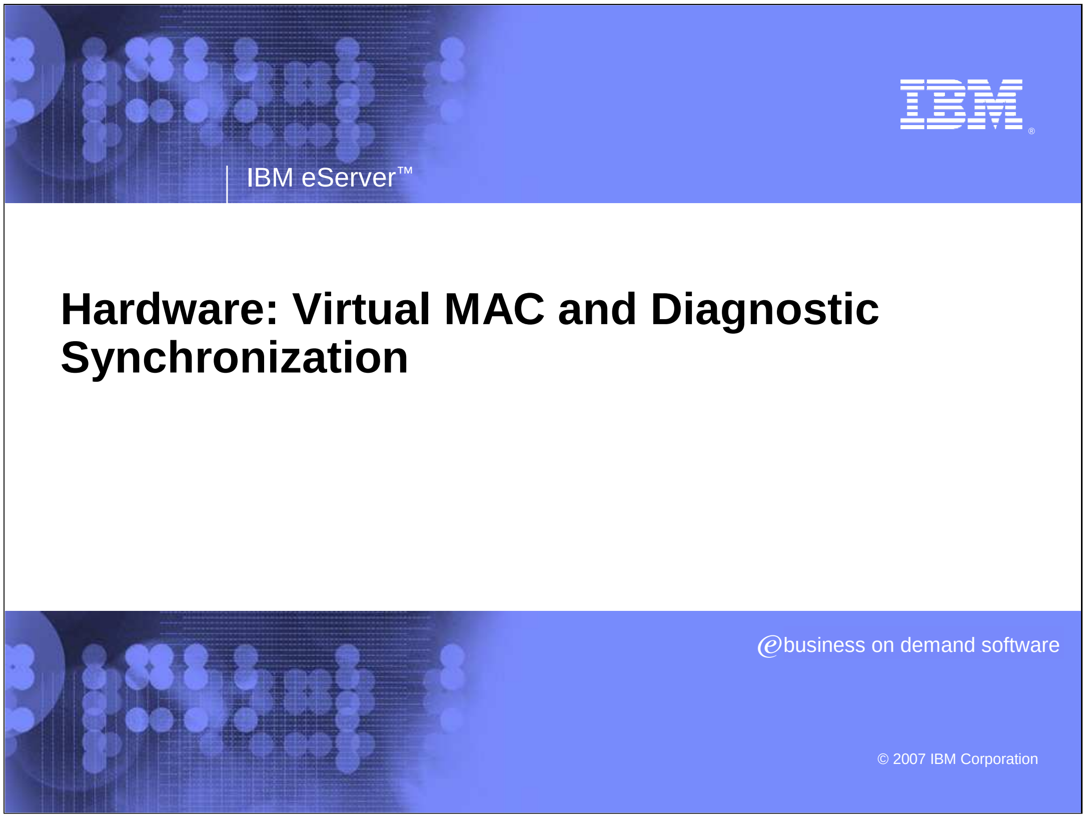
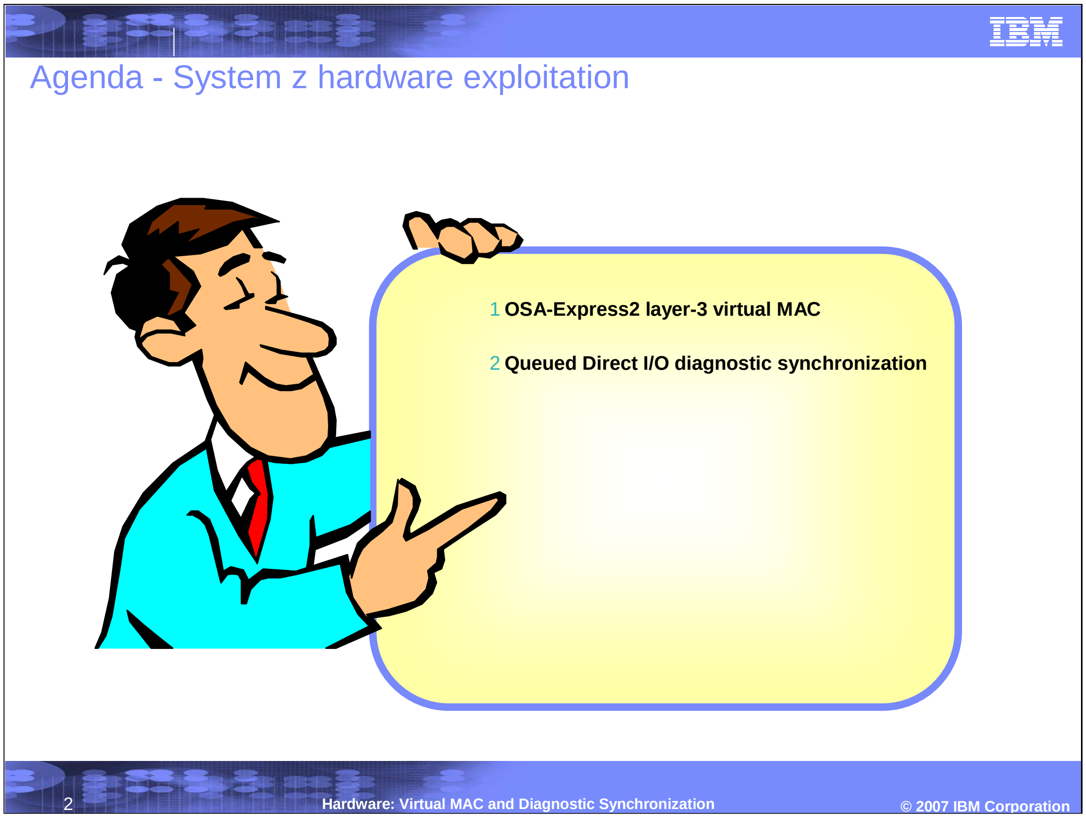


IBM eServer™

Hardware: Virtual MAC and Diagnostic Synchronization

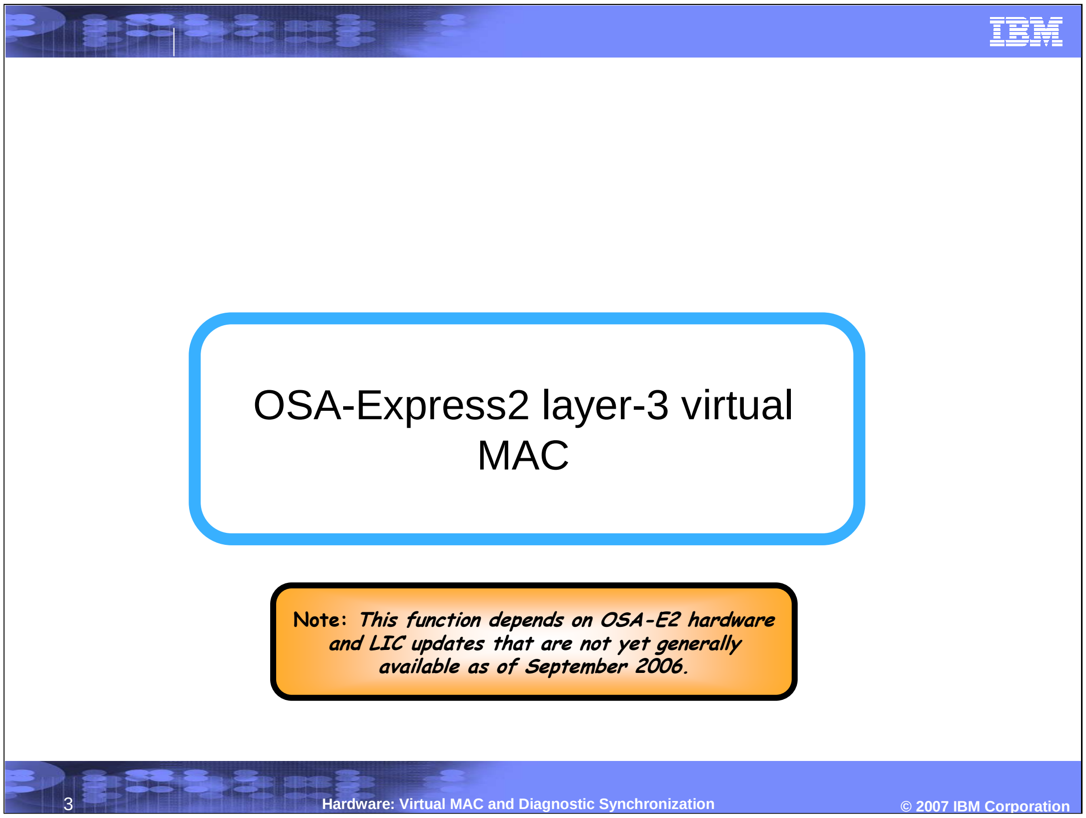
ebusiness on demand software



Agenda - System z hardware exploitation

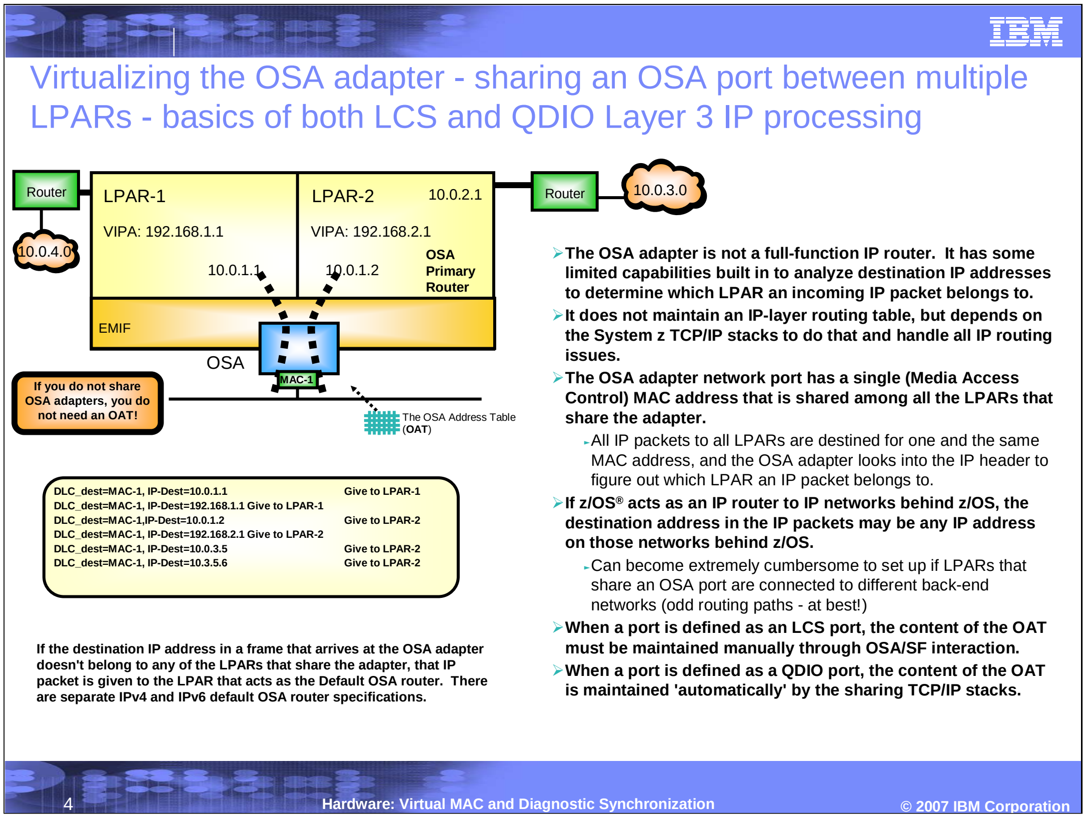
1 OSA-Express2 layer-3 virtual MAC

2 Queued Direct I/O diagnostic synchronization



OSA-Express2 layer-3 virtual MAC

Note: *This function depends on OSA-E2 hardware and LIC updates that are not yet generally available as of September 2006.*



Virtualizing the OSA adapter - sharing an OSA port between multiple LPARs - basics of both LCS and QDIO Layer 3 IP processing

If you do not share OSA adapters, you do not need an OAT!

DLC_dest=MAC-1, IP-Dest=10.0.1.1	Give to LPAR-1
DLC_dest=MAC-1, IP-Dest=192.168.1.1 Give to LPAR-1	
DLC_dest=MAC-1,IP-Dest=10.0.1.2	Give to LPAR-2
DLC_dest=MAC-1, IP-Dest=192.168.2.1 Give to LPAR-2	
DLC_dest=MAC-1, IP-Dest=10.0.3.5	Give to LPAR-2
DLC_dest=MAC-1, IP-Dest=10.3.5.6	Give to LPAR-2

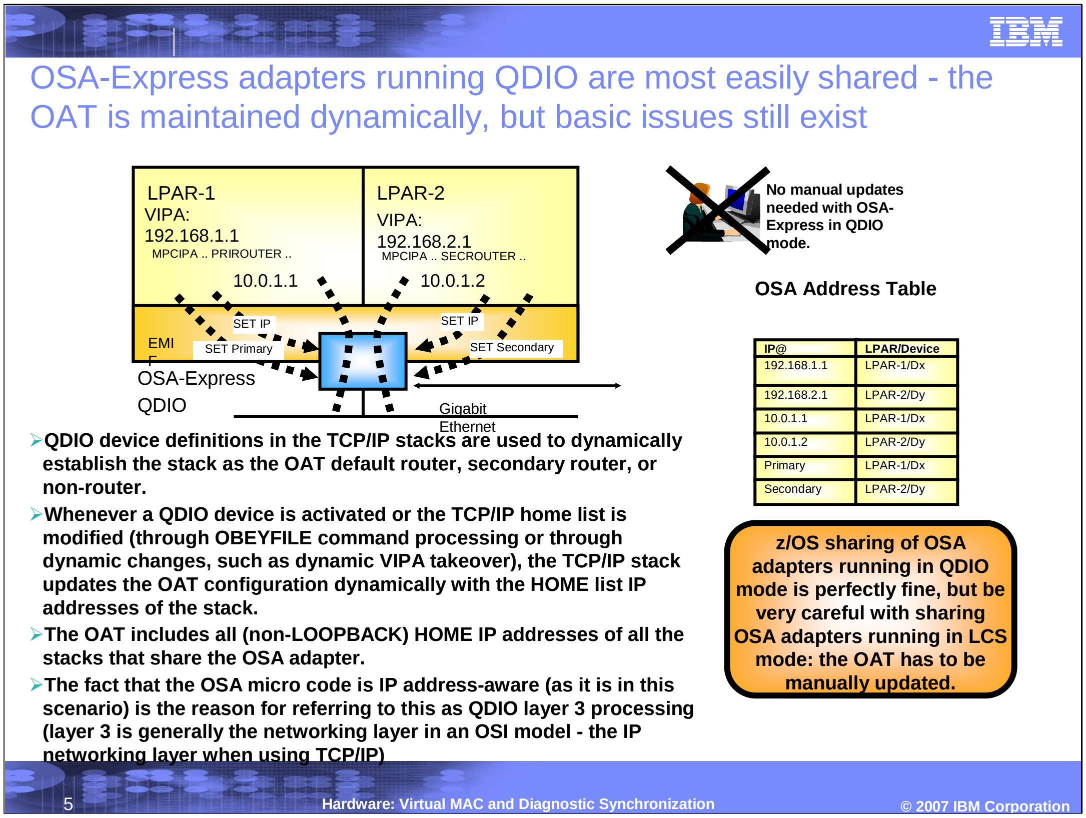
If the destination IP address in a frame that arrives at the OSA adapter doesn't belong to any of the LPARs that share the adapter, that IP packet is given to the LPAR that acts as the Default OSA router. There are separate IPv4 and IPv6 default OSA router specifications.

- **The OSA adapter is not a full-function IP router. It has some limited capabilities built in to analyze destination IP addresses to determine which LPAR an incoming IP packet belongs to.**
- **It does not maintain an IP-layer routing table, but depends on the System z TCP/IP stacks to do that and handle all IP routing issues.**
- **The OSA adapter network port has a single (Media Access Control) MAC address that is shared among all the LPARs that share the adapter.**
  - All IP packets to all LPARs are destined for one and the same MAC address, and the OSA adapter looks into the IP header to figure out which LPAR an IP packet belongs to.
- **If z/OS® acts as an IP router to IP networks behind z/OS, the destination address in the IP packets may be any IP address on those networks behind z/OS.**
  - Can become extremely cumbersome to set up if LPARs that share an OSA port are connected to different back-end networks (odd routing paths - at best!)
- **When a port is defined as an LCS port, the content of the OAT must be maintained manually through OSA/SF interaction.**
- **When a port is defined as a QDIO port, the content of the OAT is maintained 'automatically' by the sharing TCP/IP stacks.**

The primary router definitions do not apply to MPCOSA - some limitations in how the OAT can be configured for MPCOSA.

MPCOSA was really just implemented to facilitate migration from HSAS to the native TCP/IP stack.

There are no plans to allow for display of ARP cache information from the OSA-2 adapter.



IBM

OSA-Express adapters running QDIO are most easily shared - the OAT is maintained dynamically, but basic issues still exist

LPAR-1
VIPA:
192.168.1.1
MPCIPA .. PRIROUTER ..
10.0.1.1

LPAR-2
VIPA:
192.168.2.1
MPCIPA .. SECROUTER ..
10.0.1.2

SET IP
SET Primary
SET IP
SET Secondary
EMIF
OSA-Express
QDIO
Gigabit
Ethernet

No manual updates needed with OSA-Express in QDIO mode.

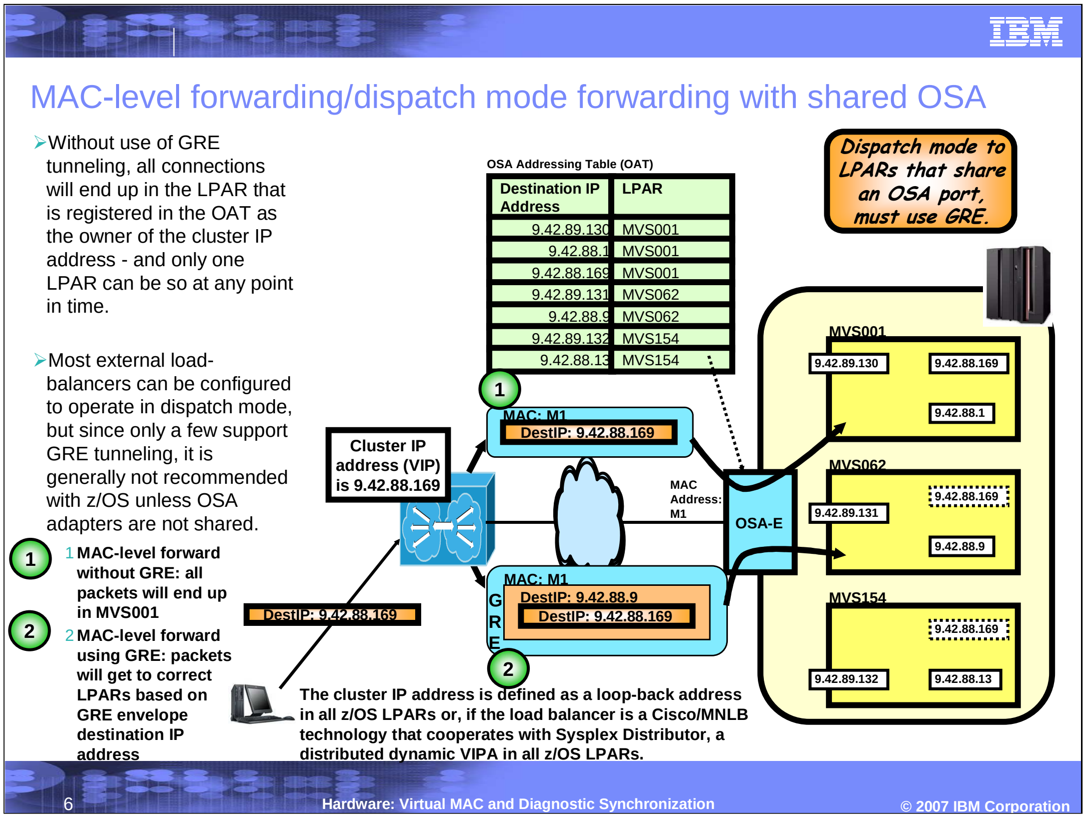
OSA Address Table

IP@	LPAR/Device
192.168.1.1	LPAR-1/Dx
192.168.2.1	LPAR-2/Dy
10.0.1.1	LPAR-1/Dx
10.0.1.2	LPAR-2/Dy
Primary	LPAR-1/Dx
Secondary	LPAR-2/Dy

- **QDIO device definitions in the TCP/IP stacks are used to dynamically establish the stack as the OAT default router, secondary router, or non-router.**
- **Whenever a QDIO device is activated or the TCP/IP home list is modified (through OBEYFILE command processing or through dynamic changes, such as dynamic VIPA takeover), the TCP/IP stack updates the OAT configuration dynamically with the HOME list IP addresses of the stack.**
- **The OAT includes all (non-LOOPBACK) HOME IP addresses of all the stacks that share the OSA adapter.**
- **The fact that the OSA micro code is IP address-aware (as it is in this scenario) is the reason for referring to this as QDIO layer 3 processing (layer 3 is generally the networking layer in an OSI model - the IP networking layer when using TCP/IP)**

z/OS sharing of OSA adapters running in QDIO mode is perfectly fine, but be very careful with sharing OSA adapters running in LCS mode: the OAT has to be manually updated.

Multiple stacks per LPAR is supported. The SET commands pass both LPAR and eveice number information to the adapter.



MAC-level forwarding/dispatch mode forwarding with shared OSA

- Without use of GRE tunneling, all connections will end up in the LPAR that is registered in the OAT as the owner of the cluster IP address - and only one LPAR can be so at any point in time.

- Most external load-balancers can be configured to operate in dispatch mode, but since only a few support GRE tunneling, it is generally not recommended with z/OS unless OSA adapters are not shared.

1 MAC-level forward without GRE: all packets will end up in MVS001

2 MAC-level forward using GRE: packets will get to correct LPARs based on GRE envelope destination IP address

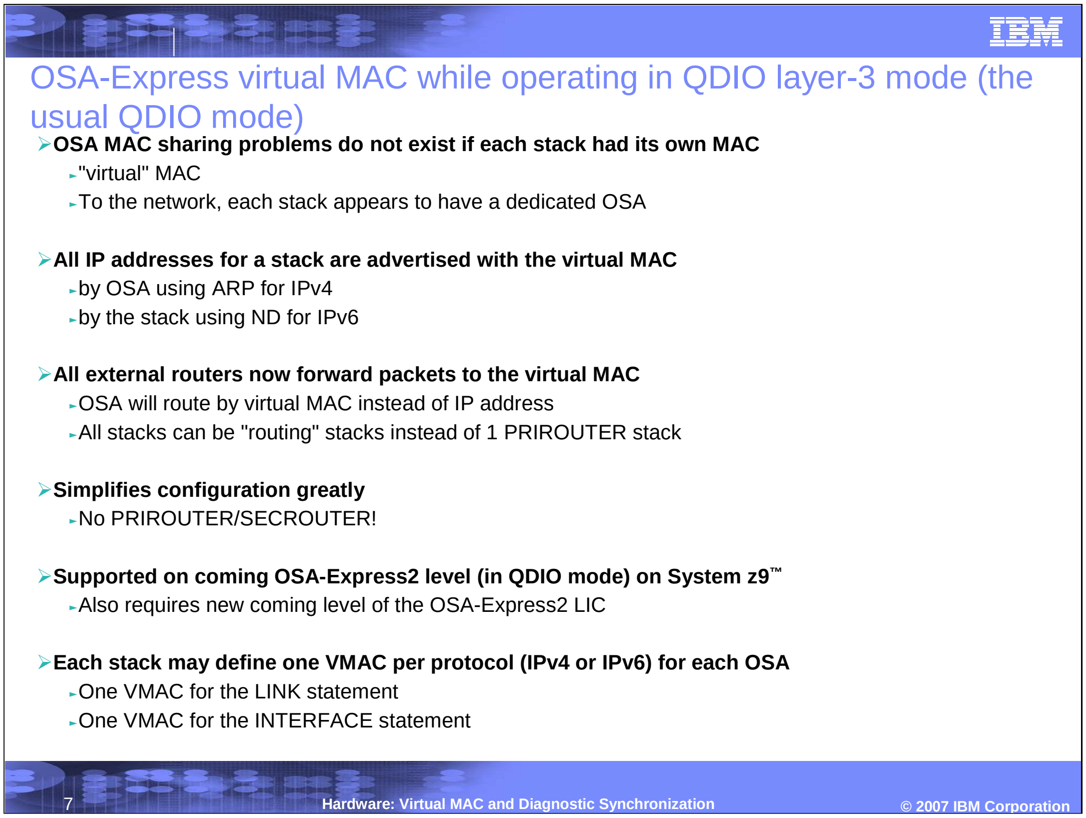
OSA Addressing Table (OAT)

Destination IP Address	LPAR
9.42.89.130	MVS001
9.42.88.1	MVS001
9.42.88.169	MVS001
9.42.89.131	MVS062
9.42.88.9	MVS062
9.42.89.132	MVS154
9.42.88.13	MVS154

Dispatch mode to LPARs that share an OSA port, must use GRE.

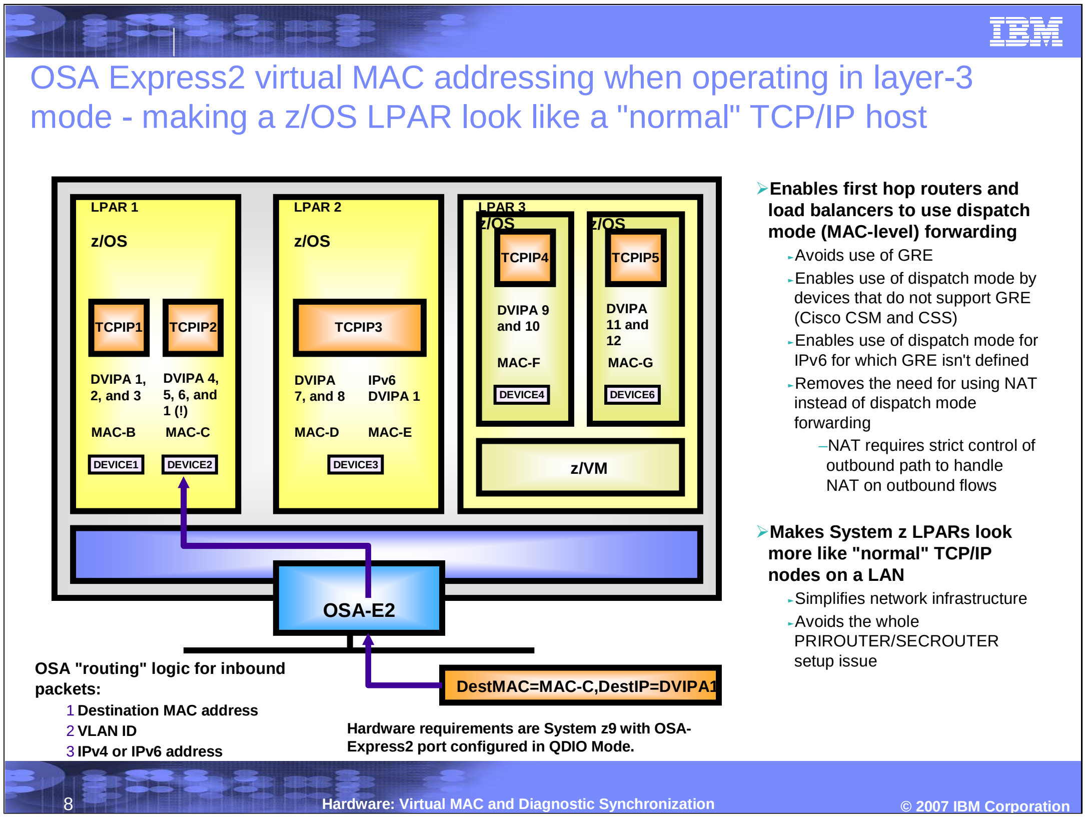
Cluster IP address (VIP) is 9.42.88.169

The cluster IP address is defined as a loop-back address in all z/OS LPARs or, if the load balancer is a Cisco/MNLB technology that cooperates with Sysplex Distributor, a distributed dynamic VIPA in all z/OS LPARs.



OSA-Express virtual MAC while operating in QDIO layer-3 mode (the usual QDIO mode)

- **OSA MAC sharing problems do not exist if each stack had its own MAC**
  - "virtual" MAC
  - To the network, each stack appears to have a dedicated OSA
- **All IP addresses for a stack are advertised with the virtual MAC**
  - by OSA using ARP for IPv4
  - by the stack using ND for IPv6
- **All external routers now forward packets to the virtual MAC**
  - OSA will route by virtual MAC instead of IP address
  - All stacks can be "routing" stacks instead of 1 PRIROUTER stack
- **Simplifies configuration greatly**
  - No PRIROUTER/SECROUTER!
- **Supported on coming OSA-Express2 level (in QDIO mode) on System z9™**
  - Also requires new coming level of the OSA-Express2 LIC
- **Each stack may define one VMAC per protocol (IPv4 or IPv6) for each OSA**
  - One VMAC for the LINK statement
  - One VMAC for the INTERFACE statement



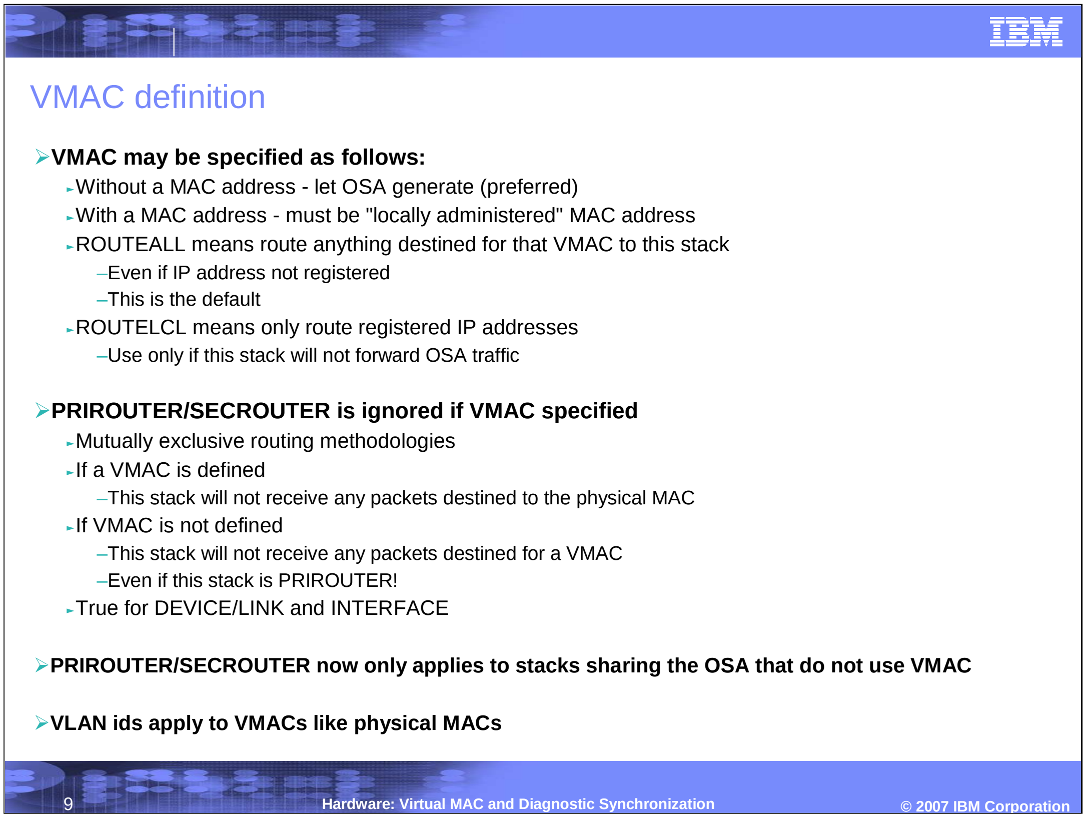
OSA Express2 virtual MAC addressing when operating in layer-3 mode - making a z/OS LPAR look like a "normal" TCP/IP host

OSA "routing" logic for inbound packets:

1. Destination MAC address
2. VLAN ID
3. IPv4 or IPv6 address

Hardware requirements are System z9 with OSA-Express2 port configured in QDIO Mode.

- **Enables first hop routers and load balancers to use dispatch mode (MAC-level) forwarding**
  - Avoids use of GRE
  - Enables use of dispatch mode by devices that do not support GRE (Cisco CSM and CSS)
  - Enables use of dispatch mode for IPv6 for which GRE isn't defined
  - Removes the need for using NAT instead of dispatch mode forwarding
    - NAT requires strict control of outbound path to handle NAT on outbound flows
- **Makes System z LPARs look more like "normal" TCP/IP nodes on a LAN**
  - Simplifies network infrastructure
  - Avoids the whole PRIROUTER/SECROUTER setup issue



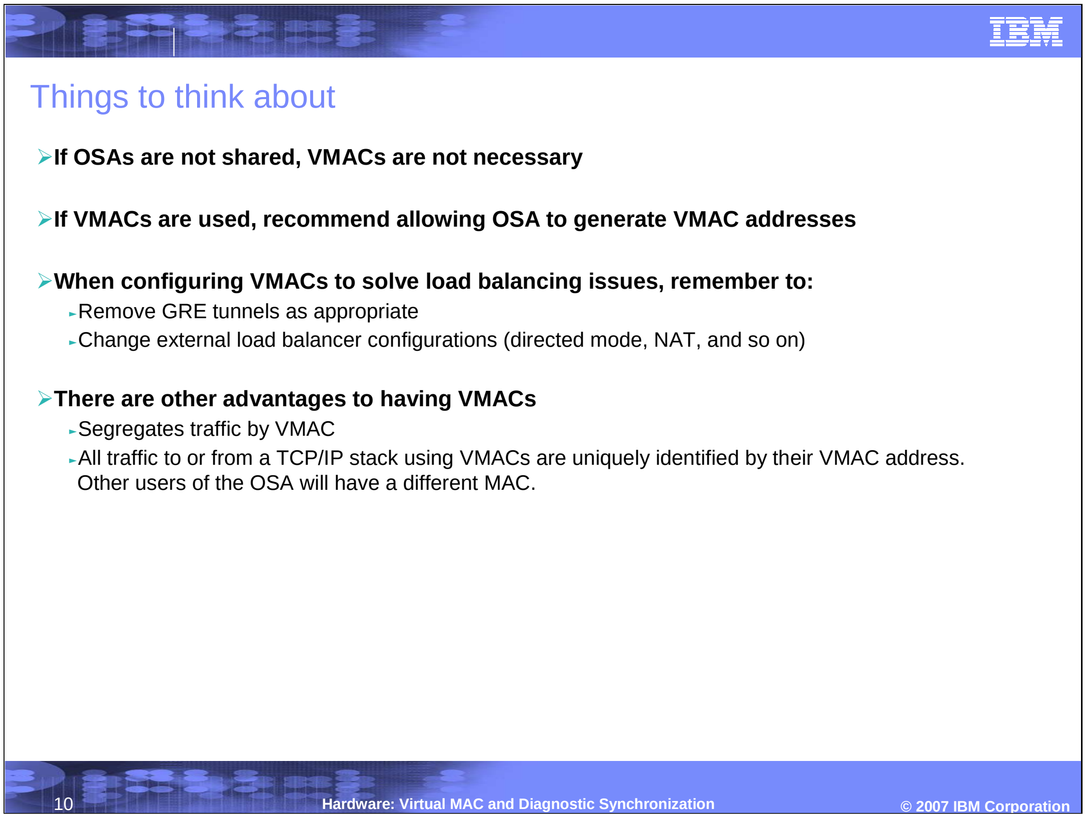
VMAC definition

- **VMAC may be specified as follows:**
  - Without a MAC address - let OSA generate (preferred)
  - With a MAC address - must be "locally administered" MAC address
  - ROUTEALL means route anything destined for that VMAC to this stack
    - Even if IP address not registered
    - This is the default
  - ROUTELCL means only route registered IP addresses
    - Use only if this stack will not forward OSA traffic

- **PRIROUTER/SECROUTER is ignored if VMAC specified**
  - Mutually exclusive routing methodologies
  - If a VMAC is defined
    - This stack will not receive any packets destined to the physical MAC
  - If VMAC is not defined
    - This stack will not receive any packets destined for a VMAC
    - Even if this stack is PRIROUTER!
  - True for DEVICE/LINK and INTERFACE

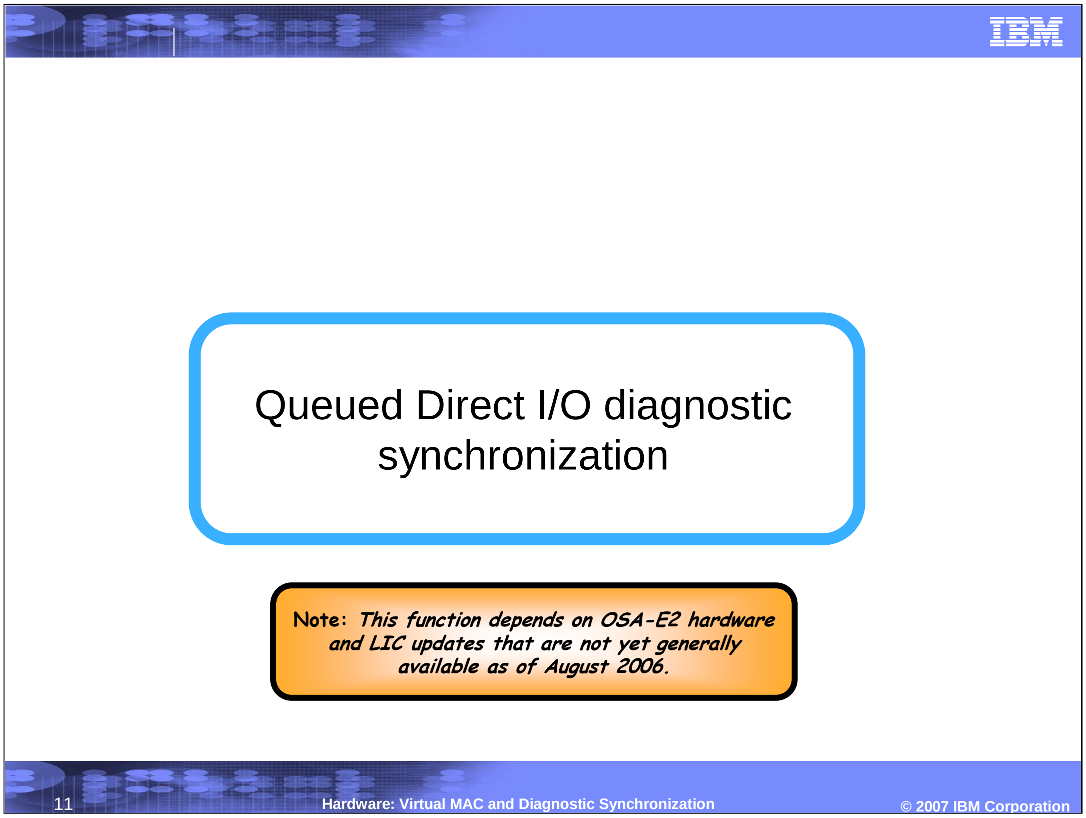
- **PRIROUTER/SECROUTER now only applies to stacks sharing the OSA that do not use VMAC**

- **VLAN ids apply to VMACs like physical MACs**



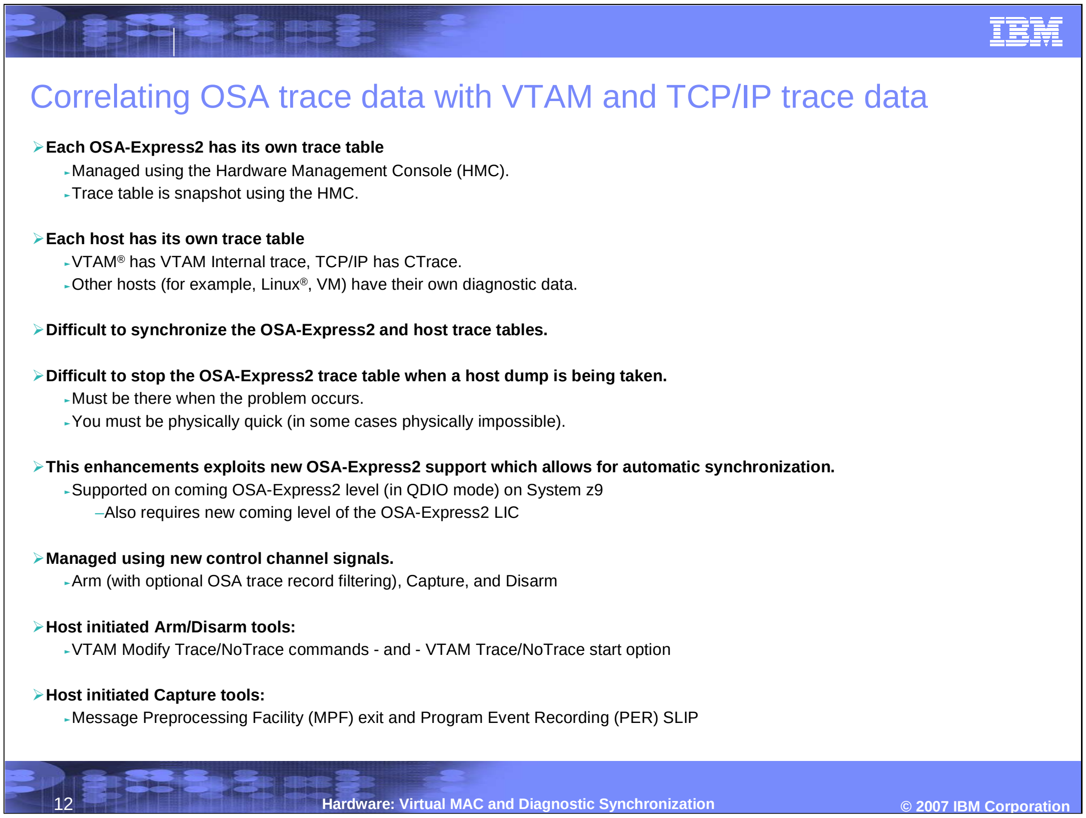
Things to think about

- **If OSAs are not shared, VMACs are not necessary**
- **If VMACs are used, recommend allowing OSA to generate VMAC addresses**
- **When configuring VMACs to solve load balancing issues, remember to:**
  - Remove GRE tunnels as appropriate
  - Change external load balancer configurations (directed mode, NAT, and so on)
- **There are other advantages to having VMACs**
  - Segregates traffic by VMAC
  - All traffic to or from a TCP/IP stack using VMACs are uniquely identified by their VMAC address. Other users of the OSA will have a different MAC.



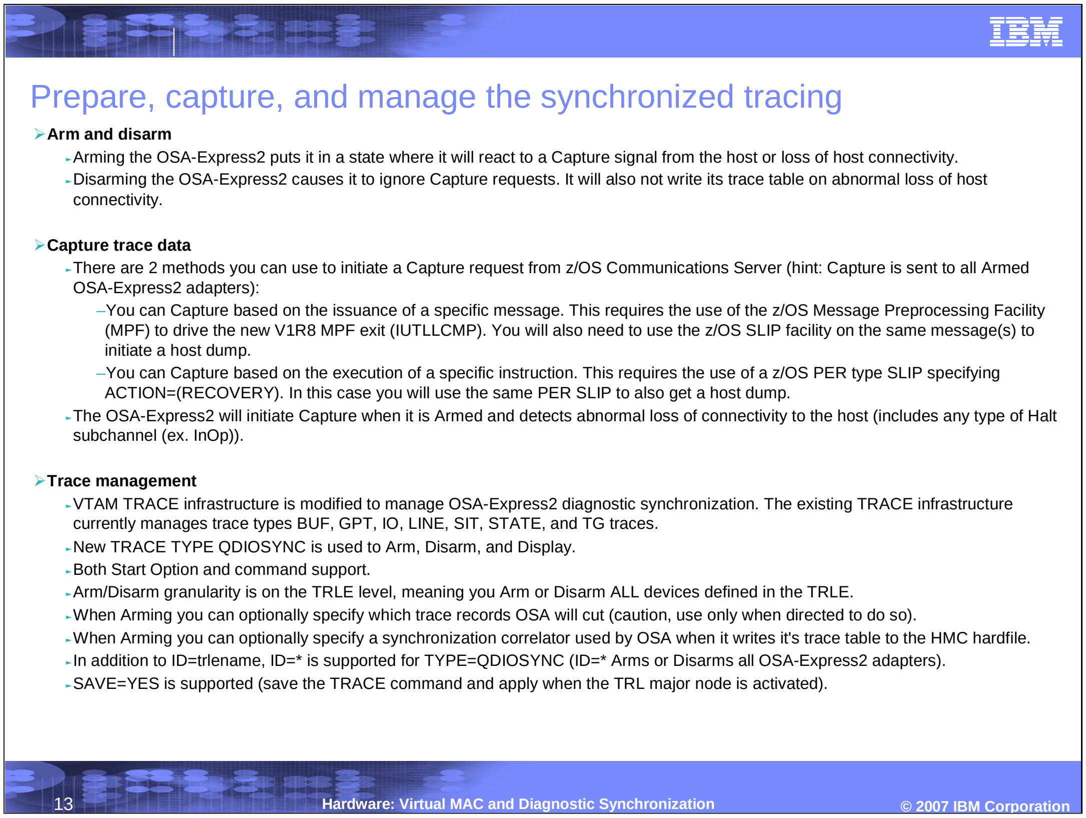
Queued Direct I/O diagnostic synchronization

Note: *This function depends on OSA-E2 hardware and LIC updates that are not yet generally available as of August 2006.*



Correlating OSA trace data with VTAM and TCP/IP trace data

- **Each OSA-Express2 has its own trace table**
  - Managed using the Hardware Management Console (HMC).
  - Trace table is snapshot using the HMC.
- **Each host has its own trace table**
  - VTAM® has VTAM Internal trace, TCP/IP has CTrace.
  - Other hosts (for example, Linux®, VM) have their own diagnostic data.
- **Difficult to synchronize the OSA-Express2 and host trace tables.**
- **Difficult to stop the OSA-Express2 trace table when a host dump is being taken.**
  - Must be there when the problem occurs.
  - You must be physically quick (in some cases physically impossible).
- **This enhancements exploits new OSA-Express2 support which allows for automatic synchronization.**
  - Supported on coming OSA-Express2 level (in QDIO mode) on System z9
    - Also requires new coming level of the OSA-Express2 LIC
- **Managed using new control channel signals.**
  - Arm (with optional OSA trace record filtering), Capture, and Disarm
- **Host initiated Arm/Disarm tools:**
  - VTAM Modify Trace/NoTrace commands - and - VTAM Trace/NoTrace start option
- **Host initiated Capture tools:**
  - Message Preprocessing Facility (MPF) exit and Program Event Recording (PER) SLIP

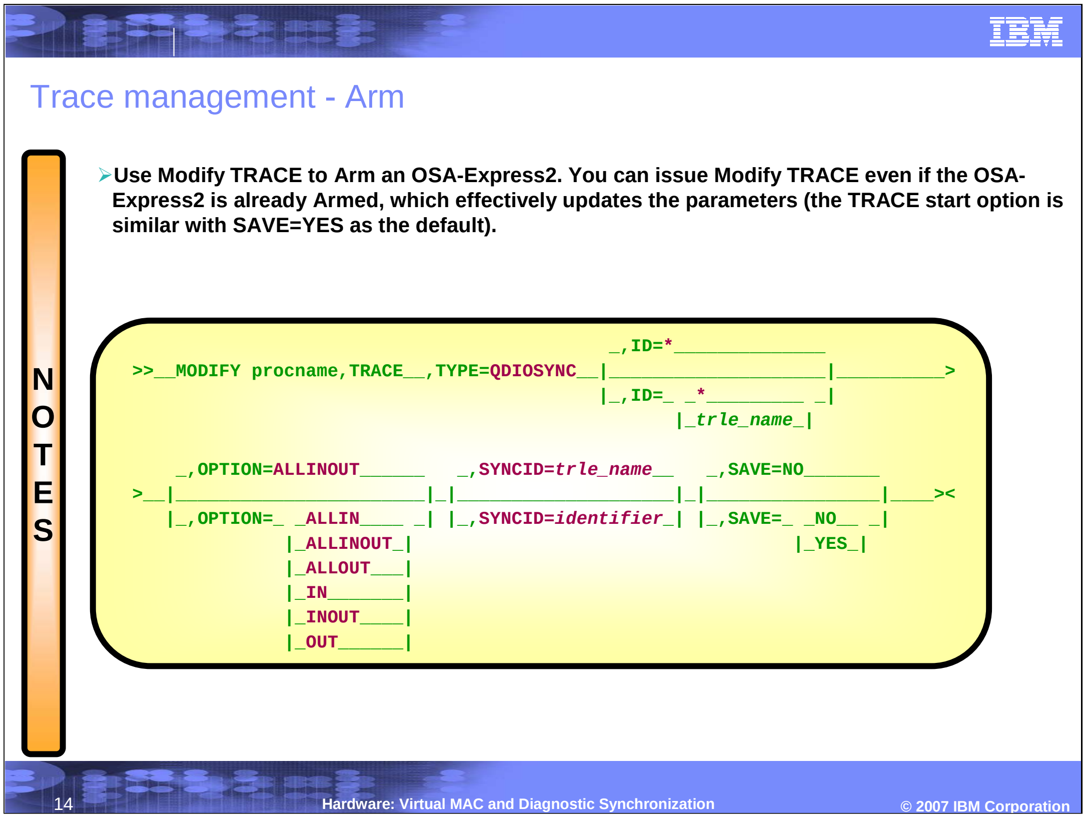


Prepare, capture, and manage the synchronized tracing

- **Arm and disarm**
  - Arming the OSA-Express2 puts it in a state where it will react to a Capture signal from the host or loss of host connectivity.
  - Disarming the OSA-Express2 causes it to ignore Capture requests. It will also not write its trace table on abnormal loss of host connectivity.

- **Capture trace data**
  - There are 2 methods you can use to initiate a Capture request from z/OS Communications Server (hint: Capture is sent to all Armed OSA-Express2 adapters):
    - You can Capture based on the issuance of a specific message. This requires the use of the z/OS Message Preprocessing Facility (MPF) to drive the new V1R8 MPF exit (IUTLLCMP). You will also need to use the z/OS SLIP facility on the same message(s) to initiate a host dump.
    - You can Capture based on the execution of a specific instruction. This requires the use of a z/OS PER type SLIP specifying ACTION=(RECOVERY). In this case you will use the same PER SLIP to also get a host dump.
  - The OSA-Express2 will initiate Capture when it is Armed and detects abnormal loss of connectivity to the host (includes any type of Halt subchannel (ex. InOp)).

- **Trace management**
  - VTAM TRACE infrastructure is modified to manage OSA-Express2 diagnostic synchronization. The existing TRACE infrastructure currently manages trace types BUF, GPT, IO, LINE, SIT, STATE, and TG traces.
  - New TRACE TYPE QDIOSYNC is used to Arm, Disarm, and Display.
  - Both Start Option and command support.
  - Arm/Disarm granularity is on the TRLE level, meaning you Arm or Disarm ALL devices defined in the TRLE.
  - When Arming you can optionally specify which trace records OSA will cut (caution, use only when directed to do so).
  - When Arming you can optionally specify a synchronization correlator used by OSA when it writes it's trace table to the HMC hardfile.
  - In addition to ID=trlename, ID=* is supported for TYPE=QDIOSYNC (ID=* Arms or Disarms all OSA-Express2 adapters).
  - SAVE=YES is supported (save the TRACE command and apply when the TRL major node is activated).



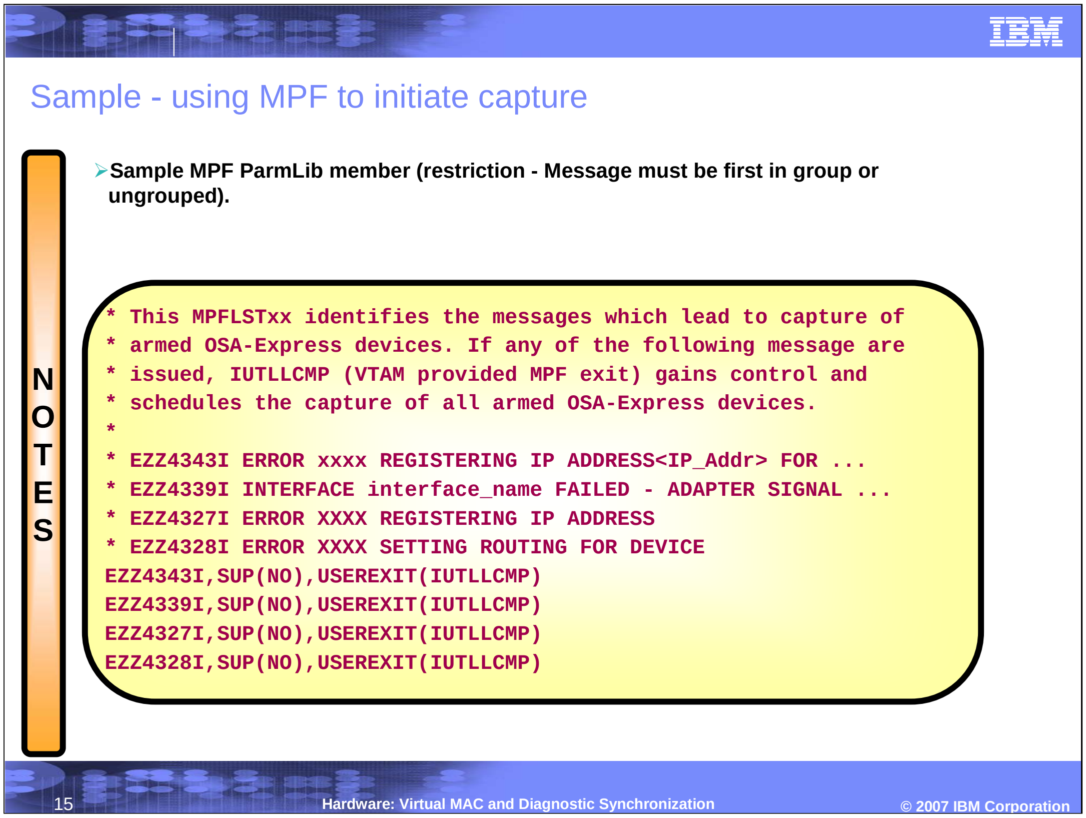
Trace management - Arm

N O T E S

- Use Modify TRACE to Arm an OSA-Express2. You can issue Modify TRACE even if the OSA-Express2 is already Armed, which effectively updates the parameters (the TRACE start option is similar with SAVE=YES as the default).

```
                                            _,ID=*______________
>>__MODIFY procname,TRACE__,TYPE=QDIOSYNC__|____________________|__________>
                                           |_,ID=_ _*_________ _|
                                                  |_trle_name_|

    _,OPTION=ALLINOUT______   _,SYNCID=trle_name__   _,SAVE=NO_______
>__|_______________________|_|____________________|_|________________|____><
   |_,OPTION=_ _ALLIN____ _| |_,SYNCID=identifier_| |_,SAVE=_ _NO__ _|
              |_ALLINOUT_|                                   |_YES_|
              |_ALLOUT___|
              |_IN_______|
              |_INOUT____|
              |_OUT______|
```



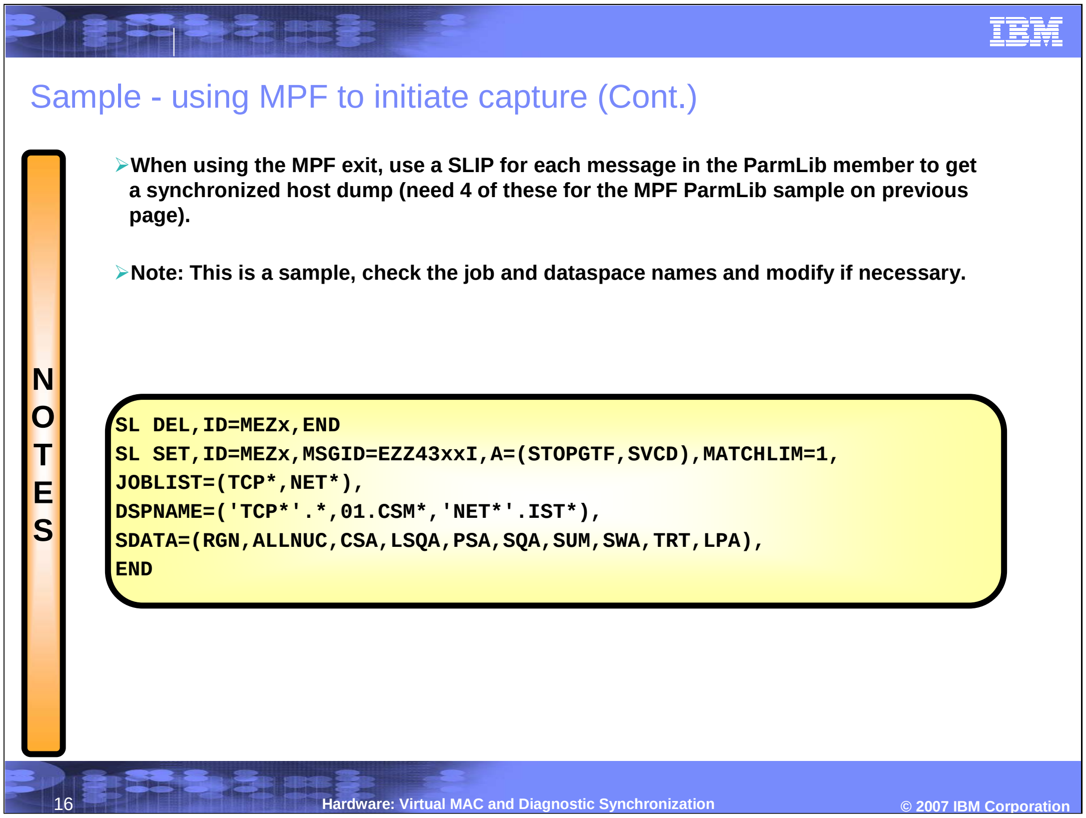
IBM

Sample - using MPF to initiate capture

NOTES

- Sample MPF ParmLib member (restriction - Message must be first in group or ungrouped).

```
* This MPFLSTxx identifies the messages which lead to capture of
* armed OSA-Express devices. If any of the following message are
* issued, IUTLLCMP (VTAM provided MPF exit) gains control and
* schedules the capture of all armed OSA-Express devices.
*
* EZZ4343I ERROR xxxx REGISTERING IP ADDRESS<IP_Addr> FOR ...
* EZZ4339I INTERFACE interface_name FAILED - ADAPTER SIGNAL ...
* EZZ4327I ERROR XXXX REGISTERING IP ADDRESS
* EZZ4328I ERROR XXXX SETTING ROUTING FOR DEVICE
EZZ4343I,SUP(NO),USEREXIT(IUTLLCMP)
EZZ4339I,SUP(NO),USEREXIT(IUTLLCMP)
EZZ4327I,SUP(NO),USEREXIT(IUTLLCMP)
EZZ4328I,SUP(NO),USEREXIT(IUTLLCMP)
```

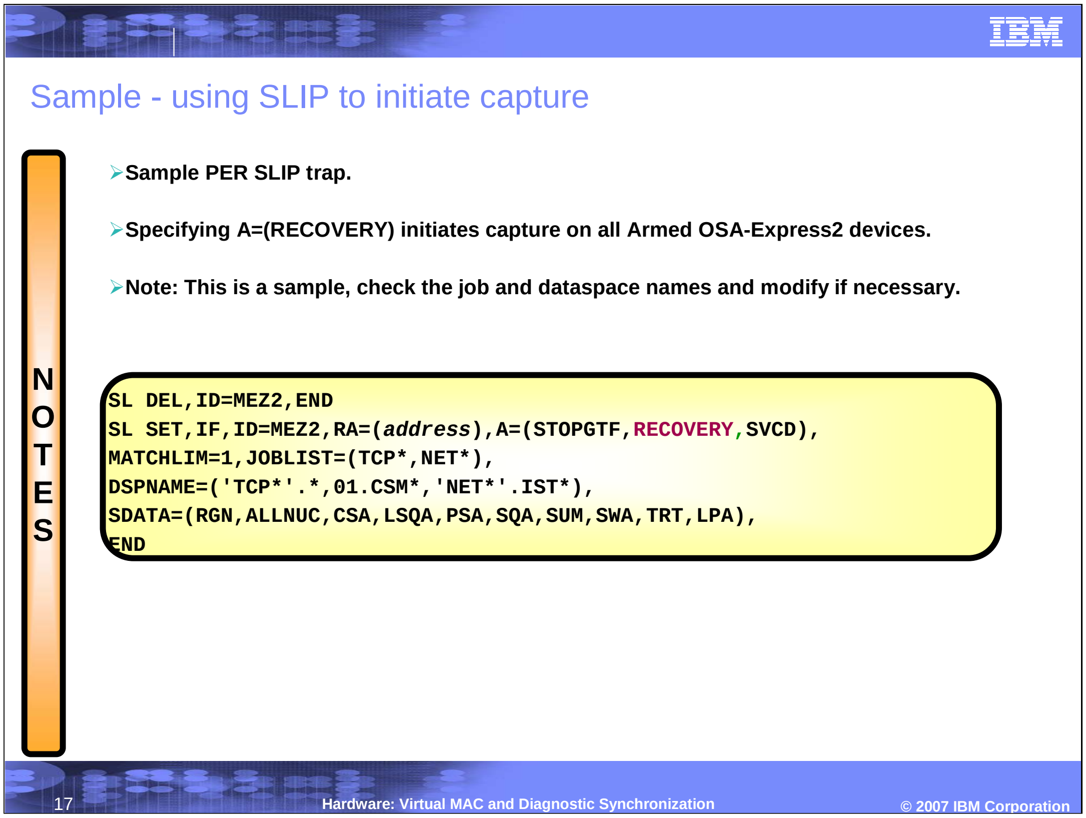


Sample - using MPF to initiate capture (Cont.)

NOTES

- When using the MPF exit, use a SLIP for each message in the ParmLib member to get a synchronized host dump (need 4 of these for the MPF ParmLib sample on previous page).
- Note: This is a sample, check the job and dataspace names and modify if necessary.

```
SL DEL,ID=MEZx,END
SL SET,ID=MEZx,MSGID=EZZ43xxI,A=(STOPGTF,SVCD),MATCHLIM=1,
JOBLIST=(TCP*,NET*),
DSPNAME=('TCP*'.*,01.CSM*,'NET*'.IST*),
SDATA=(RGN,ALLNUC,CSA,LSQA,PSA,SQA,SUM,SWA,TRT,LPA),
END
```

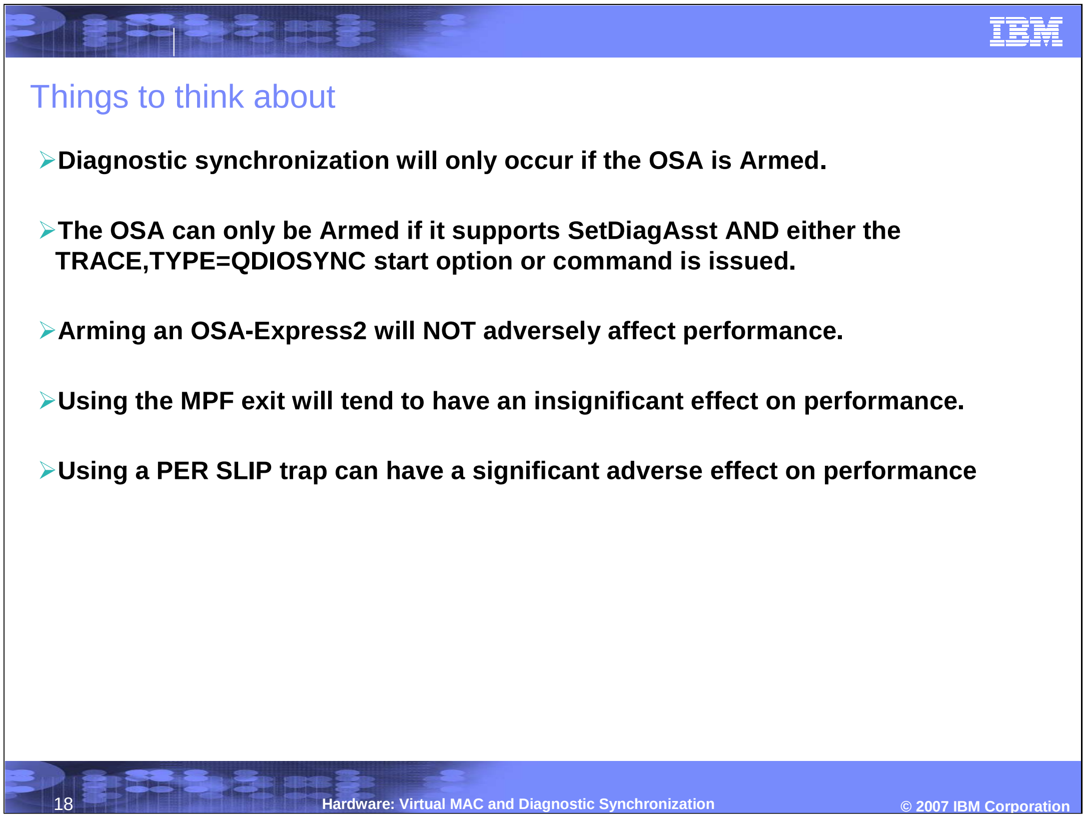


Sample - using SLIP to initiate capture

NOTES

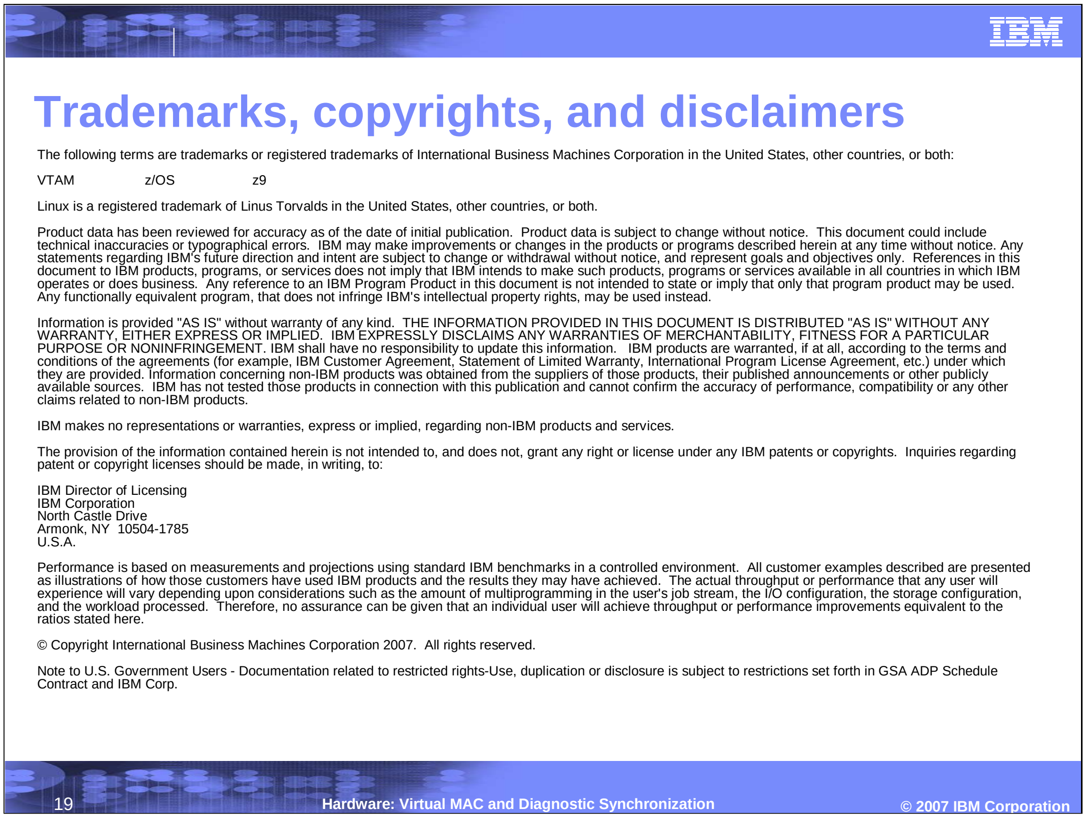
- ➢Sample PER SLIP trap.
- ➢Specifying A=(RECOVERY) initiates capture on all Armed OSA-Express2 devices.
- ➢Note: This is a sample, check the job and dataspace names and modify if necessary.

```
SL DEL,ID=MEZ2,END
SL SET,IF,ID=MEZ2,RA=(address),A=(STOPGTF,RECOVERY,SVCD),
MATCHLIM=1,JOBLIST=(TCP*,NET*),
DSPNAME=('TCP*'.*,01.CSM*,'NET*'.IST*),
SDATA=(RGN,ALLNUC,CSA,LSQA,PSA,SQA,SUM,SWA,TRT,LPA),
END
```



Things to think about

- Diagnostic synchronization will only occur if the OSA is Armed.
- The OSA can only be Armed if it supports SetDiagAsst AND either the TRACE,TYPE=QDIOSYNC start option or command is issued.
- Arming an OSA-Express2 will NOT adversely affect performance.
- Using the MPF exit will tend to have an insignificant effect on performance.
- Using a PER SLIP trap can have a significant adverse effect on performance



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:

VTAM z/OS z9

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

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.

Note to U.S. Government Users - Documentation related to restricted rights-Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract and IBM Corp.