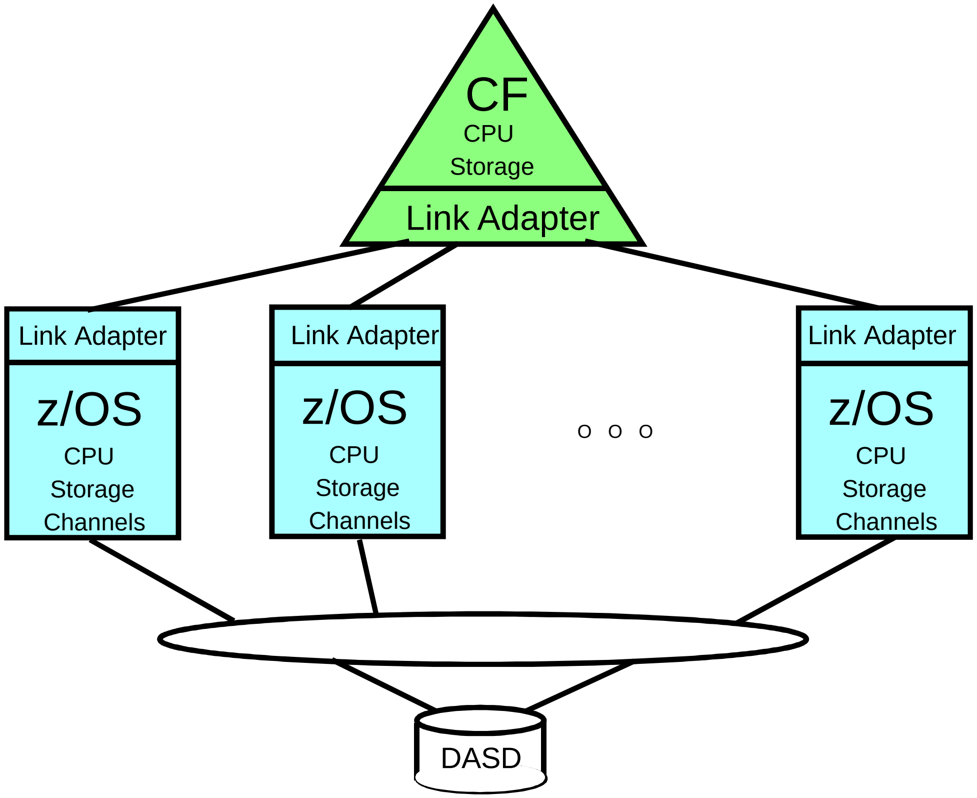


# What coupling technology is right for me?

- CF Functionality
- CF Capacity
- CF Service Time

# What coupling technology is right for me?

- CF Functionality
  - ►CFCC levels
- CF Capacity
- CF Service Time



# CFCC Level Functional Highlights

- z990/z900/z890/z800
  - CF level 1-10 functions plus ...
  - CF level 12: 64 bit exploitation, 48 tasks, SM structure duplexing
  - CF level 13: DB2 castout improvements

- z9EC/z9BC/z990/z890
  - CF level 14: CFCC dispatcher enhancements

- z9EC/z9BC/z10EC/z10BC
  - CF level 15: 112 tasks, CPU % by structure

- z10EC/z10BC
  - CF level 16: improved SM duplexing, improved LN for IMS and MQ

- z196/z114
  - CF level 17: 2047 structures, enhanced serviceability

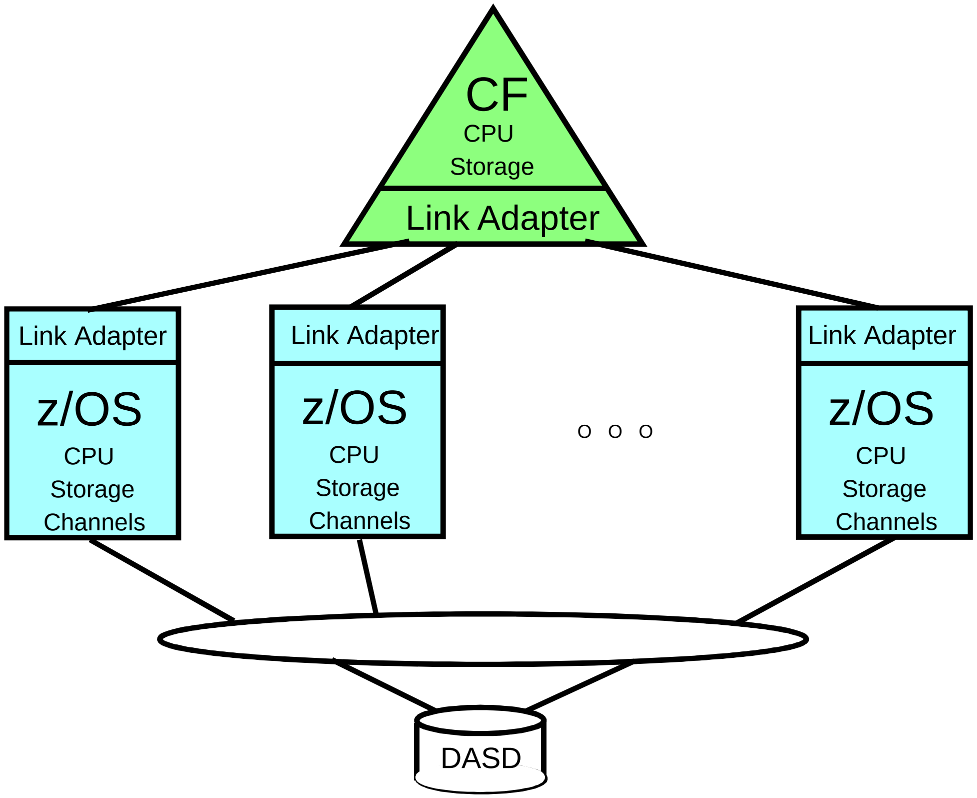
- zEC12
  - CF level 18: enhanced serviceability, improved structure size alters

- zEC12 GA2, zBC12
  - CF level 19: coupling thin interrupts

- z13
  - CF level 20: ICA SR link support, large cache structure performance improvements,
    256 coupling CHPIDS per CEC (128 per CF image)

# What coupling technology is right for me?

- CF Functionality
  - ►CFCC levels
  - ►CF Partitions
    - –dedicated versus shared processors
    - –standalone versus internal
      - •CF Duplexing



- CF Capacity

- CF Service Time

# CF Partition Options:  dedicated or shared processors?

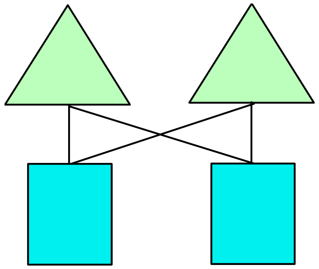
- Dedicated processor(s)
  - ►best for production

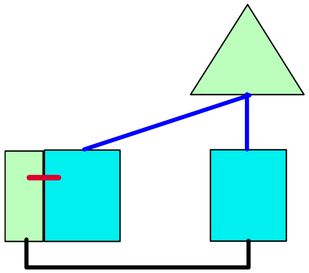
- Shared processor(s) prior to CF level 19 (zEC12 GA2, zBC12)
  - ►At best, could have 1 shared CF image with good service time, while the rest have poor to bad service time (but still often acceptable for test sysplexes)
  - ►Potential use for test or non-data-sharing production

- Shared processor(s) with CF level 19 and above
  - ►CF image dynamic dispatch setting: DYNDISP=OFF|ON|THIN
    - OFF = logical processor (LP) constantly polls for work when dispatched
    - ON = LP dynamically adjusted pattern of sleep (up to 10k mics) and polling
    - THIN = LP is awoken when work arrives, polls for work, then sleeps
  - ►Subject to PR/SM time slice management based on LP usage pattern
    - OFF will cause LP to run to end of time slice and may be undispatched
    - ON or THIN will generally have LP voluntarily sleep prior to end of time slice
  - ►THIN provides near dedicated service times to shared CF images
    - Hugely increases opportunity to share CF processors among multiple production and test/development partitions

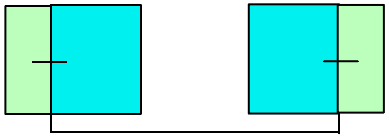
# CF Partition Options:  standalone or internal CF?

- Standalone or “logical” standalone CF in the configuration
  - Inherently provides failure isolation
  - Easier maintenance
  - More connectivity
  - Most commonly used for …
    - Large sysplexes
    - Intensive data sharing workloads



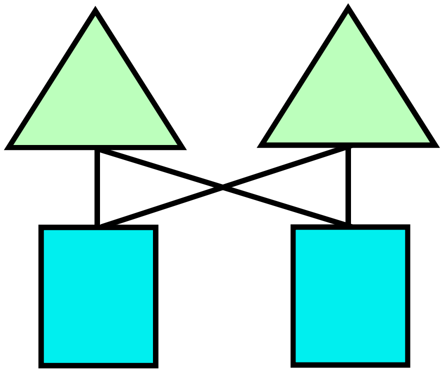


- All internal CFs in the configuration
  - Less costly than separate footprints
  - Technology upgrades simultaneously with host
  - Take advantage of internal coupling links
  - Needs SM duplexing to provide failure isolation
    - MIPS cost can be prohibitive to intensive data sharing workloads
  - Most commonly used for …
    - Smaller sysplexes
    - Resource sharing or low-intensity data sharing workloads

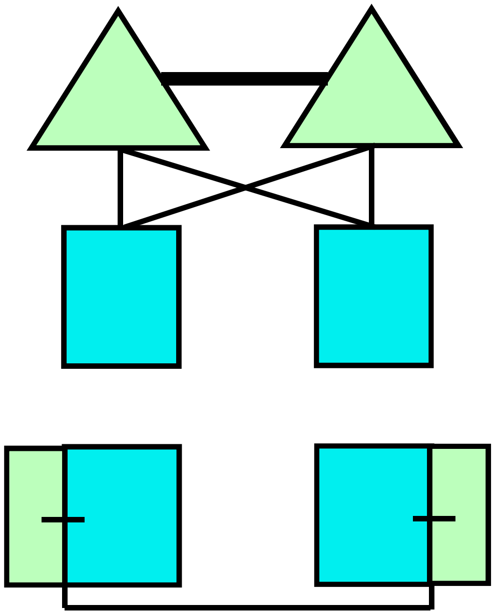


# What is CF duplexing?

- User Managed CF structure duplexing
  - ►Available only for DB2 GBPs and VSO
  - ►User (DB2 or IMS shared VSO)
    - –asks for primary/secondary structures
    - –writes updates to both
    - –synchronizes via already held locks



- System Managed CF structure duplexing
  - ►Installation selects duplexing option
    - –for specific exploiters/structures
  - ►System
    - –creates primary/secondary structures
    - –writes updates to both
    - –synchronizes via 2 CF-to-CF ops



# CF duplexing:  value vs. cost

- Value
  - ►Faster recovery from CF failures
    - –much, much faster compared to log recovery (40x)
    - –faster compared to rebuild (4x)
  - ►Provides failure isolation
    - –fully exploit ICFs

- Cost
  - ►Increased resource requirements:  host CPU, CF CPU, CF links
  - ►User Managed (DB2 GBP and VSO structures)
    - –2x times 1% to 100% (typically 20%) of simplex cost
  - ►System Managed (list and lock structures)
    - –3x to 5x times near 100% of simplex cost

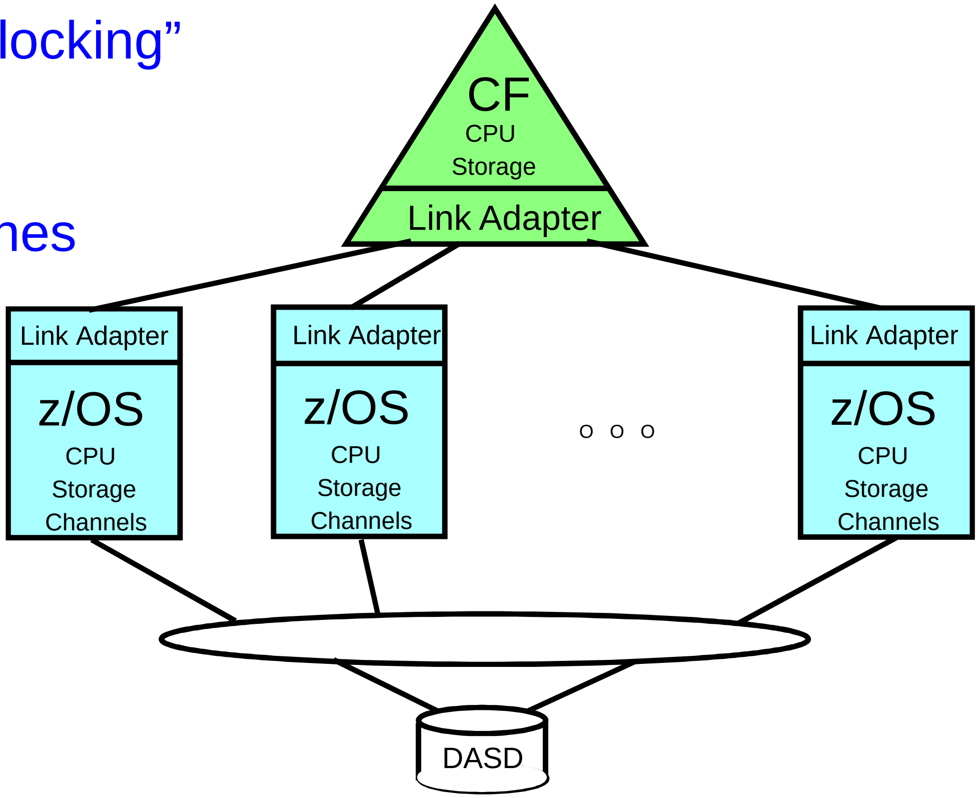
- Selectively enable when value > cost

# What coupling technology is right for me?

- CF Functionality

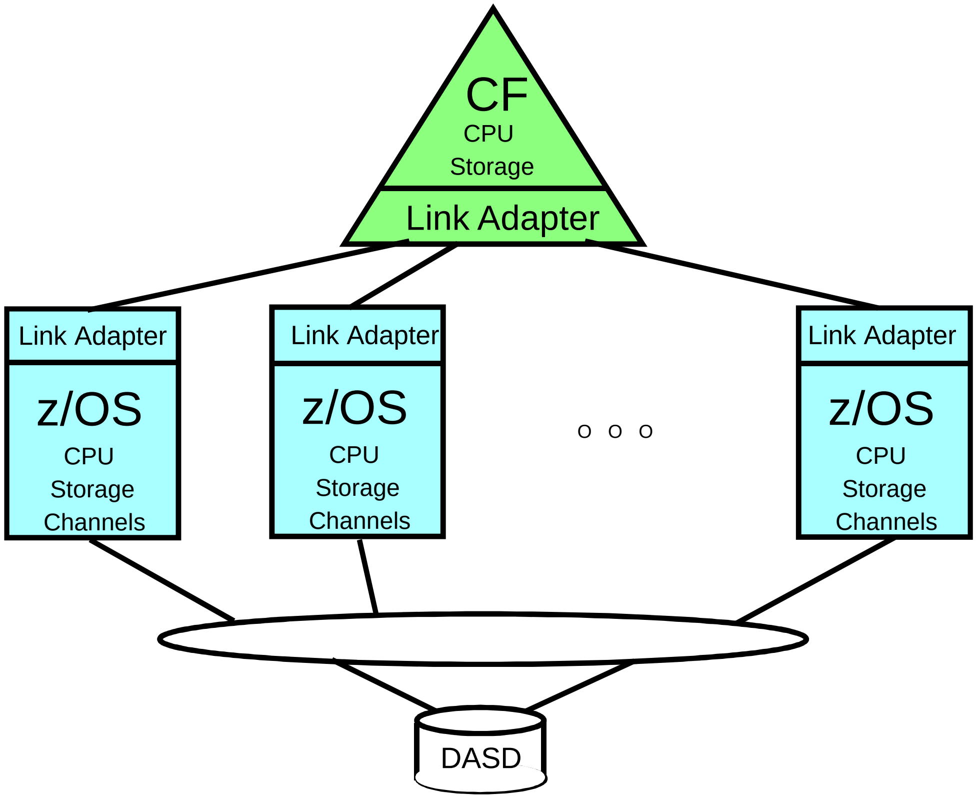
- CF Capacity
  - ►Need enough to handle the request rate (keep utilization < 50%)
    - ►Note 1way CFs <30% due to
      - •Single server queuing
      - •Long running commands “blocking” short commands
  - ►Add engines or move to faster engines

- CF Service Time



# What coupling technology is right for me?

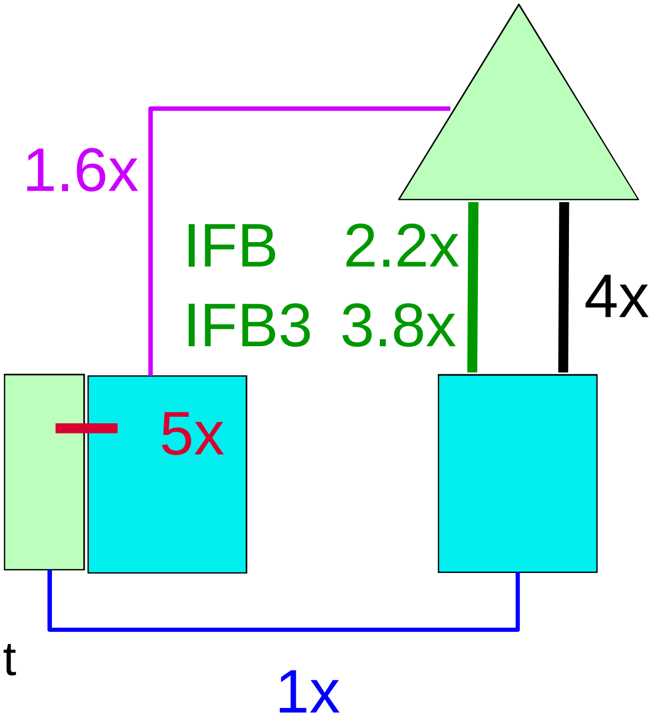
- CF Functionality
- CF Capacity
- CF Service Time
  - Affected by
    - –speed of CF engine
    - –link technology



# Coupling Link Choices - Overview

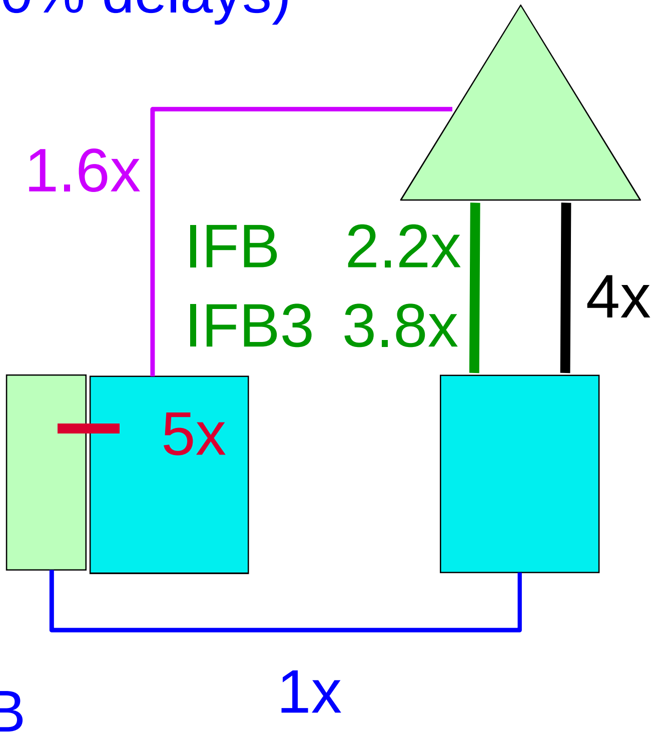
- ISC (Inter-System Channel) – NA after zEC12/zBC12
  - Fiber optics, I/O Adapter card, >10km with qualified WDM solutions
- ICB (Integrated Cluster Bus) – NA after z10EC/z10BC
  - Copper cable plugs close to memory bus, 10m max length
- IC (Internal Coupling Channel)
  - Microcode - no external connection
  - Only between partitions on same processor
- 12x IFB and 12X IFB3 (InfiniBand)
  - 150 meter max distance optical cabling
  - Supports multiple CHPIDs per physical link
  - Multiple CF partitions can share physical link
- 1x IFB
  - 10km and longer distances with qualified WDM solutions
  - Same multiple CHPIDs and sharing flexibility as 12x
  - 32 subchannels (up from 7) per CHPID (intro z196 GA2)
- ICA SR (CS5 is the CHPID name) – intro z13
  - ICA (Integrated Coupling Adapter) connects to PCIe fanout
  - 150 meter max distance
  - Supports up to 4 CHPIDs per physical link

Relative Performance
Based on avg data xfer size



# IFB and ICA SR Link Configuration Advantages

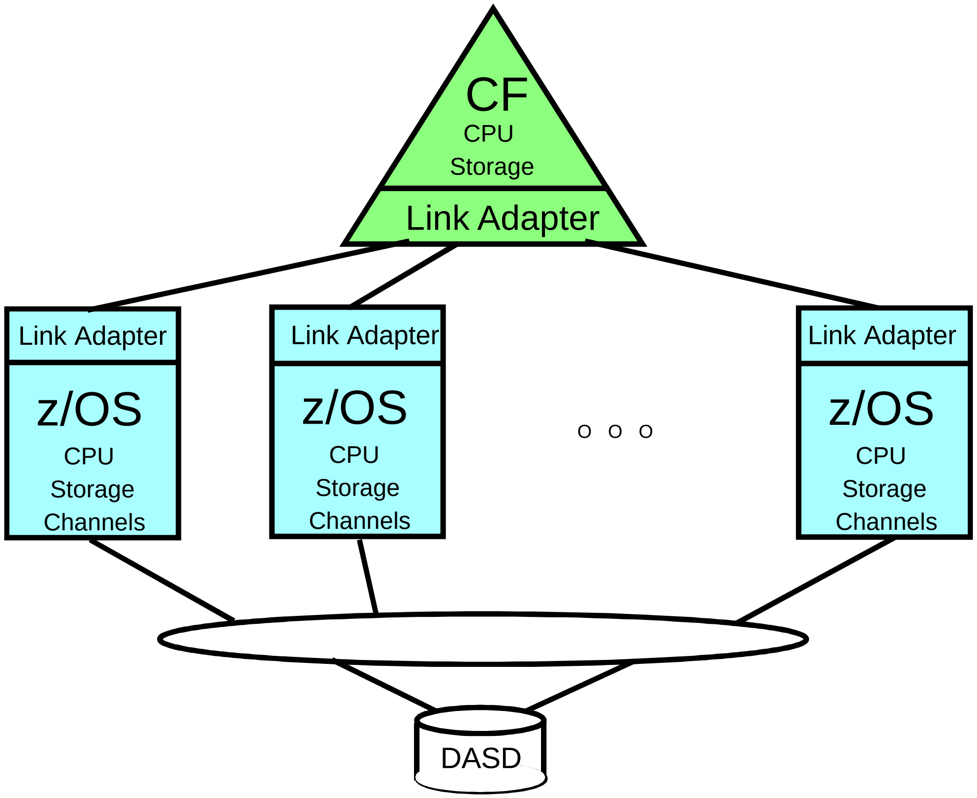
- Pure Capacity
  - 1 1x IFB = 1 ISC3
  - 1 ICA SR = 1 12x IFB(3) = 1 ICB4 = 4 ISC3s
- Eliminating subchannel and path delays (ROT <10% delays)
  - Often >2 ICB4s or >2 ISC3s are configured not for capacity but for extra subchannels/paths to eliminate delays
  - Multiple CHPID support with ICA and IFB links and 32 subchannel support with 1x IFB can often be used in lieu of adding more links beyond 2 for redundancy
- Multiple sysplexes sharing hardware
  - Production, development, test sysplexes may share hardware – each needs own ICB4 or ISC3 links
  - Multiple CHPID support with ICA and IFB links can often be used in lieu of configuring separate links for each sysplex



- Multiple CHPID recommendations for ICA and IFB
  - Most sysplexes will find 2 links defined with 2 CHPIDs each to be sufficient
    - ►Provides 28 subchannels which generally is sufficient to stay under ROT of <10% delays
  - May define up to a max of 4 CHPIDS per link for connectivity or to reduce delays for heavy loads
    - ►4 is either the practical limit (IFB) or the actual limit (ICA)

# What coupling technology is right for me?

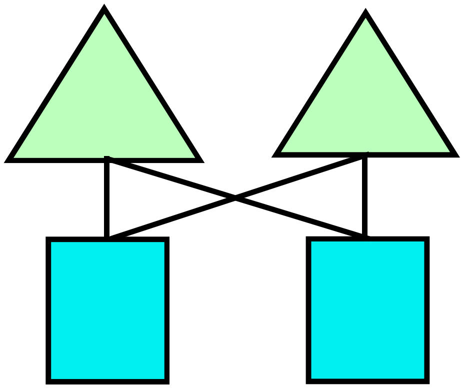
- CF Functionality
- CF Capacity
- CF Service Time
  - ►Affected by
    - –speed of CF engine
    - –link technology
  - ►Affects cost of data sharing
    - –host processor dwells
      for synchronous requests



# CF Operations:  synchronous or asynchronous

- Describes state of host processor engine issuing CF operation

- Synchronous operation
  - ►SW cost: exploiter+XES
  - ►HW cost ("dwelling time")

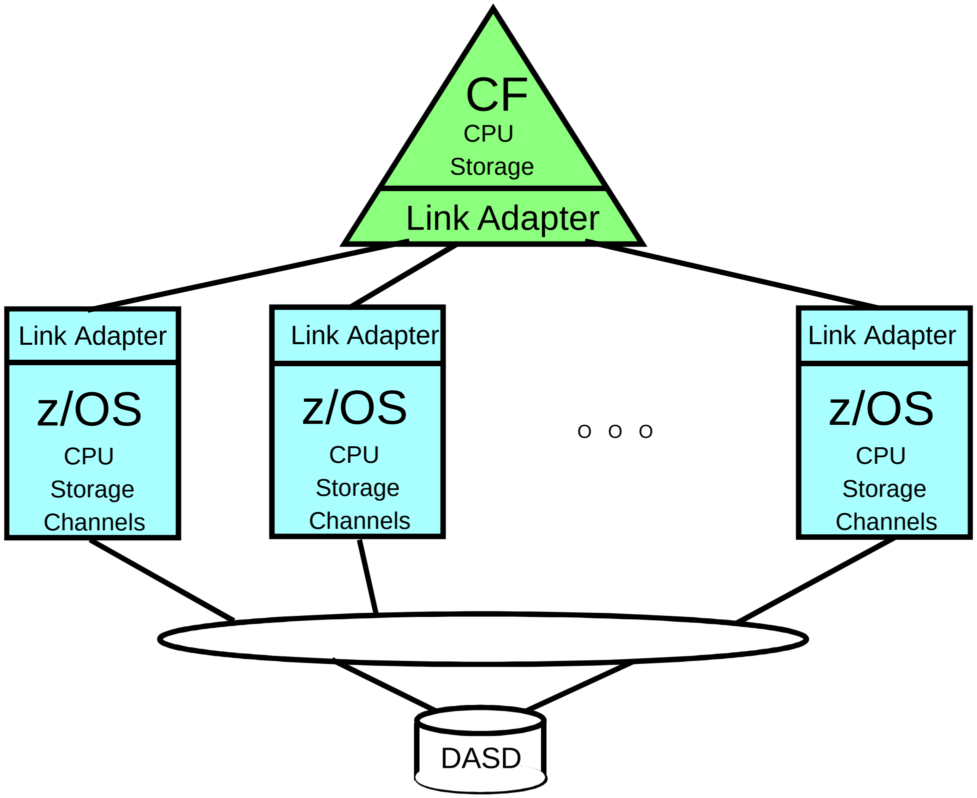


- Asynchronous operation
  - ►SW cost
    - –exploiter+XES+SRBs
    - –task switching impact on HW
  - ►HW cost - virtually none (no dwelling)
  - ►CF service time elongation
    - –added latency for XES to recognize completion of operation
    - –vastly reduced by coupling thin interrupts on zEC12 GA2 / zBC12 / z13 processors
      - • need z/OS V2R1 or proper maintenance on V1R12 and V1R13

- Which when?
  - ►Exploiter can specify synch or asynch
  - ►If synch, XES heuristic can override and issue it asynch
    - –based on measured synch service time versus "breakeven" cost of asynch
  - ►If issued synch and encounter subchannel busy, will change to asynch

# What coupling technology is right for me?

- CF Functionality
- CF Capacity
- CF Service Time
  - Affected by
    - –speed of CF engine
    - –link technology
  - Affects cost of data sharing
    - –host processor dwells for synchronous requests
  - Impact relative to ...
    - –Rate of requests to the CF
    - –Speed of the host processor



# Host Capacity Effect

- Directly related to activity to CF
  - CF request rate x SW+HW cost

- Varies based on
  - Portion of workload involved in data sharing
  - Access rate to shared data
  - Type of hardware for Host, CF and CF links

- Typical system-level effects
  - Resource Sharing:  2-3% versus single image
  - Data sharing primary production application:  5-10%

- Individual Transaction/Job effects - can have wide variation

# Production Examples

Host Effect with primary application involved in data sharing

| Industry | Trx Mgr / DB Mgr | z/OS Images | CF access per Mi | % of used capacity |
|---|---|---|---|---|
| Banking | CICS/IMS | 4 | 9 | 11% |
| Banking | CICS/IMS | 8 | 8 | 9% |
| Banking | IMS/IMS | 2 | 5 | 7% |
| Pharmacy | CICS/DB2 | 3 | 8 | 10% |
| Insurance | CICS/IMS+DB2 | 9 | 9 | 10% |
| Banking | IMS/IMS+DB2 | 4 | 8 | 11% |
| Transportation | CICS/DB2 | 3 | 6 | 8% |
| Banking | IMS/IMS+DB2 | 2 | 7 | 9% |
| Retail | CICS/DB2+IMS | 3 | 4 | 5% |
| Shipping | CICS/DB2+IMS | 2 | 8 | 9% |

# Coupling Technology versus Host Processor Speed

Host effect with primary application involved in data sharing
Chart is based on 9 CF ops/Mi – may be scaled linearly for other rates

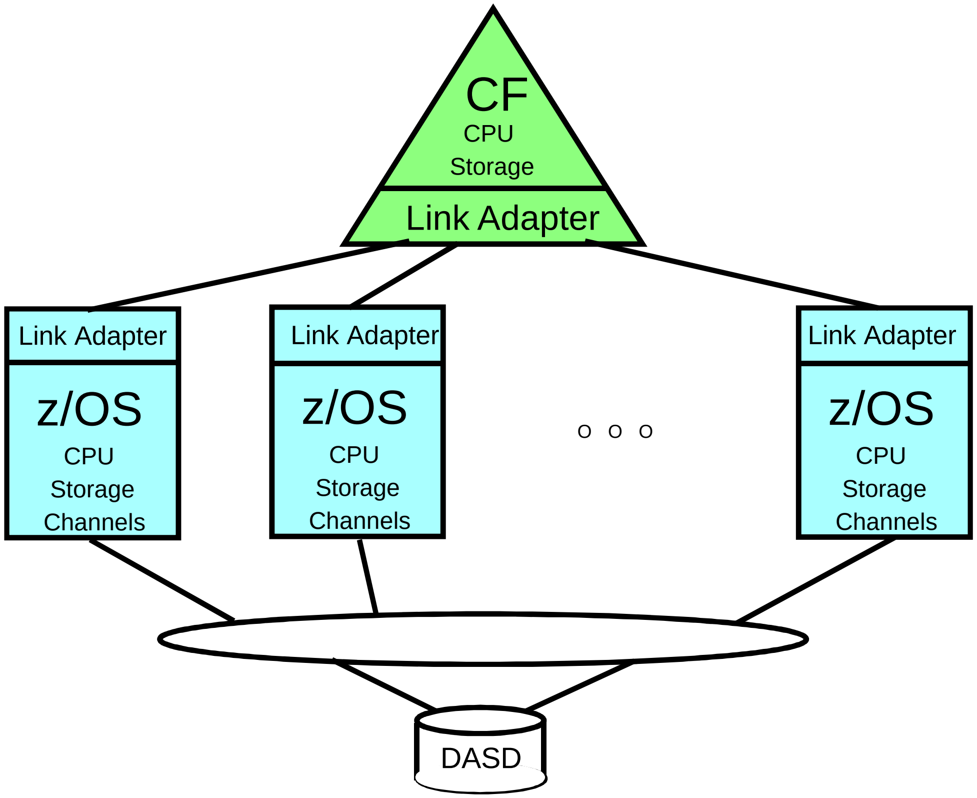
| CF\Host | z114 | z196 | zBC12 | zEC12 | z13 |
|---|---|---|---|---|---|
| z114 ISC3 | 17% | 21% | 19% | 24% | NA |
| z114 1x IFB | 14% | 17% | 17% | 21% | 22% |
| z114 12x IFB | 12% | 15% | 15% | 17% | 19% |
| z114 12x IFB3 | 10% | 12% | 12% | 13% | 14% |
| z196 ISC3 | 17% | 21% | 19% | 24% | NA |
| z196 1x IFB | 13% | 16% | 16% | 18% | 21% |
| z196 12x IFB | 11% | 14% | 14% | 15% | 17% |
| z196 12x IFB3 | 9% | 11% | 10% | 12% | 13% |
| zBC12 ISC3 | 17% | 21% | 19% | 24% | NA |
| zBC12 1x IFB | 14% | 18% | 17% | 20% | 22% |
| zBC12 12x IFB | 12% | 15% | 14% | 17% | 18% |
| zBC12 12x IFB3 | 10% | 11% | 11% | 12% | 14% |
| zEC12 ISC3 | 17% | 21% | 19% | 24% | NA |
| zEC12 1x IFB | 13% | 16% | 16% | 18% | 20% |
| zEC12 12x IFB | 11% | 13% | 13% | 15% | 17% |
| zEC12 12x IFB3 | 9% | 10% | 10% | 11% | 12% |
| z13 1x IFB | 14% | 17% | 16% | 19% | 20% |
| z13 12x IFB | 12% | 14% | 14% | 16% | 17% |
| z13 12x IFB3 | 9% | 11% | 10% | 12% | 12% |
| z13 CS5 | NA | NA | NA | NA | 11% |

With z/OS 1.2 and above, synch-> asynch conversion caps values in the table at about 18%
IC links scale with the speed of the host technology and would provide an 8% effect in each case

# What coupling technology is right for me?

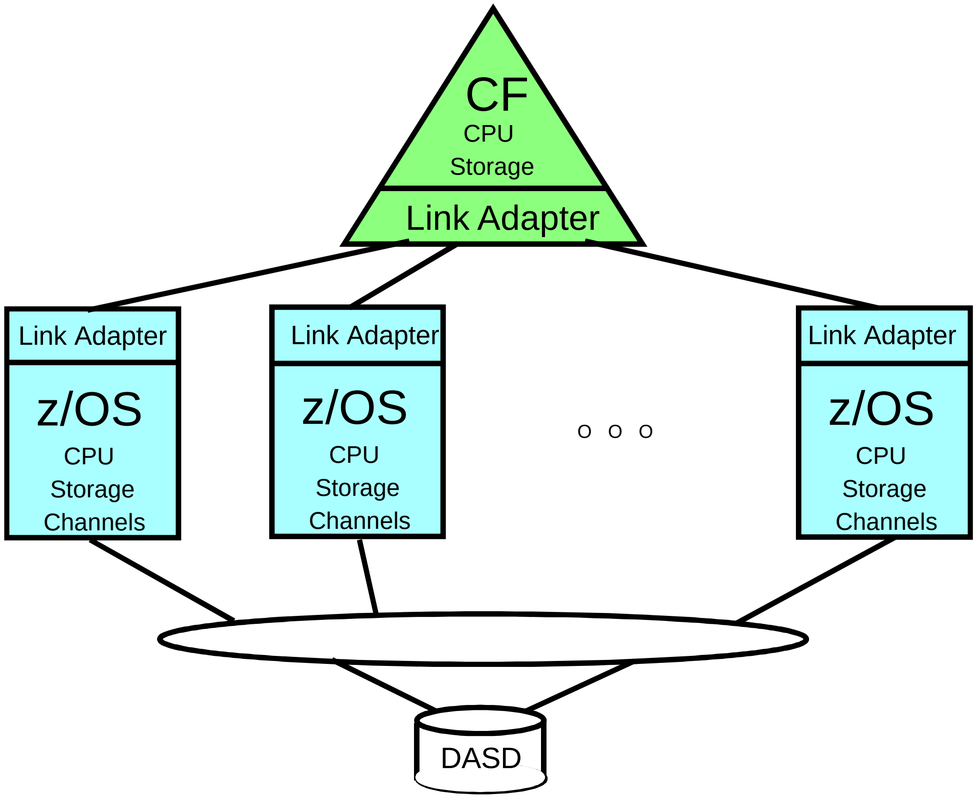
## Your Handy Dandy Checklist

- CF Functionality
  - ►CFCC level, dedicated vs shared, standalone vs internal
- CF Capacity
  - ►Need enough to handle the request rate (keep utilization < 50%)
  - ►Add engines or move to faster engines
- CF Service Time
  - ►Affected by
    - –speed of CF engine
    - –link technology
  - ►Affects cost of data sharing
    - –host processor dwells for synchronous requests
  - ►Impact relative to ...
    - –Rate of requests to the CF
    - –Speed of the host processor



# For those considering data sharing over distance ...

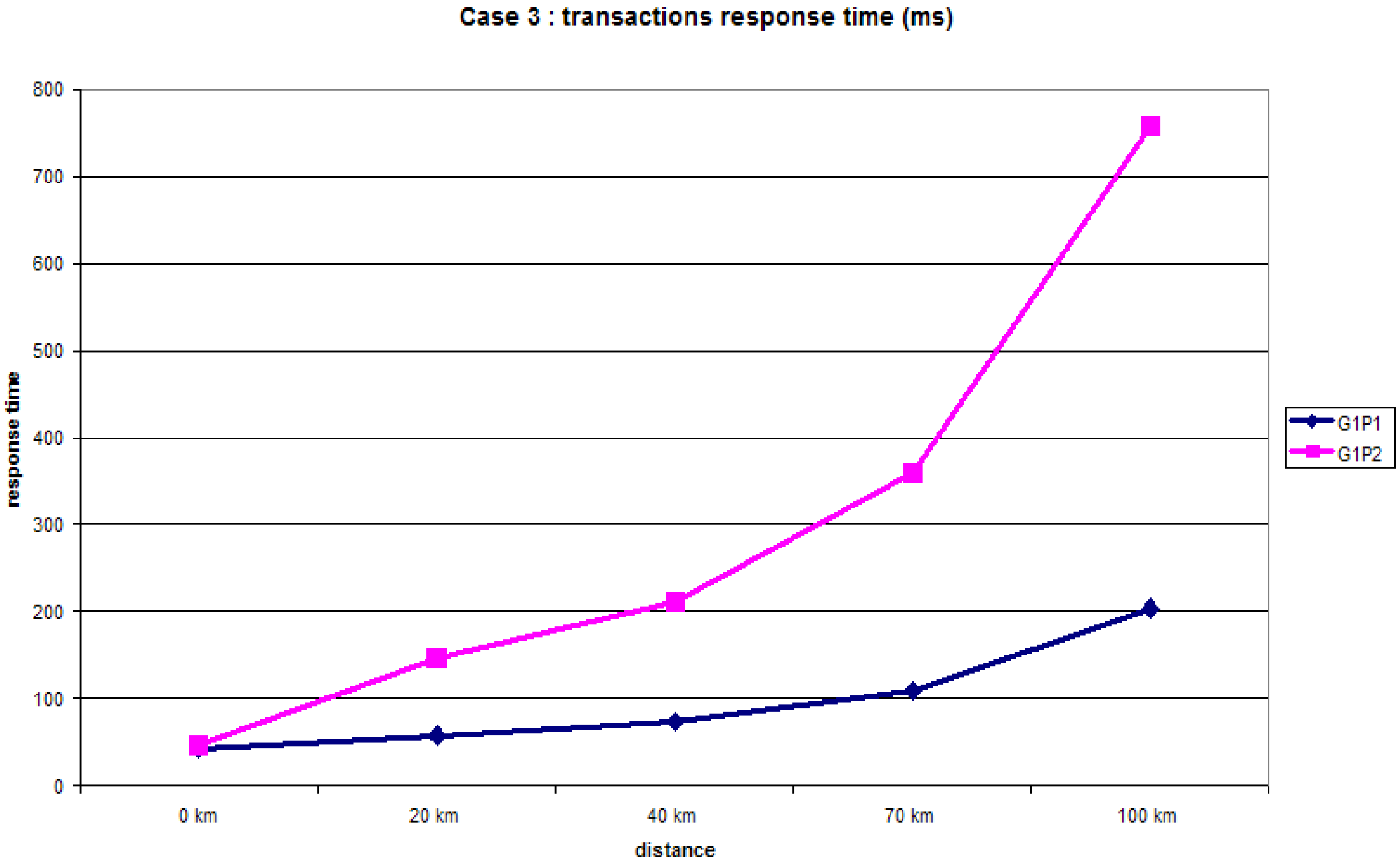
- CF Functionality
- CF Capacity
- CF Service Time
  - ►Affected by
    - –speed of CF engine
    - –link technology
    - –distance between host and CF
      - - elongates by 10 mics per km due to speed of light thru fiber
  - ►Can affect application performance
    - –transaction waits for synch and asynch requests
      - - potential impact on subsystem queues and lock contention



# Example Distance Impact

Benchmark CICS/DB2 data sharing application

- G1P1 LPAR local to CF with lock structure and primary GBPs
- G1P2 LPAR remote to CF with lock structure and primary GBPs



# Data sharing over distance summary

- CF service time elongation (versus synch request)
  - +50 mics due to change to asynch (this goes away with coupling thin interrupts)
  - +10 mics per km due to speed of light through fiber (round trip to CF)
- Host impact is capped by synch to asynch conversion
- Will likely need more link buffers (subchannels) between host and remote CF
  - link buffer (subchannel) busy grows linearly with elongated service time
  - 1x IFB multiple CHPIDs and 32 subchannels per CHPID support helps here
- Potential application performance impact
  - increased transaction response time
    - increased internal subsystem queues
    - increased lock contention
- Each application will react differently
- Difficult to predict impact
- Suggest application stress testing with simultated distance (e.g., fiber suitcases)

# References

- http://www-03.ibm.com/systems/z/advantages/pso/whitepaper.html
  - CF Configuration Options White Paper
  - System Managed CF Structure Duplexing White Paper

- http://w3.itso.ibm.com/abstracts/sg247817.html
  - System z Parallel Sysplex Best Practices

- http://www-03.ibm.com/systems/z/advantages/pso/tools.html
  - CF Structure Sizer Tool

- http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102400
  - Coupling thin interrupts white paper