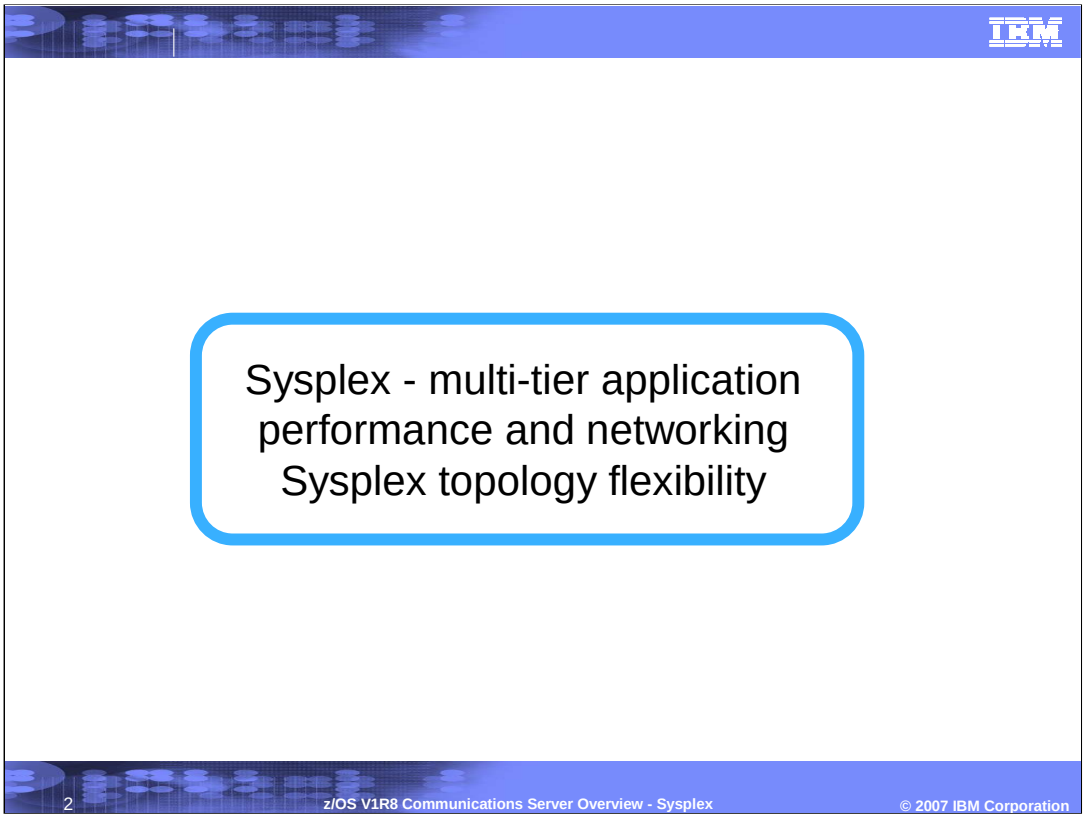
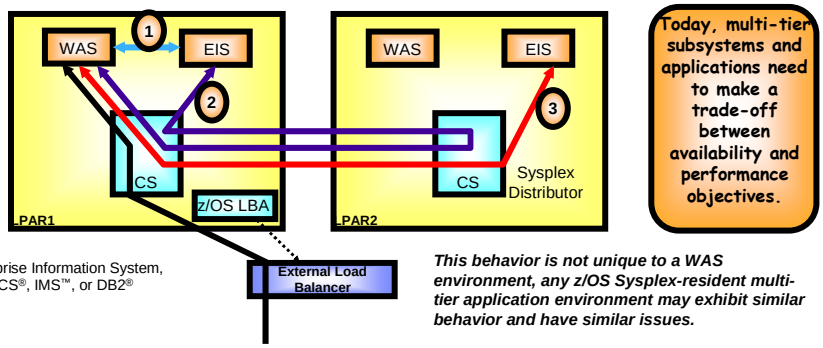
IBM eServer™

z/OS® V1R8 Communications Server Overview - Sysplex

ebusiness on demand software


IBM
Sysplex - multi-tier application performance and networking
Sysplex topology flexibility
2
z/OS V1R8 Communications Server Overview - Sysplex
© 2007 IBM Corporation

Local vs. remote connector support in today's z/OS environment



Today, multi-tier subsystems and applications need to make a trade-off between availability and performance objectives.

EIS: Enterprise Information System, such as CICS®, IMS™, or DB2®

This behavior is not unique to a WAS environment, any z/OS Sysplex-resident multi-tier application environment may exhibit similar behavior and have similar issues.

- **Local connectors (1)**
  - Optimized high-speed path (based on local services, such as cross-memory services and RRS)
  - Availability of local target of concern (no automatic switch to target on other LPAR if local is unavailable)
  - If local target becomes unavailable, WAS transactions may complete fast and WLM may in that scenario consider the LPAR a good candidate for increased workload (storm-drain issue)

Availability?

- **Remote connectors (2 and 3)**
  - Uses TCP/IP for communication
  - Sysplex Distributor (or other load balancer) selects a target among any available targets in the Sysplex
  - If target is local and Sysplex Distributor is remote, communication path is not efficient (2)
  - It is not today possible to favor a local target even if one exists and has capacity

End-to-end performance?

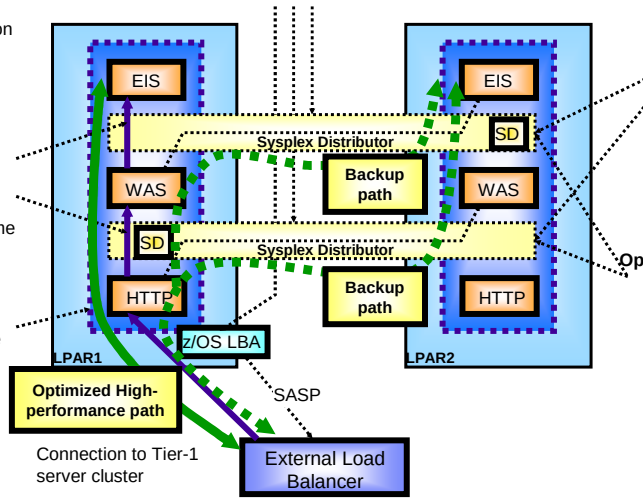
Improved multi-tier application support by Sysplex Distributor - optimized for local performance without loosing availability

Application endpoint awareness via enhanced Sysplex sockets API processing
- Avoid authentication overhead
- Avoid data conversions

Fast direct local sockets path inside the same "tower" (inside the same TCP/IP stack)

Server instances within same "tower" are **preferred targets**

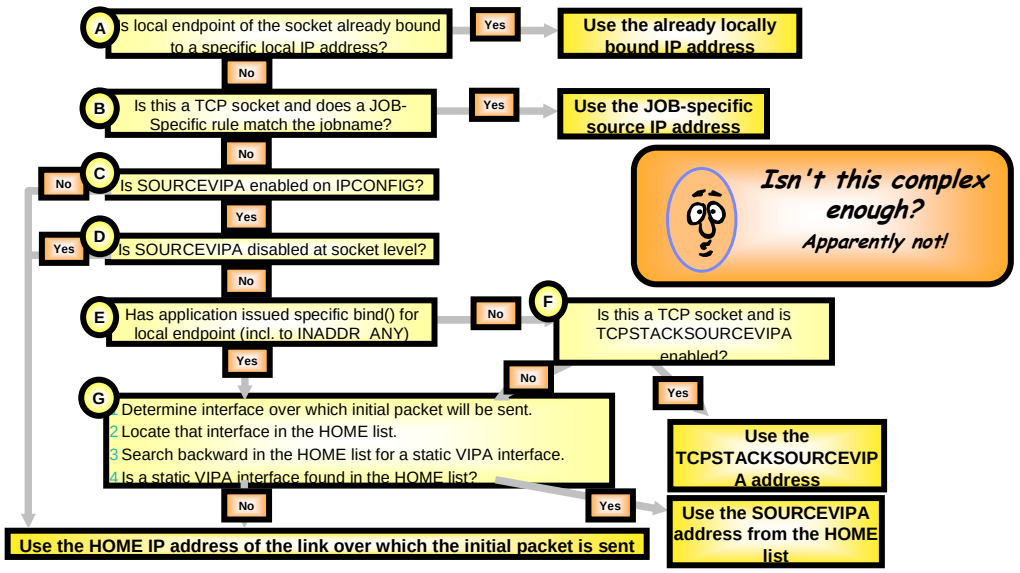
1 WLM LPAR and server-specific performance weights
2 TCP/IP stack server-specific health weights



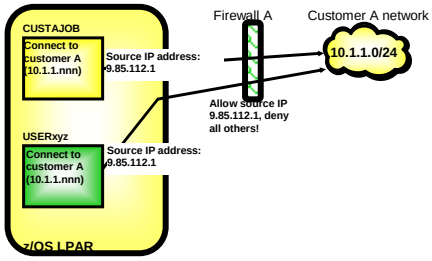
Level of **local favoritism** can be configured
- Always choose local target if target is available and healthy
- Control level of WLM weight impact on target selection

Optimized traffic flow:
- "Distributed" Sysplex Distributor logic in each stack avoids cross-LPAR flows for connection setup when local target is chosen.
- Avoids traffic routing via SD-owning LPAR for local targets

Selecting source IP address for outbound IPv4 connections or associations in CS z/OS V1R6



Adding support for yet another popular requirement: destination-based source IP address selection

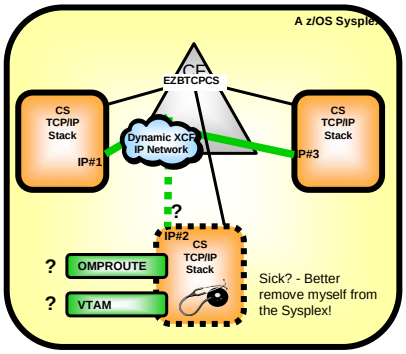



```
SRCIP
     Jobname CUSTAJOB    9.85.112.1
     Jobname CUSTBJOB    9.85.113.1
     Jobname User1*      888:555::222    ===> Wildcards allowed!
     DESTIP              10.1.1.0/24 9.85.112.1
ENDSRCIP
```

Extending configuration control over which local IP address to use for outbound connections from z/OS

- ✓Outbound connections can use same IP addresses as inbound connections to same application without application change:
  - Easier for accounting and management
  - Easier for security (firewall admin)
  - Permits source IP address selection controls for applications even when application doesn't provide for this programmatically (most don't, but some do!)
- ✓Introduced job-specific source IP addressing in z/OS V1R6
  - A new TCPIP.Profile statement SRCIP/ENDSRCIP allows the selection of a source IP address for outbound TCP connections by job name
  - Overrides TCPSTACKSOURCEVIPA and SOURCEVIPA specifications
- ✓**Introduces destination-based source IP address selection in z/OS V1R8**
  - **Extends the SRCIP/ENDSRCIP block with destination IP address-based rules**
  - **The source IP address used by a DESTIP rule cannot be a distributed DVIPA**
  - **Useful if jobnames are unpredictable or if the same jobname establishes connections to multiple business partners**

TCP/IP Sysplex autonomics in z/OS V1R6 and V1R7 reacts to and recovers dynamically from a range of error conditions



The assumption is that if a TCP/IP stack determines it can no longer perform its Sysplex functions correctly, it is better for it to leave the TCP/IP XCF group and by doing so, signal the other TCP/IP stacks in the Sysplex that they are to initiate whatever recovery actions have been defined, such as moving dynamic VIPA addresses or removing application instances from distributed application groups.

- Autonomic functions to reduce single point of failure for distributed applications in a sysplex
  - Monitor CS health indicators
    - Storage usage - CSM, TCPIP Private & ECSA
  - Monitor dependent networking functions
    - OMPROUTE availability
    - VTAM® availability
    - XCF links available
  - Monitor Communications Server component-specific functions
- Monitors determine if this TCPIP stack will remove itself from the sysplex and allow a healthy backup to take ownership of the sysplex duties (own DVIPAs, distribute workload)
- Monitoring is always done, but configuration controls in the TCPIP Profile determine if the TCPIP stack will remove itself from the sysplex.

```
GLOBALCONFIG SYSPLEXMONITOR TIMERSECS
seconds RECOVERY|NORECOVERY
DELAYJOIN|NODELAYJOIN
AUTOREJOIN|NOAUTOREJOIN
```

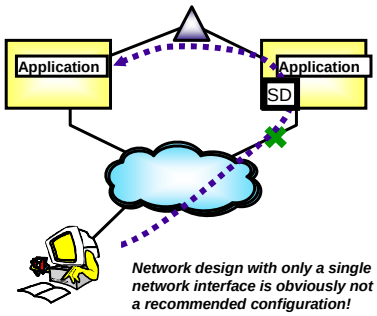
- *Timersecs* - used to determine duration of the troubling condition before issuing messages or leaving the sysplex (if Recovery)
- *RECOVERY* - TCPIP removes itself from the sysplex.
- *NORECOVERY* - TCPIP does not remove itself from the sysplex.
- *DELAYJOIN* - Delay joining Sysplex until OMPROUTE is up
- *NODELAYJOIN* - Join Sysplex immediately
- *AUTOREJOIN* - Rejoin when condition is cleared
- *NOAUTOPREJOIN* - Let an operator decide when to rejoin

Messages are always issued to the console when these conditions are detected regardless of SYSPLEXMONITOR Recovery specification
Messages are eventual action (deleted when the action is taken or problem is resolved)

New operator command is provided to allow TCPIP to leave the sysplex (ie. EZBTCPCS xcf group)
Vary TCPIP,,SYSPLEX,LEAVEGROUP

To have TCPIP rejoin the sysplex group, a Vary Obey of the TCPIP profile with sysplex configuration statements is needed.
Severe problems may require a TCPIP stack restart

TCP/IP Sysplex autonomics adds automated recovery from network outage conditions



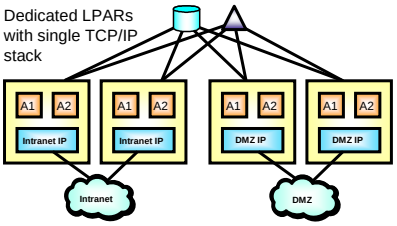
Network design with only a single network interface is obviously not a recommended configuration!

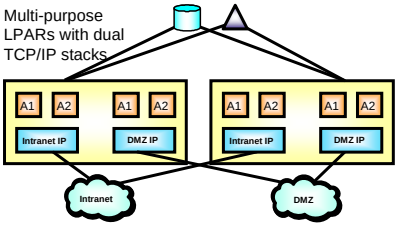
- **Assume that DynamicXCF is not an OSPF interface or that we have disabled routing through z/OS in general (NODATAGRAMFWD):**
  - Assume the downstream nodes cannot reach the SD node
    - OSA failure
    - First hop router (downstream) problems
  - All Sysplex health monitors indicate a healthy environment
    - Dynamic XCF connectivity is working
    - No storage issues
    - VTAM is operational
    - OMPROUTE is operational
    - Target Server Responsiveness Fraction indicates no SD environment health problems
  - But since there is no route from the client into the SD node, the SD functions appears unavailable

- **Network outage detection added to the Sysplex autonomics of TCP/IP**
  - Specify which network interfaces to be monitored
  - Monitor network interface itself (active or inactive)
    - To detect interface hardware issues
  - If dynamic routing is used, monitor if dynamic routes exist over the interface
    - To detect first-hop router issues
  - DELAYJOIN extended to monitor for interfaces up and dynamic routes detected

If a network outage condition is detected, the stack may remove itself from the Sysplex if requested by the configuration - allowing backup stacks in the Sysplex to take over its Sysplex responsibilities.

z/OS Sysplex connectivity to multiple security areas has been an issue every since CS began using Sysplex capabilities

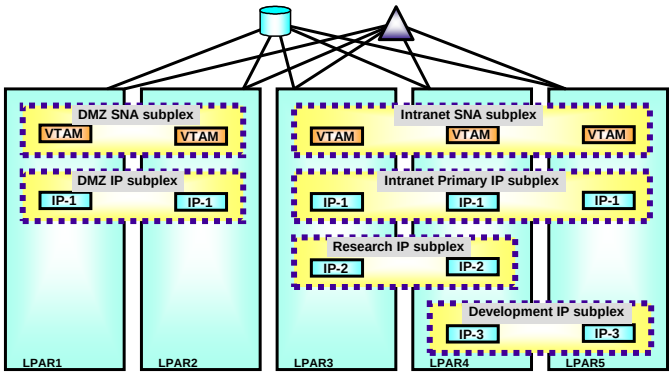




- **How to control level of automatic connectivity**
  - XCF signalling (group name) - both IP and SNA
  - IUTSAMEH (same host IP links inside an LPAR)
  - HiperSockets® (as enabled by IQDCHPID in VTAM)
- **How to control level of IP and SNA resource awareness**
  - Dynamic IP address discovery across the Sysplex
  - VTAM generic resource and MNPS resource scope spans the full Sysplex
- **How to control scope of IP workload balancing using Sysplex Distributor**
  - SD requires Dynamic XCF to be enabled, and Dynamic XCF will establish automatic IP connectivity to all stacks in the Sysplex that also have Dynamic XCF enabled

To support environments such as these, installations typically end up implementing complex resource controls and disabling many of the dynamic networking functions that are provided by TCP/IP and VTAM.

Enable use of networking Sysplex functions in a Sysplex that is connected to multiple security areas

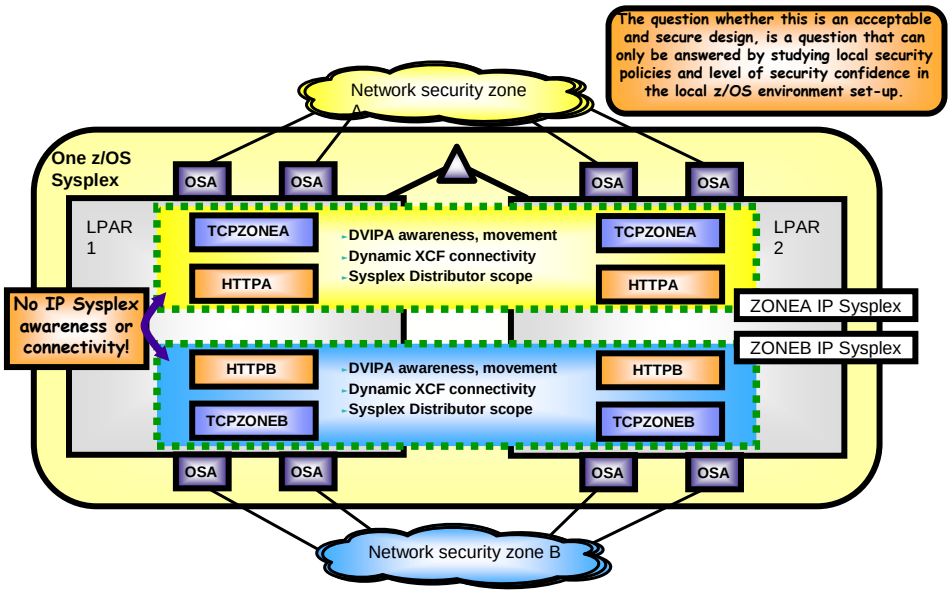


Networking Subplexing - Sysplex partitioning from a network perspective

- **Networking subplex scope:**
  - VTAM Generic Resources (GR) and Multi-Node Persistent Session (MNPS) resources
  - Automatic connectivity - IP connectivity and VTAM connectivity over XCF (including dynamic IUTSAMEH and dynamic HiperSockets based on Dynamic XCF for IP)
  - IP stack IP address (including dynamic VIPA) awareness and visibility
  - Dynamic VIPA movement candidates
  - Sysplex Distributor target candidates

- One SNA subplex per LPAR
- An IP subplex cannot span multiple SNA subplexes
- Different IP stacks in an LPAR may belong to different IP subplexes
- Standard RACF® controls for stack access and application access to z/OS resources need to be in place.

Example of LPARs connected to multiple security zones

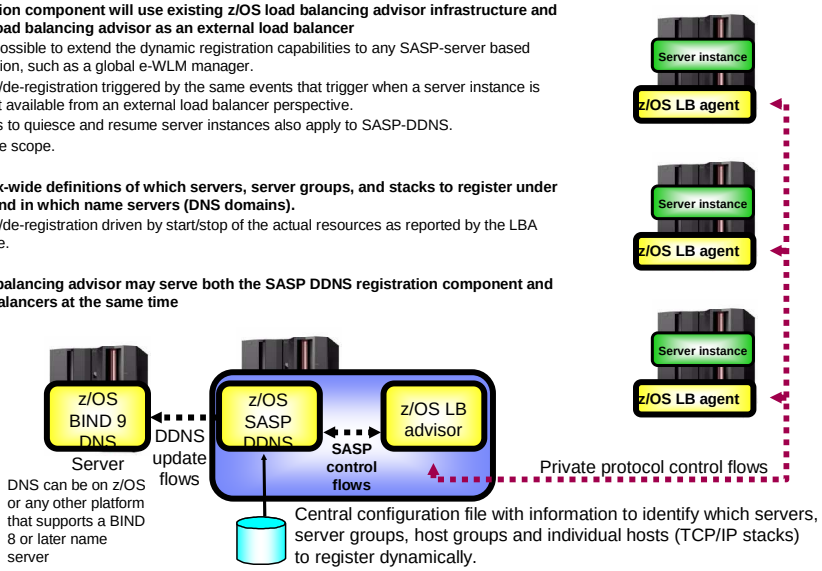


DNS/WLM - going away or not going away or what ?

- **DNS/WLM implemented two distinct functions:**
  - Dynamic name registration of servers, server groups, and TCP/IP stacks
  - Workload balancing based on name resolution requests and interaction with WLM
- **WLM-based TCP/IP workload balancing into a z/OS Sysplex is today better handled by more modern technologies, such as Sysplex Distributor or external load balancers using the z/OS load balancing advisor technology:**
  - Less overhead - balancing at connection set up time and not at name resolution time
  - Not sensitive to DNS caching
  - Better load balancing decisions - the new technologies have more metrics available than DNS/WLM had
- **However, the dynamic name registration capabilities of DNS/WLM are still very useful from an availability perspective and are not replaced by any of the currently available alternative load balancing technologies:**
  - Dynamic registration of individual application instances when they start up
  - Dynamic registration of groups of application instances when they start up
  - Dynamic registration of TCP/IP stacks when they start up
- **General dynamic registration in modern DNS servers (BIND 8 or later) is supported by a set of DNS protocols that are known as Dynamic DNS (DDNS)**
  - CS z/OS V1R8 will implement a new infrastructure that will support DDNS registration of the same type of entries that were supported by DNS/WLM
  - DDNS is a standard protocol
  - Any DDNS capable name server can be the target of the DDNS registrations

Replacing the dynamic DNS registration part of the DNS/WLM component with a Dynamic DNS (DDNS-based) solution

- **DDNS registration component will use existing z/OS load balancing advisor infrastructure and appear to the load balancing advisor as an external load balancer**
  - Potentially possible to extend the dynamic registration capabilities to any SASP-server based implementation, such as a global e-WLM manager.
  - Registration/de-registration triggered by the same events that trigger when a server instance is available/not available from an external load balancer perspective.
  - LBA controls to quiesce and resume server instances also apply to SASP-DDNS.
  - Sysplex-wide scope.
- **Central Sysplex-wide definitions of which servers, server groups, and stacks to register under which names and in which name servers (DNS domains).**
  - Registration/de-registration driven by start/stop of the actual resources as reported by the LBA infrastructure.
- **The z/OS load balancing advisor may serve both the SASP DDNS registration component and external load balancers at the same time**



Trademarks, copyrights, and disclaimers

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

CICS HiperSockets IMS RACF VTAM z/OS

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements or changes in the products or programs described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. 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. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

Information is provided "AS IS" without warranty of any kind. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (for example, IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. 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 in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products.

IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance 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 or performance improvements equivalent to the ratios stated here.