Frank Kyne
Editor and Technical Consultant
Watson & Walker

IBM

Enterprise2014

Long Distance IBM Sysplex Data Sharing

Session zHA001

Agenda

- Hi, thanks for coming
  - And double points for anyone that stays awake!
- Who I am
- What we are going to talk about:
  - WHY are you interested in a long distance sysplex?
  - Long distance sysplex/data sharing topologies
  - The relationship between distance and data sharing
  - “What is the largest distance that I can do data sharing over?”
  - Technologies
  - Tools
  - Summary
  - Reference sources
- PLEASE ask questions as I go along

What are you trying to achieve?

A full-blown multi-site Parallel Sysplex *can* offer levels of availability that no other platform can touch.

- But, it *does* have limitations that must be understood to ensure that the configuration can deliver on your expectations.

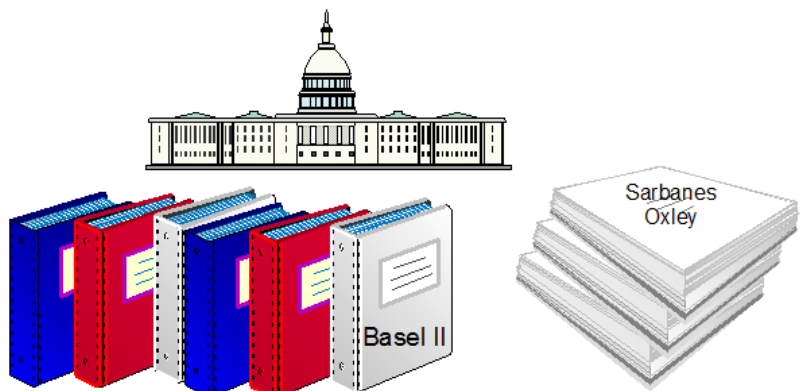
As a result, it is not unusual to see companies trying to implement multi-site sysplexes for the wrong reasons.

- The outcome of which can be disappointment, wasted time and money, missed business objectives, and disillusionment.

Instead of saying “I have two sites, what can I do with them”, the question *should be* “this is what I am trying to achieve – what is the best/most cost-effective way to do that?”

Common reasons for multi-site sysplex

Disaster recovery



Increasingly stringent government regulations for financial and other institutions – see http://www.sec.gov/news/studies/34-47638.htm

Common reasons for multi-site sysplex

Continuous Availability

IT is increasingly changing from being a tool to *support* the business to BEING the business. If your systems are down, no revenue comes into the company. WHY the systems are down is irrelevant – the line between planned and unplanned outages is disappearing.

Common reasons for multi-site sysplex

Isn't Disaster Recovery the same as Continuous Availability?

What is the "ideal" disaster recovery scenario?

Most people would say zero data loss and zero recovery time.

What does zero data loss mean? That no update is made to the Primary DASD unless it is also applied to the Secondary DASD.

If you lose connectivity between Primary and Secondary DASD, how do you ensure zero data loss? By stopping any updates to Primary DASD until connectivity is restored. During that window, you have zero application availability.

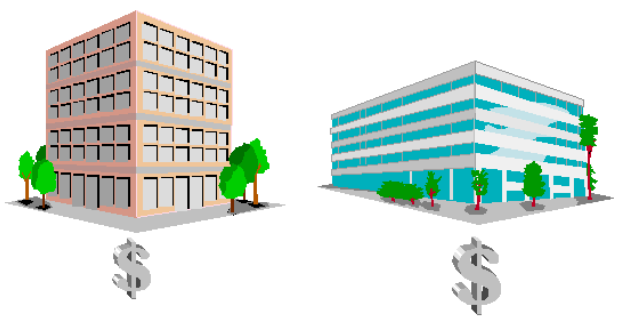
You *can* protect application availability, by allowing updates to Primary DASD during this window. But now you have updates that are not mirrored to Secondary DASD, so you don't have guaranteed zero data loss….

Common reasons for multi-site sysplex

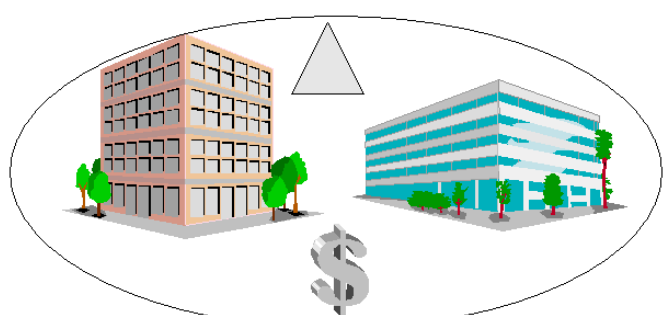
You “merged” with another company and want to take advantage of the full set of facilities that are now available to you.

Common reasons for multi-site sysplex

You have systems in two sites and want to be able to take advantage of sysplex aggregation pricing to reduce your software costs

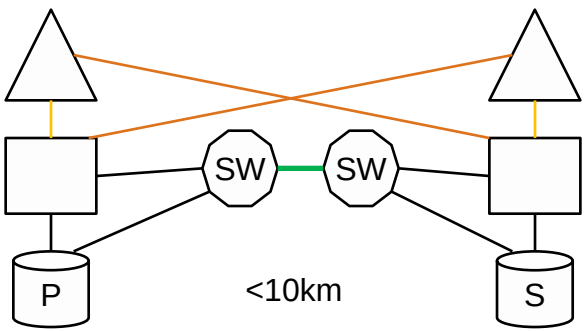


VS.



Long distance sysplex topologies

- Some typical Physical configurations – High Availability



FICON

Coupling – 12X

Coupling – 1X

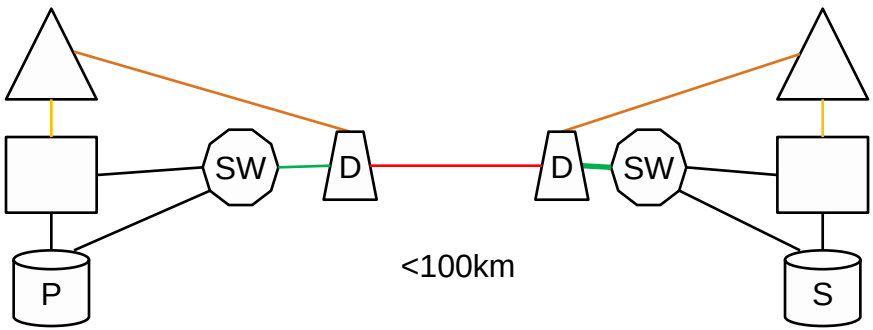
ISL

PSIFB 1X (long reach) links support up to 20km without DWDM

But modern CPCs and Storage Subsystems with 8Gb ports only have enough buffer credits to fully utilize a link up to about 10km without a switch

Long distance sysplex topologies

- Some typical Physical configurations – longer distance sysplex



FICON

Coupling – 12X

Coupling – 1X

ISL

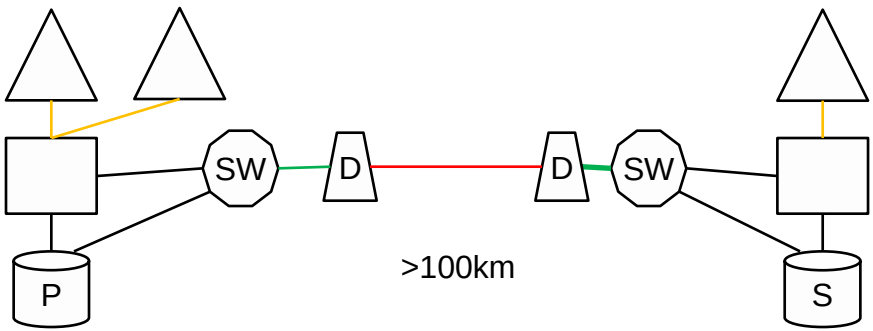
Dark Fiber

What are OTHER installations doing? I don't want to be the odd man out….

Even larger distances are possible with an RPQ – but WHY would you want to span a sysplex over such a large distance?

Long distance sysplex topologies

- Some typical Physical configurations – DR only



FICON

Coupling – 12X

ISL

Dark Fiber

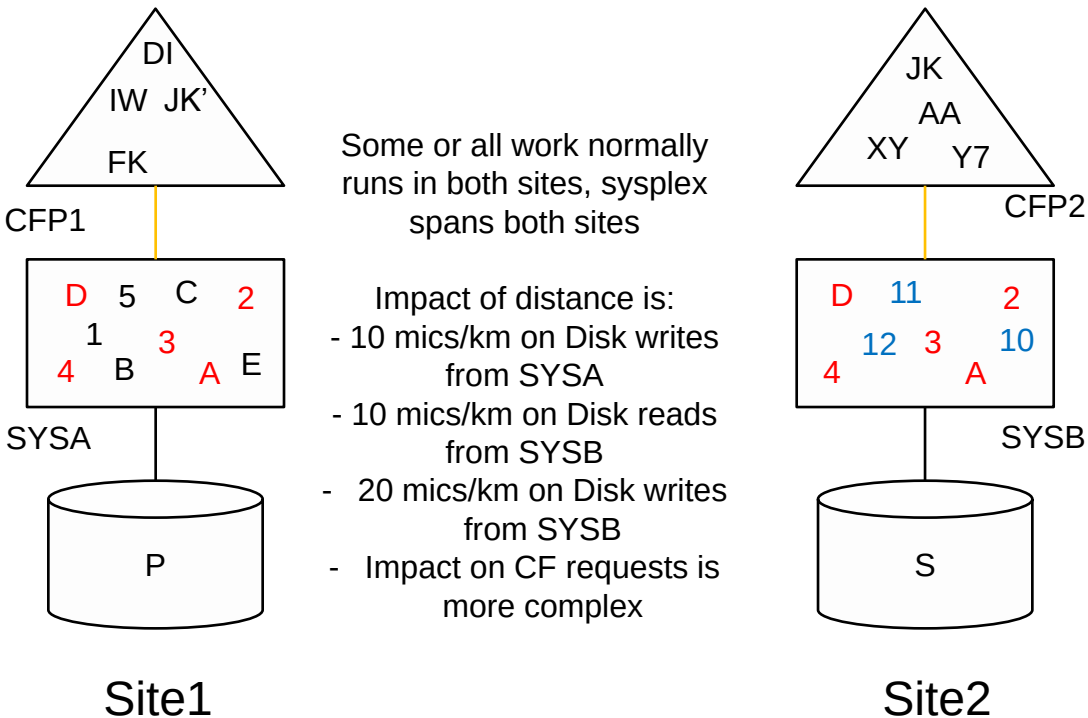
Sysplex does not span sites. Second site is only used for disaster recovery.

Long distance sysplex topologies

- Note that these distances are general *guidelines* or common practices:
  - There are people using DWDMs for less than 10km
  - There are people NOT using DWDMs for distances greater than 10km
  - There are people doing data sharing over 70km
  - There are people using a DR config over less than 100km
- IBM has published numbers for maximum supported distances
  - Just because something is SUPPORTED does not mean that it is *feasible* for you.
  - If you have a valid need to go a SMALL distance beyond the supported maximum, IBM may be willing to approve it after testing (using an RPQ 8P2340).

Long distance sysplex topologies

- Some typical *Logical* configurations – Multi-Site Workload

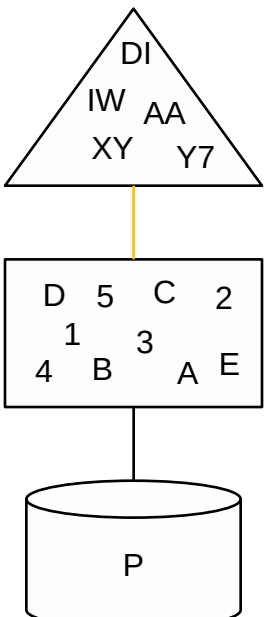


Long distance sysplex topologies

- Note that this is NOT an all or nothing situation.
  - You can aim for a Multi-Site Workload configuration without necessarily having the workload split 50/50 across the two sites. Having the sysplex span both sites provides great flexibility to move workloads between sites as necessary.
  - Typically, batch is more CF-intensive than online, so you can reduce the impact of distance by some intelligent routing of particularly CF-intensive batch jobs.

Long distance sysplex topologies

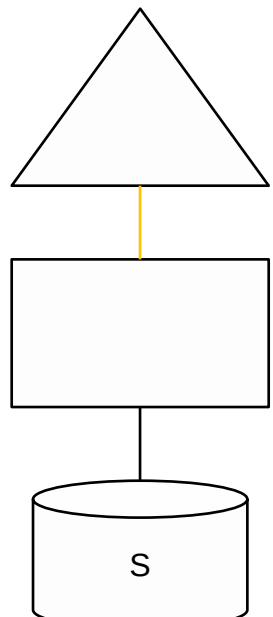
- Some typical Logical configurations – Single Site Workload



All work normally runs in the same site as primary DASD and most CF structures

Impact of distance is mainly on Disk writes

Note the difference between multi-site **sysplex** and multi-site **workload**

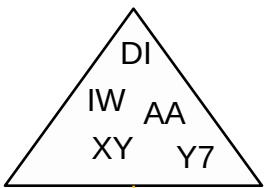


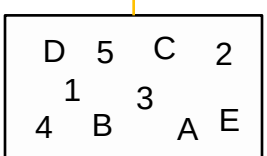
Site1

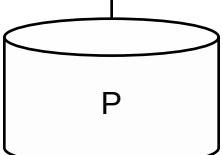
Site2

Long distance sysplex topologies

- Some typical Logical configurations – BRS config





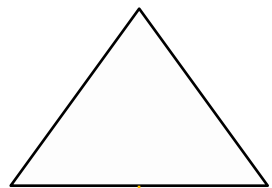


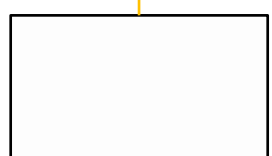
Site1

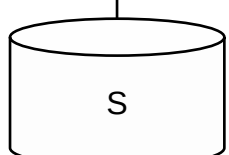
Sysplex does not span sites

No ability to non-disruptively switch between sites

Site2 only used for DR



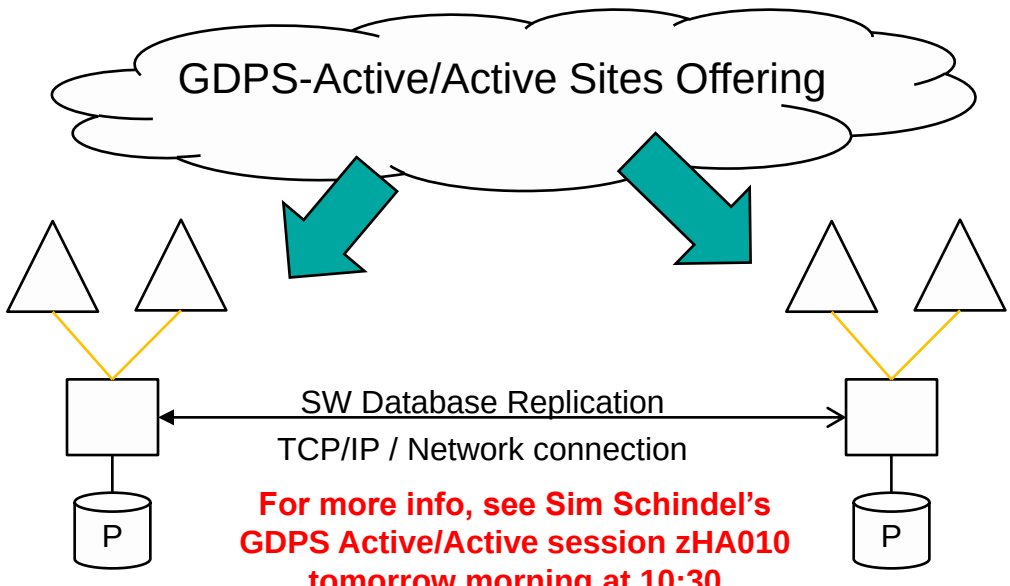




Site2

Long distance sysplex topologies

- NOT multi-site sysplex (so we won't cover it here), but does allow "data sharing" across distant sites



Long distance sysplex topologies

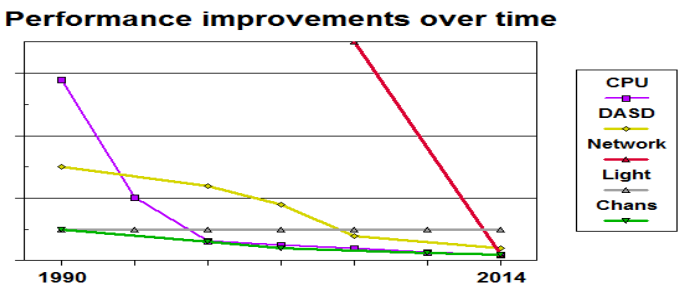
- There are multiple other flavors of multi-site sysplex:
  - Three sites, with all 3 sites in the same sysplex and within metro-distance.
    - Improved resiliency compared to two sites, but doesn't protect from regional disaster.

Over 100 3-site GDPS installations!

  - Three sites, with 1 site being "far" away (asynchronous mirroring, not in the same sysplex as first 2 sites).
    - Insurance of third, distant, site lets you place first two sites closer together
  - Four sites, made up of 2 multi-site sysplexes.
  - Multi(ish)-site sysplex. All apps run in Site1, hot standby database and transaction managers run in Site2, workload routing products control which site transactions are routed to.
    - Need to pay some cost for hot standby system idling in 2nd site.
    - However, switch to second site is much faster and less likely to encounter issues.

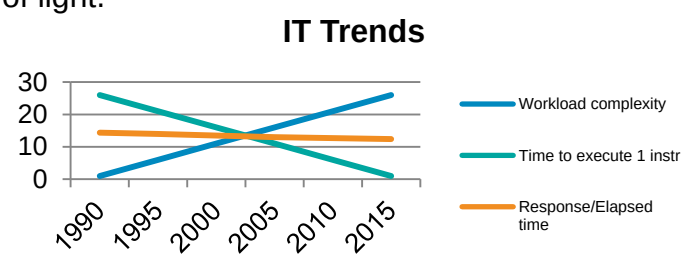
Long distance sysplex topologies

- Relationship of technology enhancements to long distance requests



Over time, as workloads and complexity have increased, technology improvements have offset those increases. But the speed of light is NOT changing and will not change in the near future.

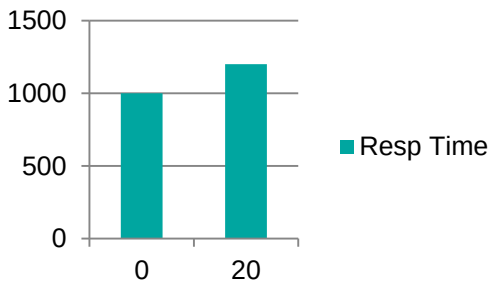
With longer distance data sharing, the bulk of your outside-the-CPU response time (disk, CF) will be related to the speed of light.



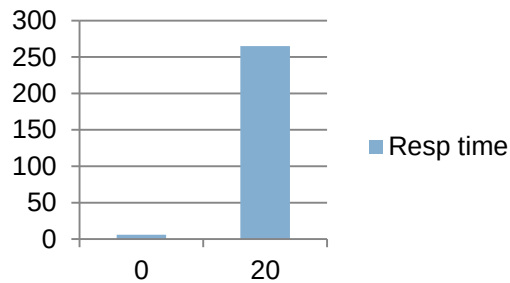
Long distance sysplex topologies

- Why is distance important in multi-site data sharing?

DASD impact of 20km



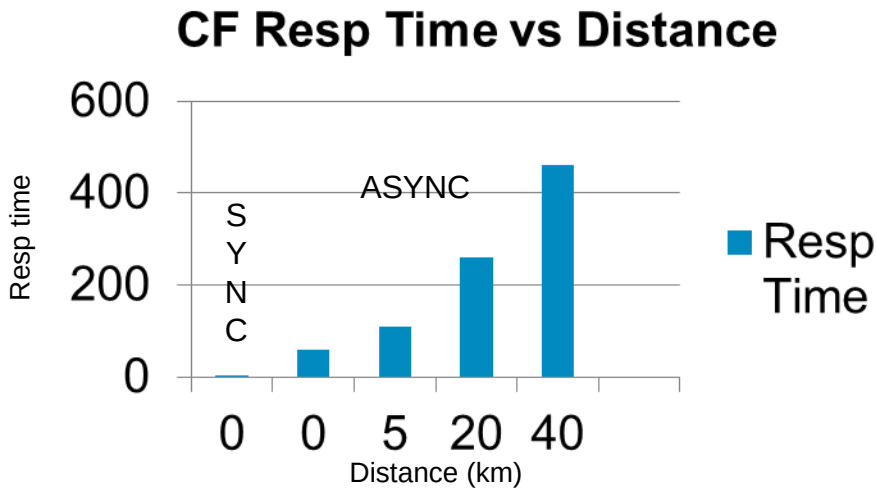
CF Impact of 20km



- In a typical data sharing environment, you do 2-3x as many CF requests as DASD I/O.
- AND, most CF requests are DURING the txn (so impact Txn response times), whereas most DASD requests are done before or after the txn (so do NOT impact Txn response times)

Long distance sysplex topologies

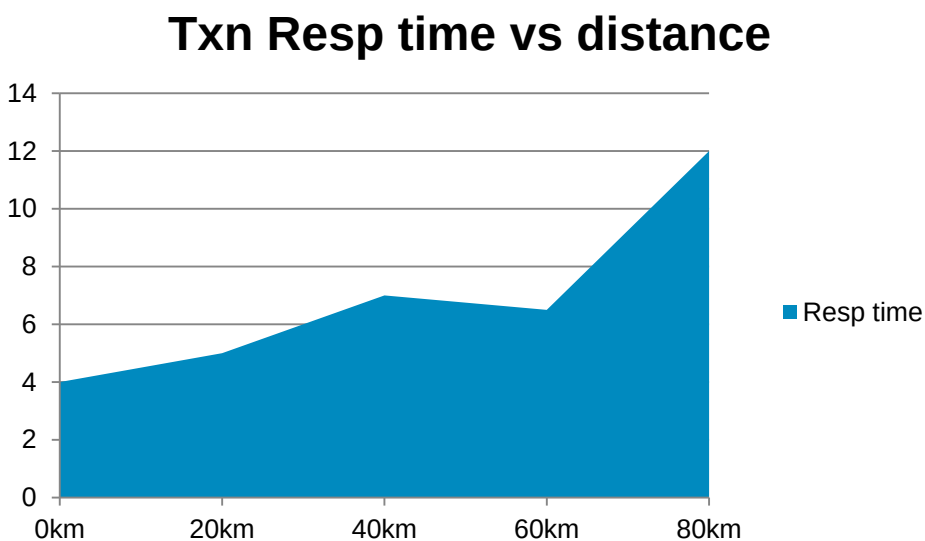
- Why is distance important in multi-site data sharing?



- At 5km, 50% of CF Resp time is due to speed of light
- At 20km, 80% of CF Resp time is due to speed of light
- At 40km, 90% of CF Resp time is due to speed of light

Some real measurements

- Impact of distance on transaction response times

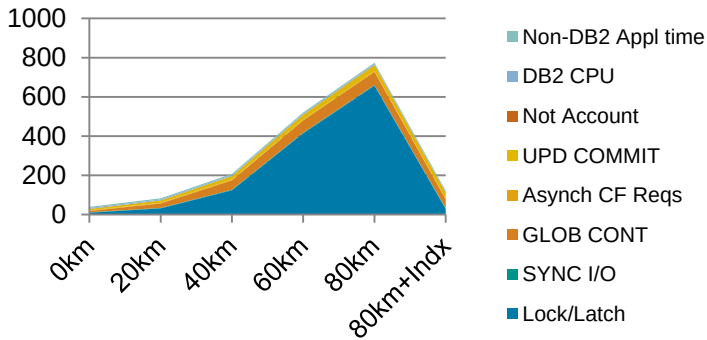


Source: IBM Redbook - Considerations for Multisite Sysplex Data Sharing, SG24-7263

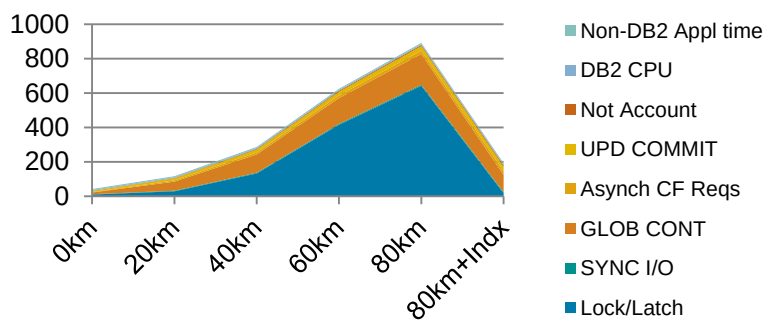
Some real measurements

- Impact of distance on a different transaction

Site1



Site2



Source: IBM Redbook - Considerations for Multisite Sysplex Data Sharing, SG24-7263

Some real measurements

- These results illustrate:
  - The non-linear response time growth as distance increases
  - The fact that response times MAY be acceptable at larger distances
  - Different applications are impacted differently by distance
  - Different *transactions* are impacted differently by distance
  - Lock contention is especially sensitive to distance
  - Applications are much more exposed to poor design and poor database design at larger distances
  - Relationship between application location and CF structure and primary DASD location *does* make a difference at larger distances

- All of these illustrate why the right answer to “how far can I do data sharing over?” is “It depends”.

What you need to consider

- You need to consider:
  - Your objective – EXACTLY what are you trying to achieve? Multi-site sysplex should deliver better overall availability but it does NOT guarantee that you will never have an outage again.
  - Your objective also helps you identify a realistic maximum distance between the sites.
  - Now that you know what you are trying to achieve, this should help you identify what *applications* you plan to run in each site.
  - Knowing what applications will run in each site helps you identify what *hardware* you will need in each site.
  - Knowing what hardware you would like to have in each site, plus the distance between the sites, helps you identify your connectivity requirements – DWDMs, Switches, Encryption, Compression, Coupling Link types, channel converters (parallel/ESCON/FICON), etc

Technologies

- All of this presentation assumes that you are using synchronous DASD mirroring.
  - HyperSwap is THE enabler of the non-disruptive site switch capability enabled by a multi-site sysplex/multi-site workload configuration.
    - HyperSwap is dependent on synchronous mirroring
    - *If you don't have HyperSwap (or equivalent) an outage is required for a planned site switch even if all applications are running everywhere.*
  - PAV and HyperPAV should be considered an absolute pre-requisite for multi-site sysplex configuration. Longer distance between CPC and Primary DASD and between Primary and Secondary DASD drive up IOSQ time which is offset by PAV.
  - Data-in-Memory exercise optimizes I/O by eliminating or reducing the number of DASD I/O operations. The fastest I/O is the I/O that is never started.
  - MIDAW does not appear to have any impact on distance, one way or the other.

Technologies

- More DASD
  - zHPF can *hurt* write performance at larger distances.
    - Investigate zHPF Extended Distance option to bring long distance response times closer to native FICON levels – see your vendor storage specialist for more info.
  - Use the highest bandwidth adapters (FICON Express8S, 8Gb switch ports, 16Gb ISL links, 8Gb DASD Subsystem adapters) available.
  - Work with switch vendor to ensure you have sufficient buffer credits.
  - Use Disk Magic and RMF Magic to ensure you have sufficient CPC-to-Primary CU and Primary CU-to-Secondary CU connectivity/bandwidth and to get accurate projections of response times and channel/adapter utilizations.
    - This will vary depending on whether you will use HyperSwap and whether you need to run with applications in one site and primary DASD in the other.

Technologies

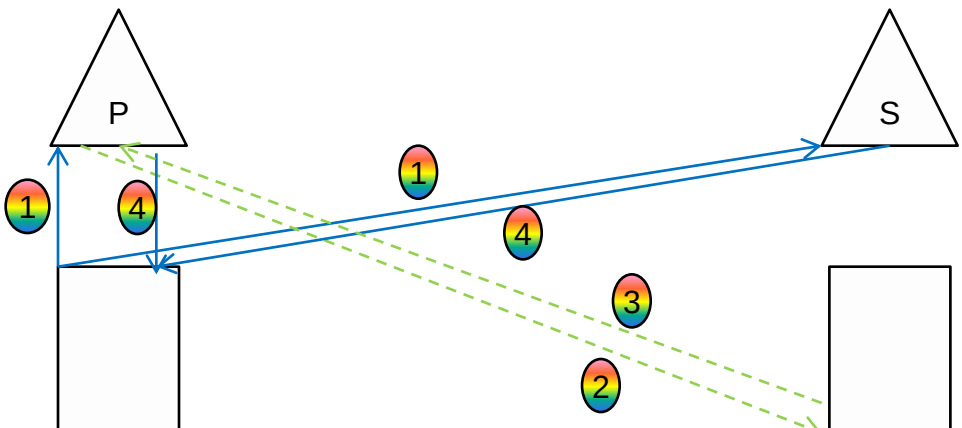
- Coupling Facility considerations:
  - There are some structures that MUST be shared by every member of a sysplex – GRS, XCF, Enhanced Catalog Sharing – basically any sysplex exploiter that doesn't support subplexing.
  - HIGHLY recommend exploiting Coupling Thin Interrupts if you have a multi-site sysplex – help minimize asynchronous response times.
  - Depending on the distance to the CF, you MIGHT consider using the SYNCASYNC function to stop requests being converted to asynchronous.
  - Remember that adding distance between z/OS and the CF will SIGNIFICANTLY increase CF Link subchannel utilization. CF subchannels are busy for the whole response time, so moving a lock structure 10km from connector could increase response time from 5 mics to about 150 mics, a 30x increase in subchannel utilization.

Technologies

- Coupling Facility considerations:
  - To help offset the impact on subchannel utilization, PSIFB 1X links support 32 subchannels per CHPID – ideal for long distance links. But you MUST explicitly define the CHPID to use 32 subchannels – the default is 7.
    - PSIFB 12X links support a max distance of 150 meters.
    - PSIFB 1X can go 10km unrepeated, 20km unrepeated with RPQ 8P2197.
      - Remember that 1X links are slower than 12X links, especially for large requests (Cache, some List) even before you take the increased distance into account.
    - Beyond 20km, some form of repeater is required. The maximum distance varies by DWDM device and feature.
    - Use of an IBM Qualified configuration is HIGHLY recommended. It is NOT just a rubber-stamp. See ResourceLink for information on GDPS Qualification process for Switches and DWDMs.

Technologies

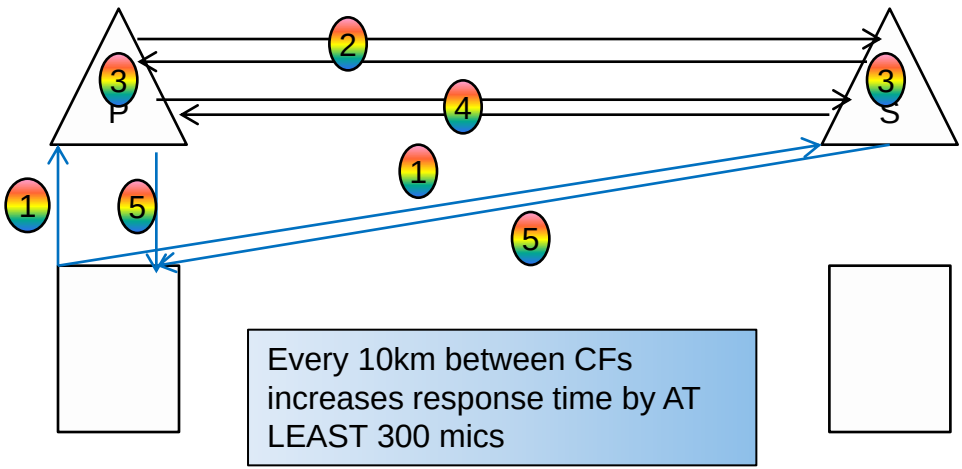
- Coupling Facility considerations:
  - User-Managed Duplexing (only used by DB2 GBPs)



1. Send update to GBP
2. CF sends cross-invalidate to other CPC
3. Check that XI was received
4. Report Update Complete back to DB2

Technologies

- Coupling Facility considerations:
  - System-Managed Duplexing



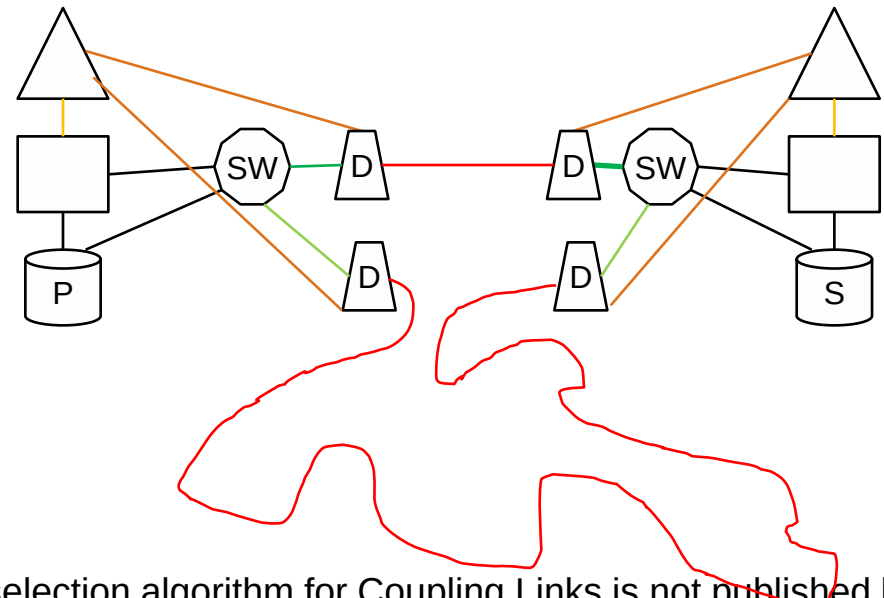
1. Send update to GBP
2. CFs send Ready to Execute to peer CF
3. Execute request
4. CFs send Ready to Complete to peer CF
5. CFs send response back to z/OS

Enable DUPLEXCF16 to reduce impact of step 4.

Technologies

- Coupling Facility structure duplexing considerations:
  - Need to think about why you are duplexing each structure? If it is to protect from CF failure, that is fine (but where should CFs be?)
  - If it is to protect from site failure, then remember that duplex copy of structure cannot be used for database restart if you don't have a FREEZE=STOP policy. This is because there is no coordination between DASD Freeze and structure duplexing.
    - CF Duplexing might stop BEFORE the DASD freeze.
      - In this case structure will revert to simplex (in some site), copy in other site is deleted, but system processing and writing to DASD (Primary AND Secondary) will continue.
    - DASD Mirroring might stop first.
      - If you have FREEZE=GO policy, both copies of structure might continue to be updated even though remote DASD are now frozen in time.

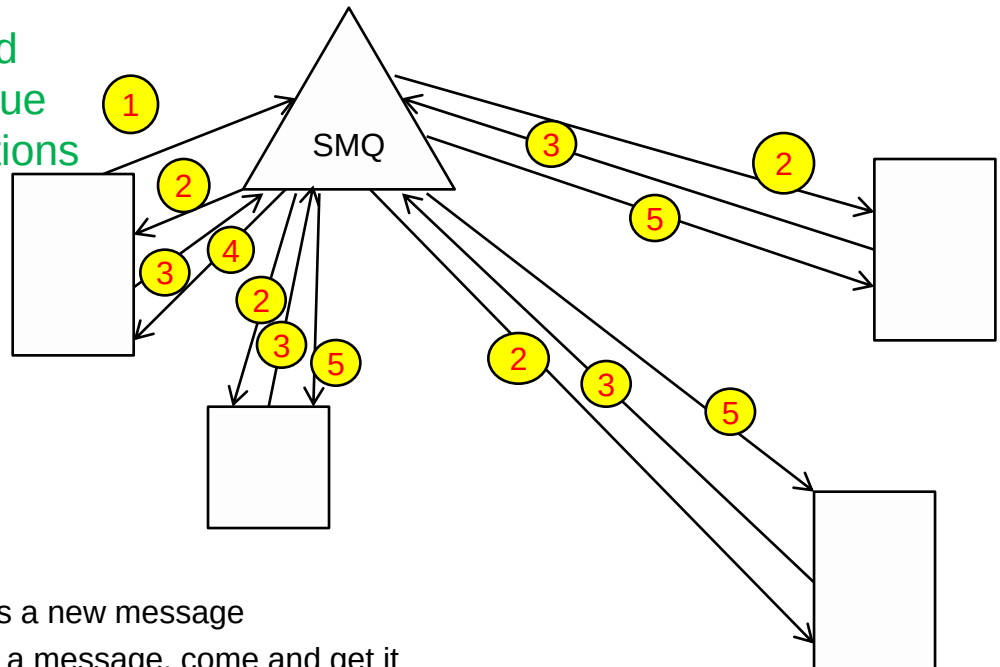
Technologies



- Path selection algorithm for Coupling Links is not published by IBM (so they have the flexibility to enhance it based on experience)
  - But if you have a large discrepancy in fiber distance for your different paths, talk to IBM and see if they can give you some guidance

Technologies

Distance and Shared Queue implementations



1. Here is a new message
2. I have a message, come and get it
3. Let me have it
4. Here it is
5. Sorry, you were too slow

SUBNOTIFYDELAY function for IMS SMQ lets you force more balance

Other devices

- Take an inventory of all devices connected to System z and determine:
  - Which ones you want in BOTH sites
  - Which will only exist in one site (and what is the backup if that site is lost)
  - Which systems will connect to each device
    - Is channel extension a consideration? If so, are there special considerations (such as parallel or ESCON connection)?

Other devices

- Some examples:
  - Tape (and consider tape mirroring)
  - Network connectivity
  - Check sorters
  - Printers
  - Consoles
  - Encryption devices

- Also need to think about CPC capabilities. If CPCs in one site have these, do the CPCs in other site have or need these?
  - zIIPs and zAAPs
  - zEDC
  - Crypto
  - STP (locations of Primary, Backup, Arbiter)

Availability considerations

- Before making commitments about levels of availability, think carefully about how various outages could impact you:
  - CPC failure
  - CF failure
  - Primary DASD failure (with and without HyperSwap)
  - Secondary DASD failure
  - Failure of mirroring links
  - Failure of entire site (from a recovery perspective, this is like having all your z/OS systems and all your CFs in that site in a single CPC)
  - Loss of connectivity between the sites
  - DWDM failure
  - Switch failure

Tools

- RMF Magic and Disk Magic (by Intellimagic) provide excellent modeling capabilities on the disk aspects of distance.
- IBM's zCP3000 Capacity Planning tool has support for modeling the impact of distance on CF response times.
- Switch vendors have tools to help identify buffer credit requirements.
- RMF Coupling Report (new Channel Path Details report) shows the length of each link (in tenths of a kilometer)!!
- But there are NO tools to model impact of distance on transaction response times and batch job elapsed times.
  - Official IBM position is that proposed configurations should be benchmarked.

Summary

- Multi-site sysplex is not a single destination – it is a huge range of configurations, varying from everything running in just one site to everything running everywhere, and every combination in between.
  - Because the underlying topology is similar for Single-site Workload and full Multi-site Workload, you can always fine tune your workload distribution over time based on your experiences:
    - You might start with Single-Site Workload and move to Multi-Site Workload for critical applications.
    - You might start with Multi-Site Workload and find the impact unacceptable for some applications and move those back to a single site.
- If done well, results can be impressive – I don't know of any company that went to multi-site sysplex and then went back to single site. There IS a cost, but they all feel that the benefits outweigh the cost.

Summary

- Vital to sit down with your executives and ensure that they understand what is possible (in terms of distance), the difference between CA and DR, and that the end result will *not* be that you will never ever have an outage again.
  - You also need to stress the difference between “supported” and “feasible”.
- IF your objective is better availability, suggest that you exhaust all single-site availability capabilities (for example, do all critical apps support data sharing and dynamic workload routing?) before you spend significant time and money on a multi-site sysplex.
  - How many of your outages would have been avoided if you just dropped your current environment onto a multi-site sysplex?
  - Even in a single site, use synchronous mirroring and HyperSwap

Summary

- If your objective is DR, what does a multi-site sysplex buy you over a “BRS” configuration?
  - It COULD buy you a lot, but only if you are configured to exploit all the possibilities.
- Read up on the IBM qualification program and use qualified devices in your configuration, unless you LIKE being in the middle of an x-way finger-pointing situation.
- Planning and consulting with subject matter experts is critical. Especially for the connectivity equipment, the technology is constantly changing so you need a true expert on your side. GDPS has been around for a LONG time, and yet there are still bugs from time to time and IBM are still learning – this stuff is NOT trivial.

Summary

- It is hard to exaggerate the importance of having NO single points of failure in the connectivity configuration.
- Don’t forget your network – having zero users logged on after a site switch is not the best way to address possible response time issues…

Reference information

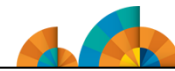
- IBM Redbook – *zEC12 IBM zEnterprise EC12 Technical Guide*, SG24-8049
- IBM Redbook - *Considerations for Multisite Sysplex Data Sharing*, SG24-7263
- IBM Redbook - *System z End-to-End Extended Distance Guide*, SG24-8047
- IBM Redbook - *Implementing and Managing InfiniBand Coupling Links on IBM System z*, SG24-7539
- IBM Redbook - *GDPS Family: An Introduction to Concepts and Capabilities*, SG24-6374
- Multiple IBM RedPapers about Qualified DWDMs – search for “IBM System z Qualified” on www.redbooks.ibm.com
- IBM’s experts in this area are in the GDPS teams

Reference information

There are a number of related other sessions that you might be interested in:

- zHA009 - *What's New in IBM GDPS 3.11?* – **Sim Schindel** – Wed 14:30
- zHA010 - *IBM GDPS/Active-Active - The Future of Continuous Application Availability on IBM System z* – **Sim Schindel** – Fri 10:30
- zSN004 - *GM, zGM, XRC, PPRC, MM, FC, CC, VCC: Understanding the Alphabet Soup of IBM Copy Services* - **Lisa Gundy** – Mon 14:30
- zSN026 - *Multi-Target Replication with IBM DS8870* - **Warren Stanley** – Tues 13:00
- zSN047 - *What's New? Gen 5 SAN Solutions with IBM's Enterprise Servers* - **Tim Jeka (Brocade)** – Tues 16:15
- zSN052 - *Moving to zEC12 or zBC12 and Need Support for Critical ESCON and Bus/Tag Devices? Optica's Prizm Makes it Easy!* - **Sean Seitz (Optica Technologies)** – Wed 09:00

Reference information



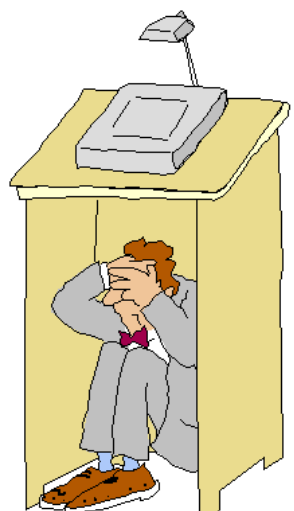
I also have a couple of other sessions that you might be interested in:

- zPE007 – *The Skinny on Coupling Thin Interrupts* - **Frank Kyne** – Wed 13:00
- zPE008 – *Why Is the CPU Time for a Job so Variable?* - **Frank Kyne** – Fri 09:00

Also, if you like SMF data (and who doesn't??!!), please see our new AND IMPROVED(!) *SMF Reference Summary* at
www.watsonwalker.com/references.html

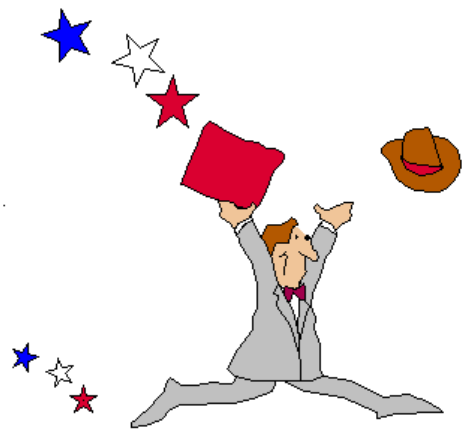
Frank Kyne
Editor and Technical Consultant
Watson and Walker

Any questions?



Frank Kyne
Editor and Technical Consultant
Watson and Walker

Thank you for coming



Please remember to complete an evaluation
Session number is zHA001

Growing your IBM skills – a new model for training

Enterprise2014

Meet the authorized IBM Global Training Providers in the Enterprise Solution Showcase

- Access to training in more cities local to you, where and when you need it, and in the format you want
  - Use IBM Training Search to locate training classes near to you
- Demanding a high standard of quality / see the paths to success
  - Learn about the New IBM Training Model and see how IBM is driving quality
  - Check Training Paths and Certifications to find the course that is right for you
- Academic Initiative works with colleges and universities to introduce real-world technology into the classroom, giving students the hands-on experience valued by employers in today's marketplace
- www.ibm.com/training is the main IBM training page for accessing our comprehensive portfolio of skills and career accelerators that are designed to meet all your training needs.

Global Skills Initiative

Avnet Academy Global Training Provider

Global Knowledge

INGRAM MICRO

LearnQuest