Server Time Protocol Recovery Considerations (STP-only CTN with 3 or more servers)

Noshir Dhondy (dhondy@us.ibm.com)

Agenda

- **STP Recovery Concepts**
  - Recovery design rules and terminology
  - Switch to Local Timing mode
- **STP-only CTN recovery (BTS and Arbiter assigned)**
  - Arbiter Assisted Recovery
  - Console Assisted Recovery
- **STP-only CTN recovery with Internal Battery Feature (IBF)**
- **Site failure scenarios**
- **GDPS/PPRC considerations**

STP-only CTN Terminology

- **CTN**
  - Collection of servers that are time synchronized to a time value called Coordinated Server Time (CST)

- **Server/CF roles**
  - **Preferred Time Server/CF (PTS)**
    - Server that is preferred to be the Stratum 1 server
  - **Backup Time Server/CF (BTS)**
    - Role is to take over as the Stratum 1 under planned or unplanned outages, without disrupting synchronization capability of STP-only CTN
  - **Current Time Server/CF(CTS)**
    - Active S1 Server/CF
      - **Only one S1 allowed**
      - Only the PTS or BTS can be assigned as the CTS
      - Normally the PTS is assigned the role of CTS – Active S1
      - BTS typically is the Inactive S1
      - BTS can take over as Active S1 or assigned Active S1 for planned actions
      - PTS is the Inactive S1 in those cases
  - **Arbiter**
    - Provides additional means to determine if BTS should take over as the CTS under unplanned outages

ETR/STP availability/recovery requirements

- **Availability**
  - When primary source of time fails, applications that depend on time synchronization can continue processing **with data integrity**.
    - Parallel Sysplex
    - GDPS customers having multi-site sysplex require Site 2 systems to continue processing when Site 1 fails and vice versa
    - z/OS Global Mirror (XRC) that uses time stamps associated with data updates to make sure secondary copy of the data is consistent
    - Non-sysplex applications that may use other than coupling links for messaging

- **ETR/STP recovery must ensure data integrity when time consistency cannot be maintained**
  - Availability can be compromised but not data integrity
  - Current designs (ETR and STP) have failure scenarios where availability is compromised, resulting in z/OS systems posting a WTOR

STP recovery design rules and overview

- **CANNOT have two Stratum 1 servers in timing network**
- **Backup Time Server (BTS) can take over as Current Time Server (CTS), active Stratum 1, only if either:**
  - Preferred Time Server (PTS) can indicate it has “failed”
    - PTS, if operational MUST surrender role of CTS
  - BTS can unambiguously determine the PTS has “failed”

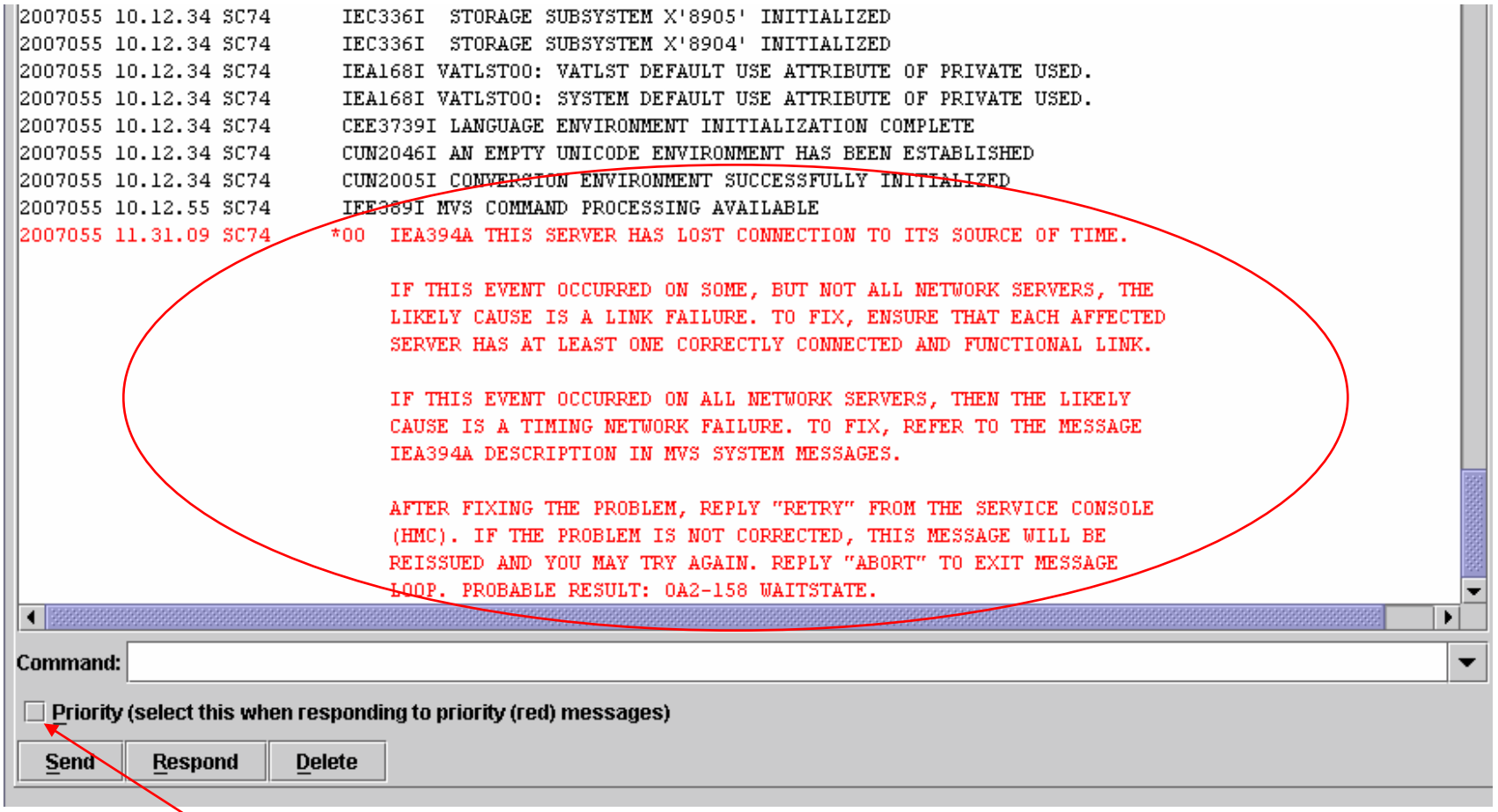
Switch to Local Timing Mode

- **Server in ETR network or CTN becomes unsynchronized (S0 in CTN):**
  - z/OS system images running in ETR or STP timing mode switch to local timing mode.
  - Impact of switching depends on
    - PLEXCFG parameter in IEASYSxx, and
    - ETRMODE or STPMODE specified in CLOCKxx.
  - z/OS systems that specify:
    - PLEXCFG=MULTISYSTEM or PLEXCFG=ANY in IEASYSxx, and
    - ETRMODE YES or STPMODE YES in CLOCKxx
  - Issue a WTOR message to allow operator intervention to resolve the problem before a wait state is loaded
    - z/OS systems that specify ETRMODE YES and are running in ETR timing mode issue WTOR message IEA015A.
    - z/OS systems that specify STPMODE YES and are running in STP timing mode issue WTOR message IEA394A.

WTOR – IEA394A

- **WTOR allows time window to correct the problem and respond *"RETRY"* if problem corrected or *"ABORT"* if problem cannot be corrected**
  - *"ABORT"* will load wait state 0A2-158
- **Backup Time Server or another operational server in the CTN can be reconfigured to be the Current Time Server (CTS) before**
  - WTOR messages responded to with *"RETRY"*
- **New function in z/OS 1.7 for SFM to recognize that WTOR IEA394A issued**

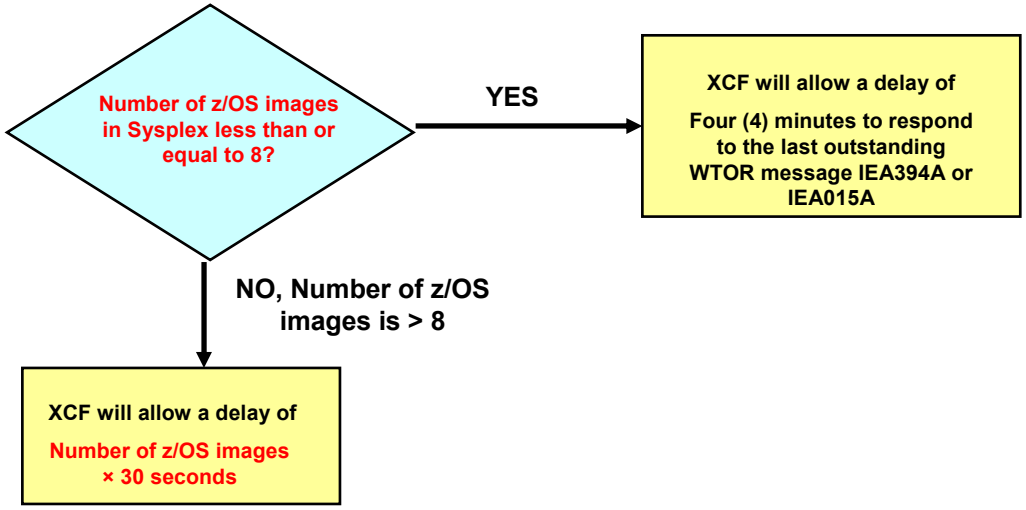
IEA394A WTOR



Important: Priority message checkbox must be selected when responding to WTOR

Sysplex Failure Management (SFM) considerations

- **SFM allows installation to code a policy to define the recovery actions to be automatically initiated following detection of a Parallel Sysplex failure.**
  - Actions include fencing off the failed image that prevents access to shared resources, logical partition deactivation, or dynamic storage reconfiguration.
- **New function in z/OS 1.7 and higher for SFM to recognize that WTOR IEA015A or IEA394A issued**
  - WTOR message issued by all the z/OS images in the sysplex, the user is not time constrained to do timing network reconfiguration before replying to IEA0394A or IEA015A.
  - Once WTOR on the first system image responded to with *"RETRY"*,



- **z/OS system images will enter disabled-wait states should the user not be able to respond to the IEA394A or IEA015A WTOR message in the allotted time**.
- If the message is issued only on a subset of participating sysplex images, the SFM settings specified in the SFM Policy must be considered

STP Recovery terminology

- **Coordinated Server Time**
  - Coordinated Server Time (CST) represents the time for the CTN and is the time at a Stratum 1 server
- **Synchronization check threshold**
  - Server/CF considered to be in synchronized state if TOD clock within synchronization check threshold of CST
  - STP synchronization check threshold 50 microseconds
  - If TOD clock differs from CST by more than +/- 50 microseconds, server/CF becomes unsynchronized
    - Can become a Stratum 0 (S0) server/CF
- **Freewheel Interval**
  - Amount of time a Stratum 2 or Stratum 3 server can remain synchronized without receiving messages from its clock source
    - Approximately 1 second (Mixed-CTN)
    - Approximately 10 seconds (STP-only CTN)

Agenda

- **STP Recovery Concepts**
  - Recovery design rules and terminology
  - Switch to Local Timing mode
- **STP-only CTN recovery (BTS and Arbiter assigned)**
  - Arbiter Assisted Recovery
  - Failure scenarios
- **STP-only CTN recovery with Internal Battery Feature (IBF)**
- **Site failure scenarios**
- **GDPS/PPRC considerations**

STP-only CTN with 3 or more servers/CFs

- **CTN has a PTS, BTS, and Arbiter assigned**
- **Assumption: PTS also assigned the CTS role**
- **CANNOT have two Stratum 1 servers in timing network**
- **Backup Time Server (BTS) can take over as Current Time Server (CTS), active Stratum 1, only if either:**
  - Preferred Time Server (PTS) can indicate it has “failed” or
  - BTS can unambiguously determine the PTS has “failed”
- **PTS, if operational MUST surrender role of CTS**
- **Arbiter Assisted Recovery used to determine if BTS can take over as CTS**

Arbiter Assisted Recovery

- Arbiter provides additional means to determine if BTS can take over as the CTS
- If BTS loses communication on all established paths to CTS
  - BTS and Arbiter communicate to establish if Arbiter also has lost communication on all established paths to CTS
- If both BTS and Arbiter cannot communicate with CTS
  - BTS takes over as CTS (S1)
- Failure also implies PTS cannot communicate with BTS and Arbiter
  - Since only 1 CTS (S1) can exist,
  - PTS initially surrenders role of CTS
    - Has to assume that BTS has taken over as CTS

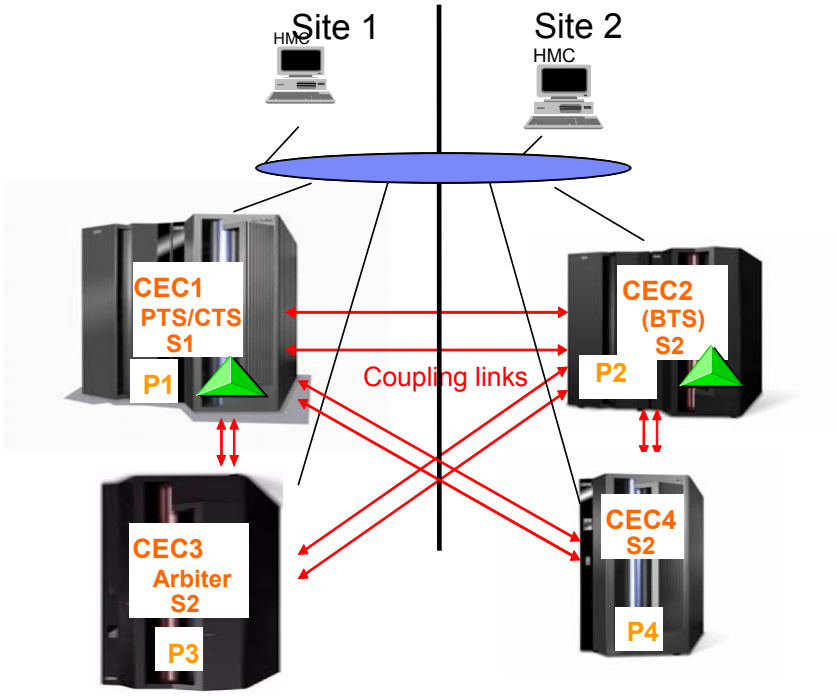
CAR – BTS and Arbiter assigned

- **Console assisted recovery uses HMC/SE LAN to determine**
  - BTS can take over as CTS (initiated by BTS case below)
  - PTS can retake role of CTS (initiated by PTS case below)
- **BTS initiates Console Assisted recovery process when:**
  - BTS has lost communication with the CTS, **and**
  - BTS cannot communicate with the Arbiter to initiate the Arbiter Assisted recovery process.
- **PTS initiates Console Assisted recovery process when:**
  - PTS has lost communication with the BTS and the Arbiter
  - PTS initially surrenders role of CTS
    - **Has to assume that BTS has taken over as CTS**
  - Needs to determine if BTS failed or operational
  - If BTS determined to have failed
    - **PTS retakes its role of CTS**
  - If BTS state good (BTS is capable of taking over as CTS) or “indeterminate”
    - PTS either becomes a S3 server if a S2 clock source is available or
    - PTS becomes unsynchronized at end of Freewheel period

Agenda

- **STP Recovery Concepts**
  - Recovery design rules and terminology
  - Switch to Local Timing mode
- **STP-only CTN recovery (BTS and Arbiter assigned)**
  - Arbiter Assisted Recovery
  - Failure scenarios
- **STP-only CTN recovery with Internal Battery Feature (IBF)**
- **Site failure scenarios**
- **GDPS/PPRC considerations**

Power outage of data center (Site 1) with PTS and Arbiter



- With Internal Battery Feature (IBF) on CEC1 and CEC3
  - Site 1 power outage
  - CEC1 and CEC3 enter IBF state
  - CEC1 and CEC3 notify CEC2 it is running on IBF
  - CEC2 waits for 30 seconds to take action
    - Could be a power glitch
    - If notified within 30 seconds that CEC1 back to “normal power”, no further action
  - If CEC1 and CEC3 in IBF state > 30 seconds,
    - CEC2 takes over as the CTS
      - Not dependent on Arbiter Assisted Recovery
    - CEC1 becomes S2 until IBF no longer functional and power drops
  - CEC1 power resumes
    - Automatic re-takeover as PTS/CTS

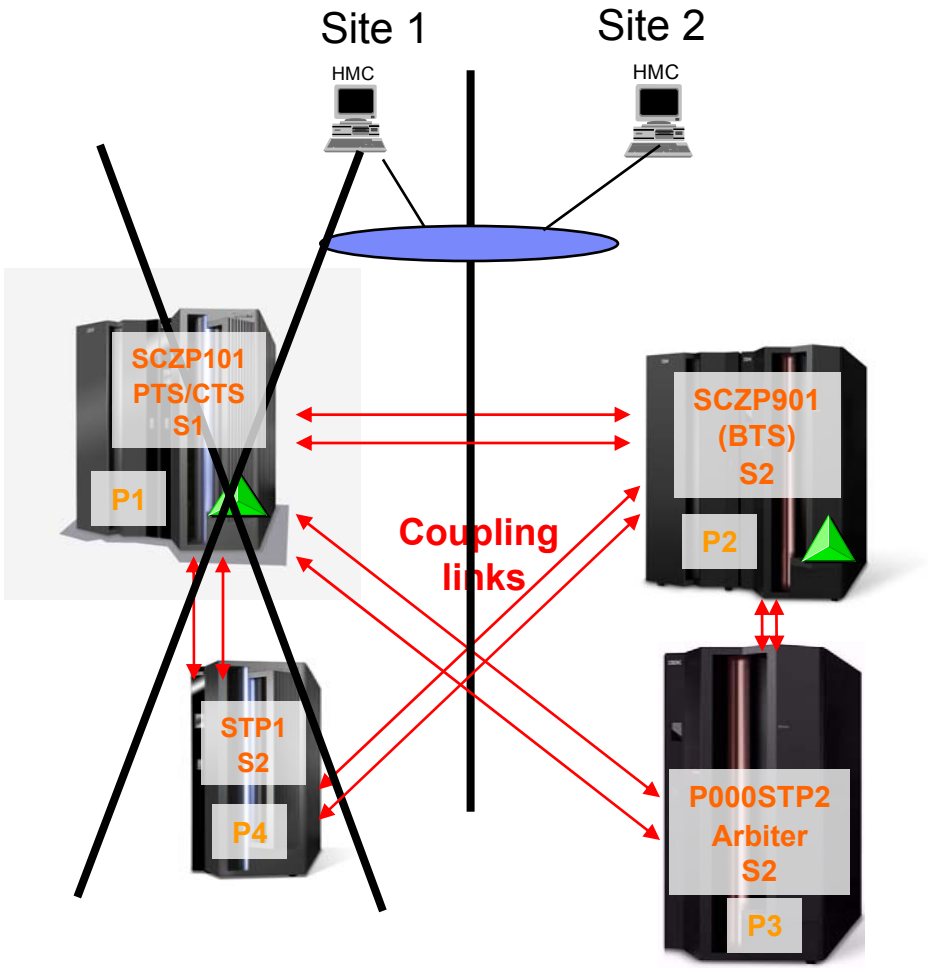
IBF Recommendations

- **Single data center**
  - IBF only protects for server power outage
  - CTN with 2 servers, install IBF on at least the PTS/CTS
    - Also recommend IBF on BTS to provide recovery protection when BTS is the CTS
  - CTN with 3 or more servers IBF not required to recover from CTS power outage, if Arbiter configured
- **Two data centers**
  - IBF protects for both server and site power outage scenarios
  - CTN with 2 servers (one in each data center) install IBF on at least the PTS/CTS
    - Also recommend IBF on BTS to provide recovery protection when BTS is the CTS
  - CTN with 3 or more servers, install IBF on CTS and Arbiter (in same site as CTS)
    - Also recommend IBF on BTS to provide recovery protection when BTS is the CTS

Agenda

- **STP Recovery Concepts**
  - Recovery design rules and terminology
  - Switch to Local Timing mode
- **STP-only CTN recovery (BTS and Arbiter assigned)**
  - Arbiter Assisted Recovery
  - Failure scenarios
- **STP-only CTN recovery with Internal Battery Feature (IBF)**
- **Site failure scenarios**
- **GDPS/PPRC considerations**

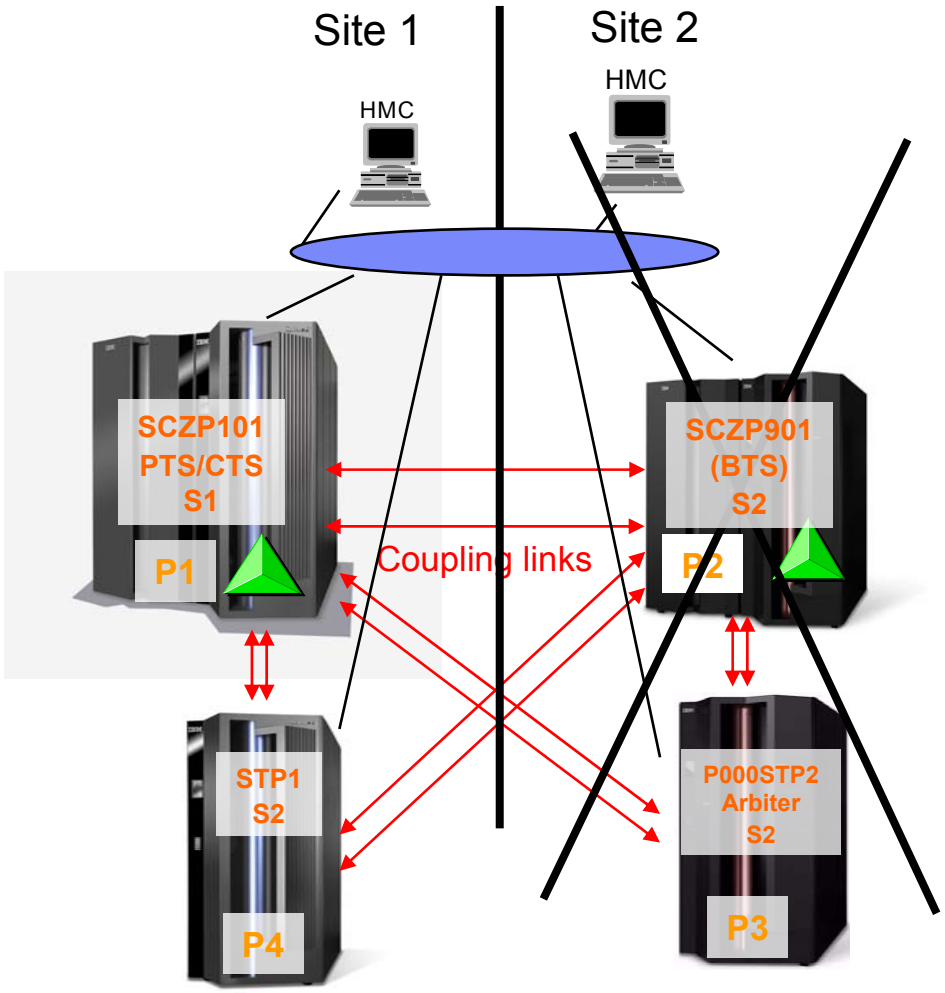
STP-only CTN (Preferred, Backup, and Arbiter assigned) Site 1 Failure – Arbiter in same site as BTS



CTNID=ITSOPOK -
P1, P2, P3, P4 in Parallel Sysplex

- **BTS (SCZP901) loses all communication with CTS (SCZP101)**
- **BTS and Arbiter communicate to establish if Arbiter also cannot communicate with CTS**
  - Both cannot communicate
- **BTS takes over as CTS (S1)**
- **z/OS systems in Site 2 requiring STPMODE YES not affected**
- **STP User actions**
  - None

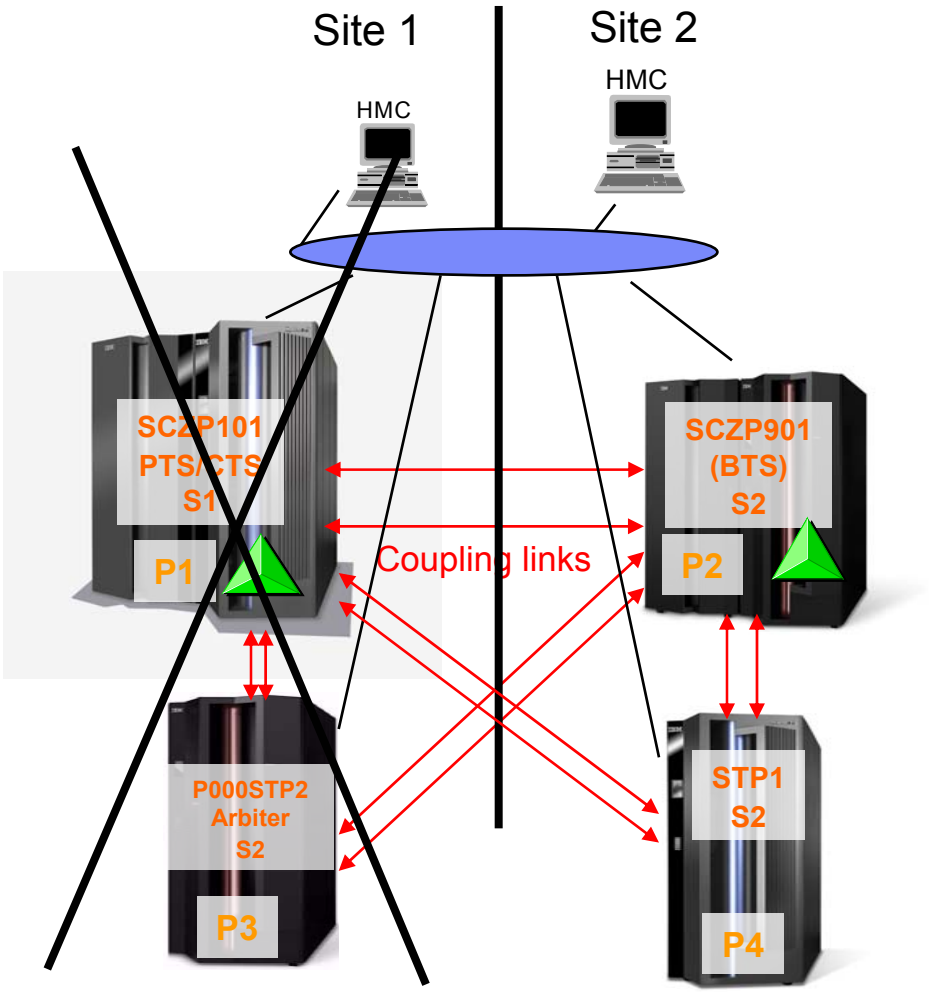
STP-only CTN (Preferred, Backup, and Arbiter assigned) Site 2 Failure – Arbiter in same site as BTS



P1, P2, P3, P4 in Parallel Sysplex

- **PTS/CTS (SCZP101) loses communication with both BTS and Arbiter**
- **PTS surrenders role of CTS**
- **PTS initiates "Console assisted recovery" to determine if BTS failed or operational**
- **Results of "Console assisted recovery"**
  - BTS state most probably indeterminate
- **PTS CANNOT retake role of CTS**
- **All z/OS systems in site 1 post WTOR (IEA394A)**
- **STP User actions**
  - Reassign PTS as CTS
  - Respond with Retry to WTOR

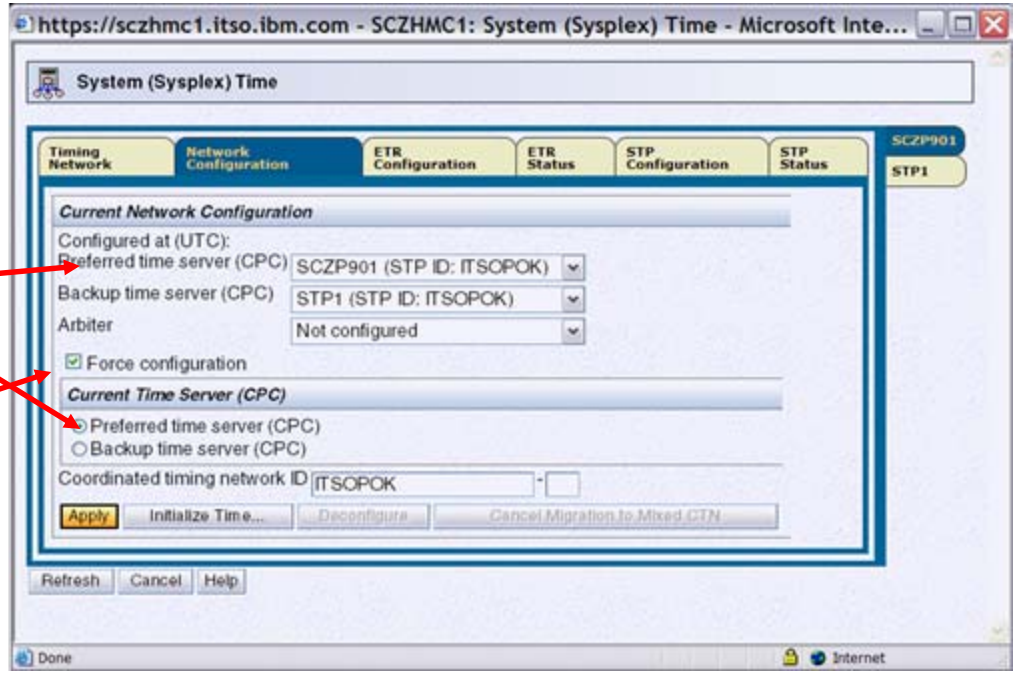
STP-only CTN (Preferred, Backup, and Arbiter assigned) Site 1 Failure – Arbiter located in same site as PTS/CTS



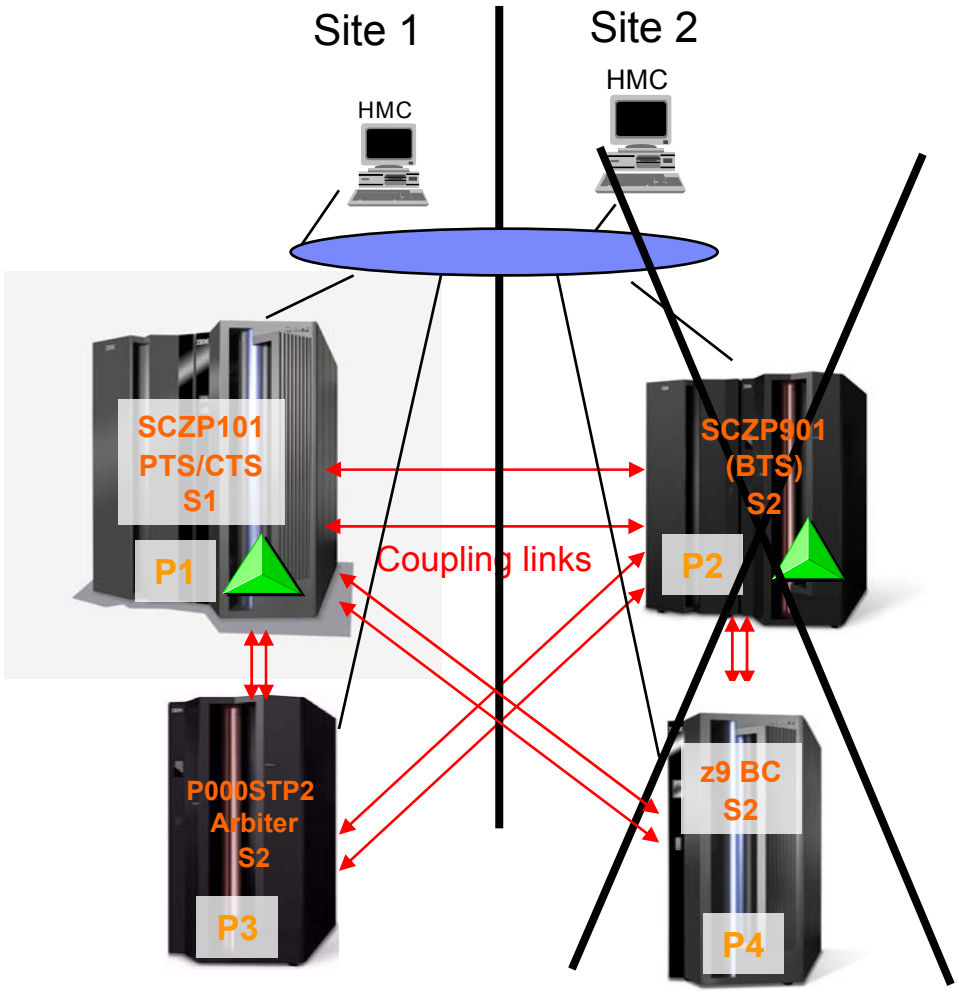
- **BTS loses all communication with CTS**
- **BTS cannot communicate with Arbiter**
- **BTS initiates "Console assisted recovery"**
- **Results of "Console assisted recovery"**
  - CTS state most probably indeterminate
- **BTS CANNOT take over as S1**
- **BTS eventually becomes unsynchronized**
  - z/OS systems (STPMODE YES) in site 2 post WTOR (IEA394A)
    - Similar to case with only PTS and BTS assigned
- **STP User actions**
  - Reassign BTS as CTS
  - Respond with Retry to WTOR

STP-only CTN (Preferred, Backup, and Arbiter assigned) - Reconfiguration after Site 1 Failure

- **Select System (Sysplex) Time task of SCZP901**
  - Server that will become the new CTS after reconfiguration
- **Select Network Configuration tab**
- **Assign SCZP901 as PTS and CTS**
- **Assign STP1 as BTS**
- **Select "Force configuration"**
  - Since starting from Stratum 0
- **Respond "Retry" to each WTOR (IEA394A) posted**
  - Note that after responding to the first WTOR, the remaining WTORs in the Sysplex have to be responded to within approximately 4 minutes if up to 8 z/OS images (additional 30 secs per image if more than 8 images)



STP-only CTN (Preferred, Backup, and Arbiter assigned) Site 2 Failure – Arbiter located in same site as PTS/CTS



- **CTS loses communication with only the BTS**
- **CTS maintains communication with Arbiter**
- **PTS maintains role of CTS (S1)**
- **STP-only CTN servers in Site 1 stay synchronized to CTS (S1)**
- **z/OS systems in Site 1 requiring STPMODE YES not affected**
- **STP User actions**
  - None

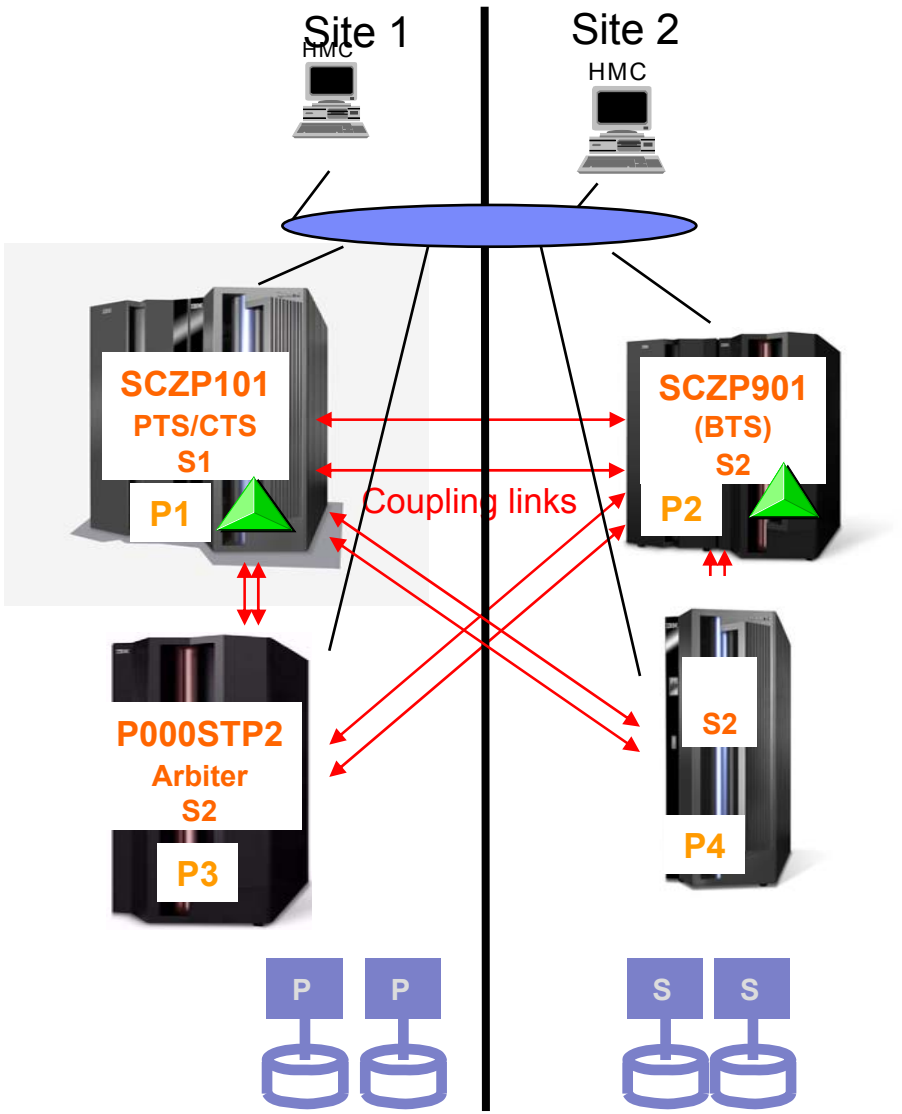
Agenda

- **STP Recovery Concepts**
  - Recovery design rules and terminology
  - Switch to Local Timing mode
- **STP-only CTN recovery (BTS and Arbiter assigned)**
  - Arbiter Assisted Recovery
  - Failure scenarios
- **STP-only CTN recovery with Internal Battery Feature (IBF)**
- **Site failure scenarios**
- **GDPS/PPRC considerations**

Improved GDPS Controlling System (K System) availability - applies to both STP and ETR

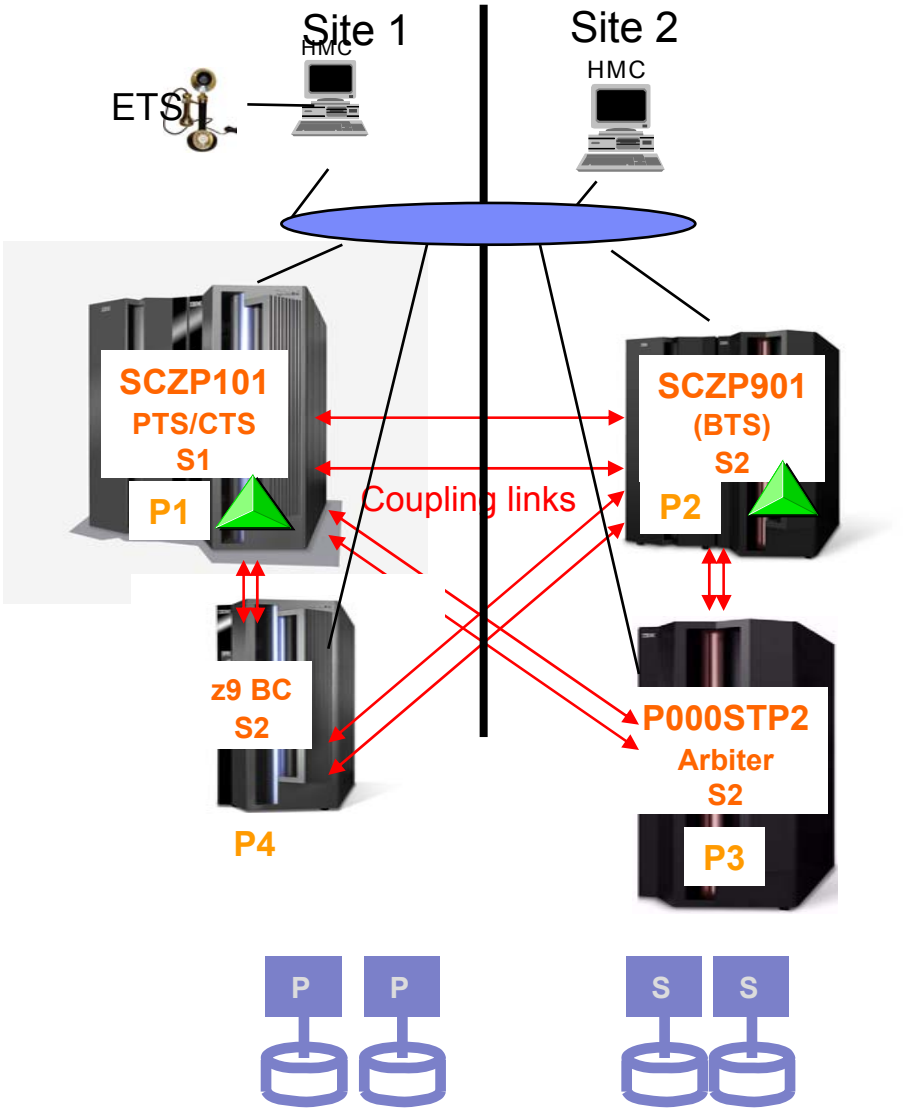
- The normal course of action when a system loses time synchronization is to either issue a synchronous WTOR or to enter a wait state.
- Most important role of a GDPS K system is to manage recovery in case of a failure:
  - If the GDPS K system is stopped it can't drive the recovery processing
- The need for time synchronization is to protect integrity of the shared data.
  - However, because the K system should never access any data that is used by the managed systems, it should be OK for the K system to run in local time mode for a short time – 80 minutes.
- The intent is that the K system will be available to complete any Freeze or HyperSwap actions it has initiated and to help you check the status of sysplex resources.
- Requires:
  - GDPS 3.6
  - z/OS 1.11 or APARs OA28323 and OA26085 on z/OS 1.9 and 1.10
  - XCF APAR OA32236 to disable SFM actions on K-sys for 80 minutes
    - Else, K-sys continues to perform normal SFM actions
  - GDPS APAR PM10041 to automate response to WTOR IEA394A once K-sys detected in local timing mode
    - Planned release date August 31, 2010

STP-only CTN (Preferred, Backup, and Arbiter assigned) Freeze=GO or SWAP,GO



- Recommend
  - PTS/CTS and Arbiter in site with primary disk
  - BTS in site with secondary disk
    - K-sys on BTS
- Site 1 Failure
  - z/OS systems in site 2 post WTOR
  - K-sys goes to local timing mode
  - Reconfigure BTS to be CTS
  - Respond “Retry” to z/OS messages
- Site 2 Failure
  - z/OS systems in site 1 not affected

STP-only CTN (Preferred, Backup, and Arbiter assigned)
Freeze=STOP or SWAP,STOP



- Recommend
  - PTS/CTS in site with primary disk
  - BTS and Arbiter in site with secondary disk
    - K-sys on BTS
- Site 1 Failure
  - BTS takes over as CTS
  - All systems in sysplex reset
  - No STP CTN reconfiguration required since K-sys still up
- Site 2 Failure
  - PTS has to surrender role of CTS
  - All servers in CTN become unsynchronized (S0)
    - Post WTOR (IEA394A)
  - All systems in sysplex reset
  - CTN reconfiguration required prior to responding “Retry” to WTORs

Summary

Summary – STP-only CTN

- Configure for link redundancy
- Initialize configuration with the PTS assigned as the Current Time Server
  - PTS, CTS must be assigned
- Assign at least a Backup Time Server
  - Can take over as CTS - active S1
- If 3 or more servers in CTN, assign BTS and Arbiter
- For configuration across 2 sites
  - Provide redundant routes for fiber links between sites
  - Use only qualified DWDMs
  - Locate the Arbiter in same site as CTS
- Install XCF and GDPS related APARs

Additional Information

- **Redbooks®**
  - Server Time Protocol Planning Guide SG24-7280
  - Server Time Protocol Implementation Guide SG24-7281
  - Server Time Protocol Recovery Guide SG24-7380
- **Education**
  - Introduction to Server Time Protocol (STP)
    - Available on Resource Link™
    - www.ibm.com/servers/resourcelink/hom03010.nsf?OpenDatabase
- **STP Web site**
  - www.ibm.com/systems/z/pso/stp.html
- **Systems Assurance**
  - The IBM team is required to complete a Systems Assurance Review (SAPR Guide SA06-012) and to complete the Systems Assurance Confirmation Form via Resource Link
- **Techdocs and WSC Flashes**
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/Web/Techdocs
    - Search on “STP”

Reference Material - Terminology

APAR	Authorized Program Analysis Report
ARB	Arbiter
BTS	Backup Time Server
CF	Coupling Facility
CTS	Current Time Server
CTN	Coordinated Timing Network
DWDM	Dense Wave Division Multiplexer
ETR	External Time Reference
ETS	External Time Source
FC	Feature Code
HMC	Hardware Management Console
HCA	Host Channel Adapter
ICB	Integrated Cluster Bus
IPL	Initial Program Load
ISC	InterSystem Coupling Channel
LAN	Local Area Network
LIC	Licensed Internal Code
LPAR	Logically Partition
NTP	Network Time Protocol
PR/SM	Processor Resource / Systems Manager
PSIFB Infiniband	Parallel Sysplex
PTF	Temporary Program Fix
PTS	Preferred Time Server
SW and	Software (programs operating systems)
SE	Support Element
TPF	Operating System
UTC	Coordinated Universal Time
zVM	Operating System
zVSE	Operating System
z/OS	Operating System
z/VM	Operating System

Questions?

Mercés

Catalan

Dank u

Dutch

Спасибо

Russian

Gracias

Spanish

Merci

French

شكراً

Arabic

감사합니다

Korean

Tack så mycket

Swedish

धन्यवाद

Hindi

תודה רבה

Hebrew

Obrigado

Brazilian Portuguese

谢谢

Chinese

Dankon

Esperanto

Thank You

ありがとうございます

Japanese

Trugarez

Breton

Tak

Danish

Danke

German

Grazie

Italian

நன்றி

Tamil

děkuji

Czech

ขอบคุณ

Thai

go raibh maith agat

Gaelic

Trademarks

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

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA,

WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, System z10®, BladeCenter®

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.