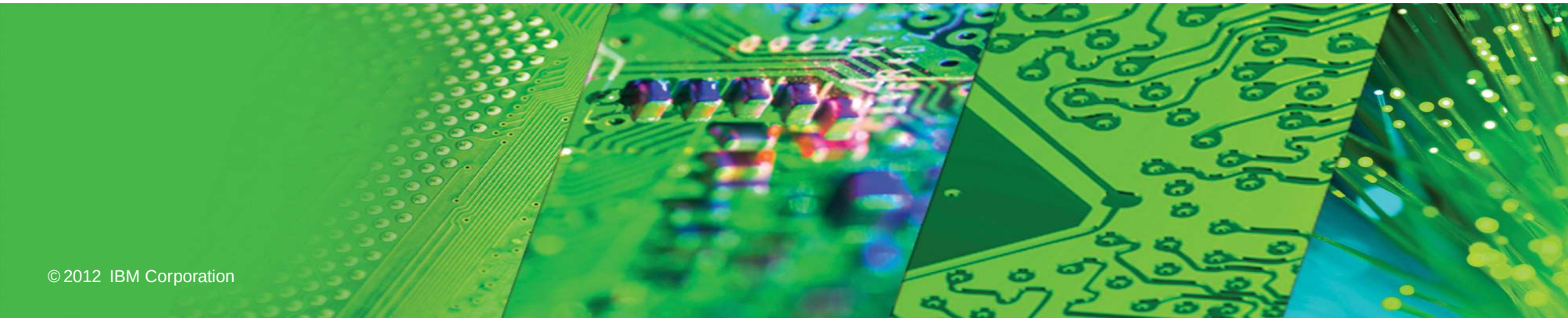
2012
IBM System z Technical University

Enabling the infrastructure for smarter computing

Networking with Linux on System z

zLG21

Dr. Stefan Reimbold



© 2012 IBM Corporation

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at Copyright and trademark information at www.ibm.com/legal/copytrade.shtml.

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.

Agenda

- Overview
- Basic Network Setup
- Advanced Network Setup
- Networking Tools
- Problem Determination / Debugging

© 2012 IBM Corporation

Agenda

- Overview
- Basic Network Setup

© 2012 IBM Corporation

Networking Options

- OSA
- HiperSockets
- Virtual NIC
  - Guest LAN
  - VSWITCH
- LCS
- CTC
- NETIUCV

© 2012 IBM Corporation

Networking Drivers

- QETH
- LCS
- CTC (functionally stable)
- NETIUCV (functionally stable)

© 2012 IBM Corporation

Networking Drivers

- CTC - Channel-To-Channel Connection
- IUCV - Inter User Communication Vehicle
- device driver deprecated (kernel 2.6)
- still available for backwards comptibility
- migration path
  - Virtual CTC and IUCV $\Rightarrow$ Guest LAN
  - CTC in LPAR $\Rightarrow$ HiperSockets
  - CTC $\Rightarrow$ OSA-Express

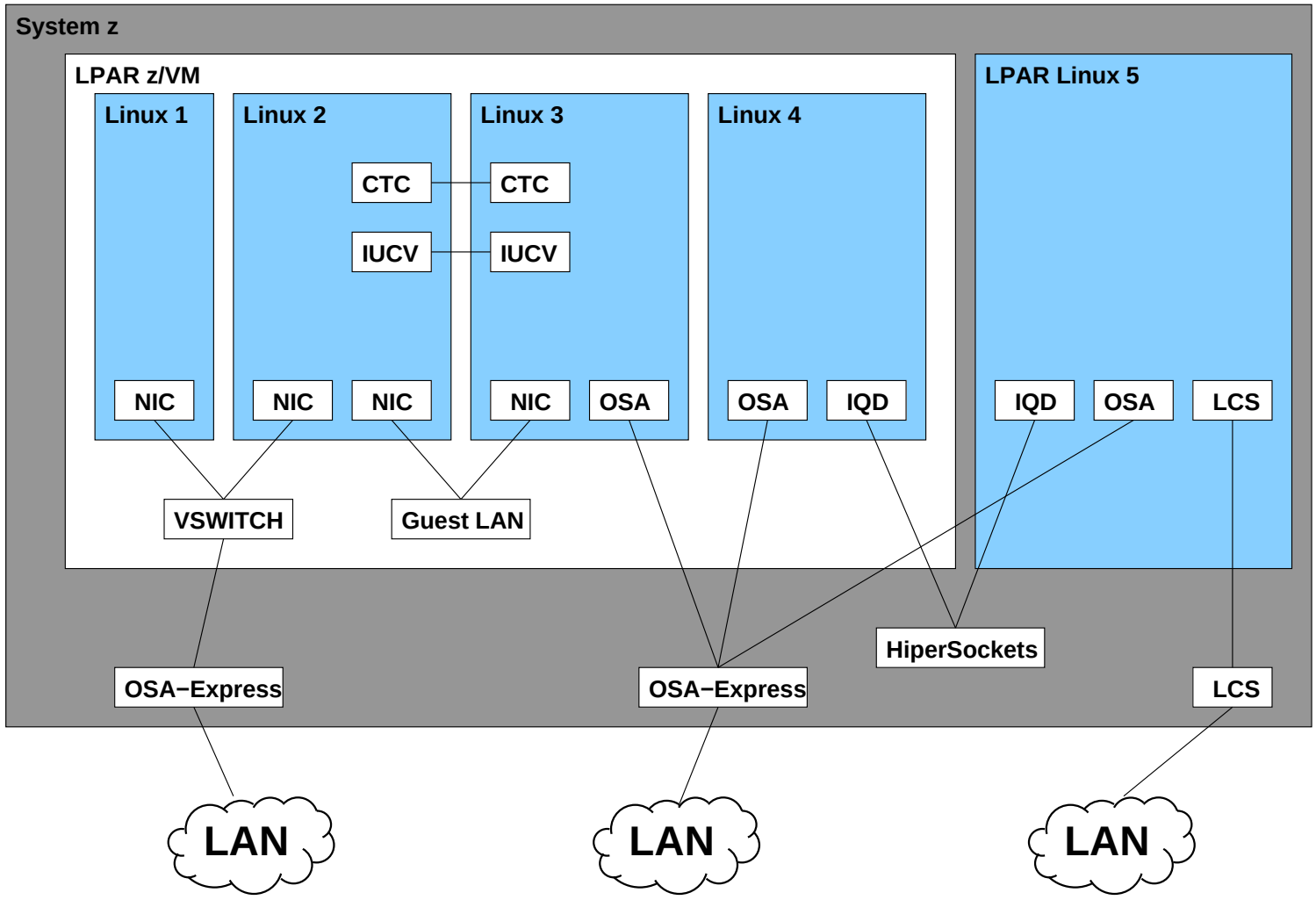
© 2012 IBM Corporation

QETH Device Driver

- supports
  - OSA-Express
  - HiperSockets
  - Guest LAN
  - VSWITCH
- primary network driver for Linux on System z

© 2012 IBM Corporation

System z Network



© 2012 IBM Corporation

OSA

- OSA Express (all adapters)
- 10 GbE, GbE, 1000BASE-T
- 640 TCP/IP stacks per CHPID
- jumbo frame support
- concurrent LIC update

© 2012 IBM Corporation

OSA

- OSA Express2
  - 2 CHPIDs with 1 port per CHPID (10 GbE 1 CHPID)
- OSA Express3
  - multimode or single mode fiber or copper cable
  - 2 CHPIDs with 2 port per CHPID (10 GbE 1 port)
  - DMA
  - hardware data router
- OSA-Express4S
  - multimode or single mode fiber or copper cable
  - 1 CHPID shared by 2 ports
  - DMA
  - hardware data router
  - IEDN support
  - checksum offload for IPv4 and IPv6

© 2012 IBM Corporation

HiperSockets

- up to 32 HiperSockets Networks
- shared by 30 LPARs
- communication through memory
- IPv4 and IPv6 support
- Layer 2 and Layer 3 support
- VLAN
- Sniffer support

© 2012 IBM Corporation

Frame Size

- 4 different frame sizes

frame size	MTU size
16k	8k
24k	16k
40k	32k
64k	56k

- frame size for HiperSockets network specified

© 2012 IBM Corporation

HiperSockets Sniffer

- enabled from Service Element or HMC
- 1 device receives all packets
- device filtering can be enabled

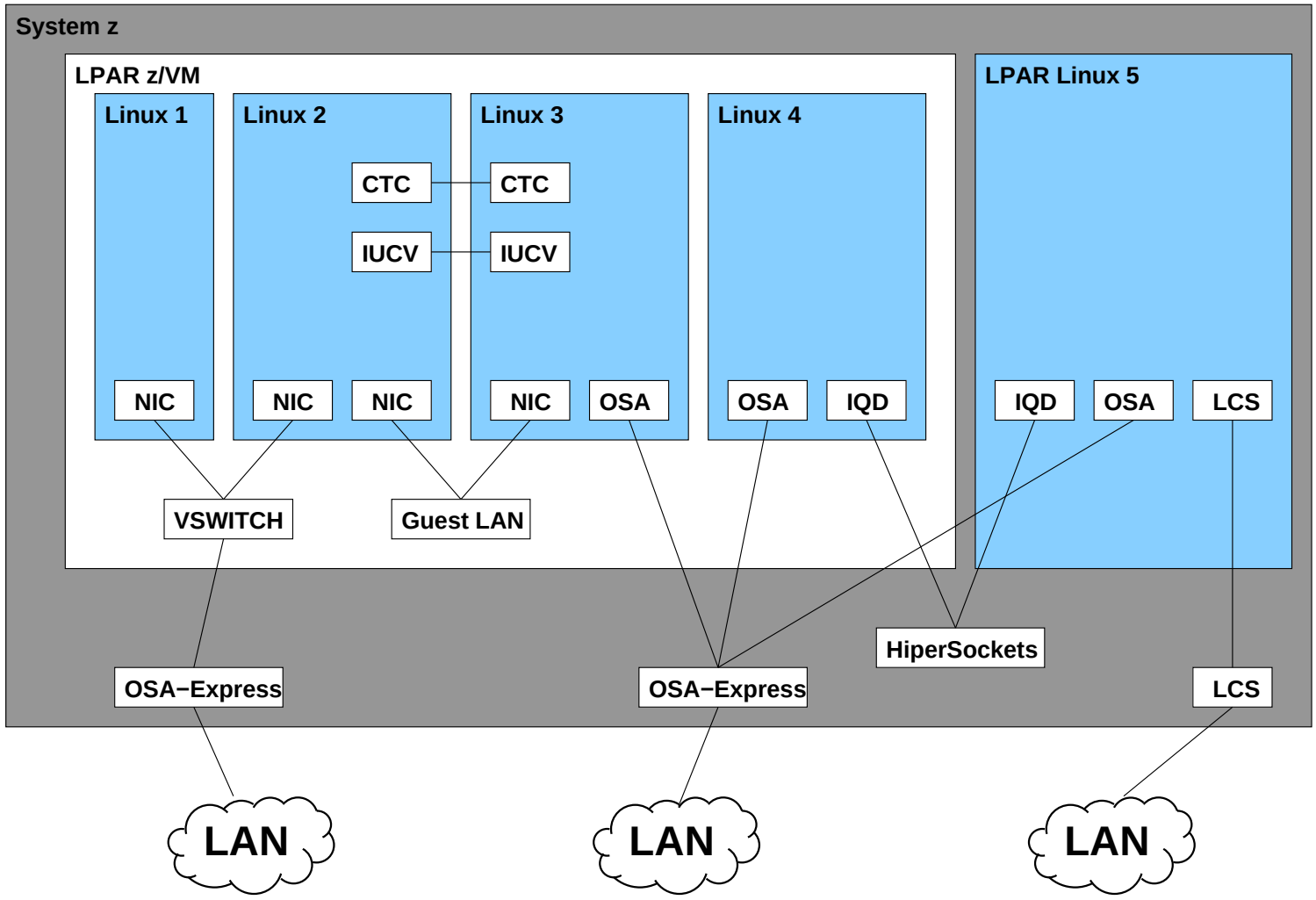
© 2012 IBM Corporation

Guest LAN

- simulated LAN segment
- used with virtual Network Interface Card (NIC)
- Types
  - QDIO Layer3
  - HiperSockets Layer3
  - Ethernet Layer2
- no physical connection
- unrestricted / restricted
- supports IPv4 and IPv6

© 2012 IBM Corporation

Guest LAN



© 2012 IBM Corporation

Guest LAN

- create Guest LAN

```
# modprobe vmcp
# vmcp define lan mylan ownerid r3515038 type qdio
LAN R3515038 MYLAN is created
```

- create virtual Network Interface

```
# vmcp define nic 0700 qdio
NIC 0700 is created; devices 0700-0702 defined
```

- connect virtual Network Interface to Guest LAN

```
# vmcp couple 0700 to r3515038 mylan
RPIMGR031E RESOURCE R3515038.MYLAN SPECIFIED BY COUPLE COMMAND NOT FOUND
NIC 0700 is connected to LAN R3515038 MYLAN
```

© 2012 IBM Corporation

Guest LAN

```
# vmcp query virtual osa
OSA   0700 ON NIC  0700  UNIT 000 SUBCHANNEL = 0018
      0700 DEVTYPE OSA          VIRTUAL CHPID 03 OSD
      0700 MAC 02-35-15-00-00-77 CURRENT
      0700 QDIO-ELIGIBLE        QIOASSIST-ELIGIBLE
OSA   0701 ON NIC  0700  UNIT 001 SUBCHANNEL = 0019
      0701 DEVTYPE OSA          VIRTUAL CHPID 03 OSD
      0701 QDIO-ELIGIBLE        QIOASSIST-ELIGIBLE
OSA   0702 ON NIC  0700  UNIT 002 SUBCHANNEL = 001A
      0702 DEVTYPE OSA          VIRTUAL CHPID 03 OSD
      0702 QDIO-ELIGIBLE        QIOASSIST-ELIGIBLE
```

© 2012 IBM Corporation

Guest LAN

- remove Guest LAN

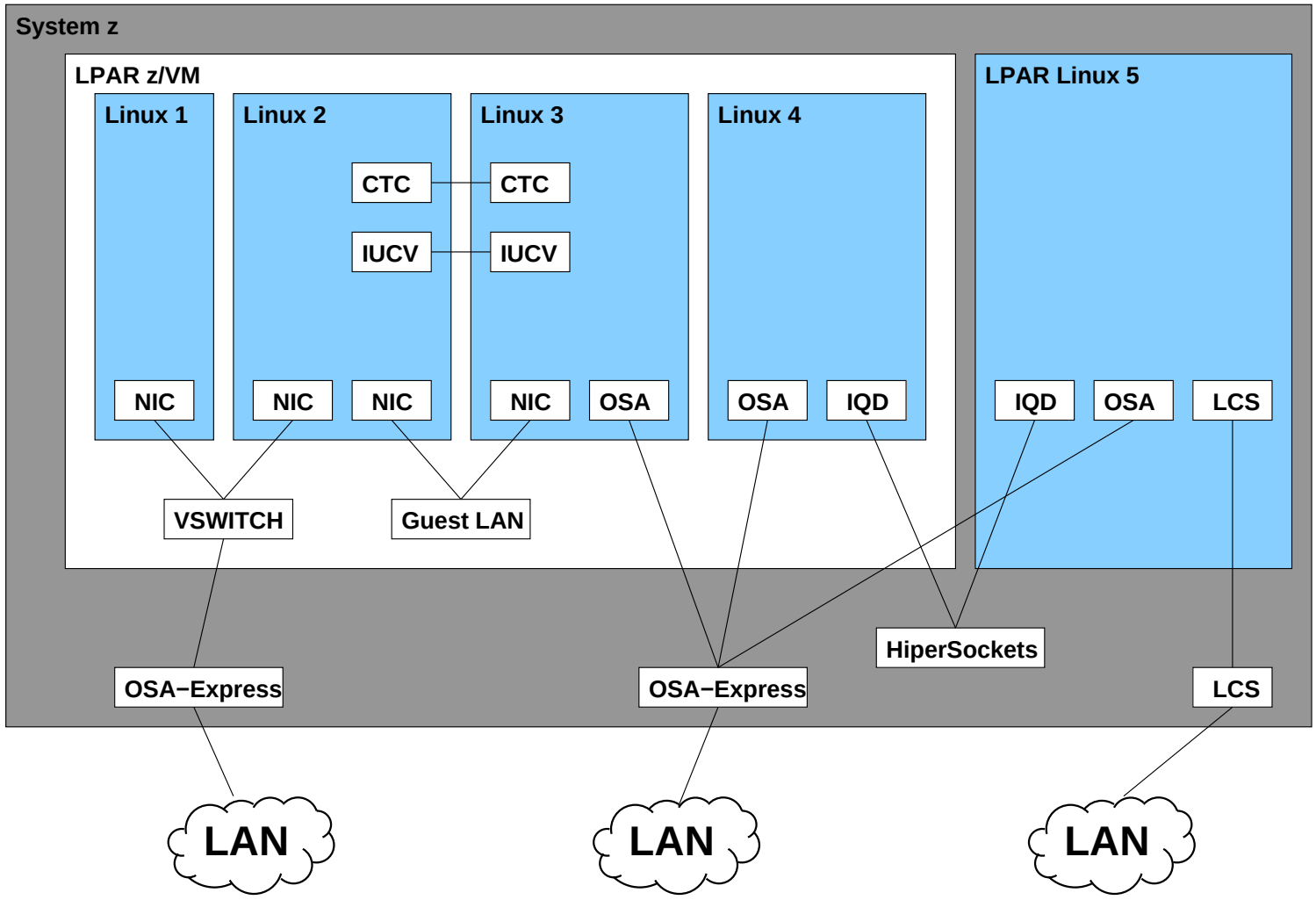
```
# vmcp uncouple 0700
NIC 0700 is disconnected from LAN R3515038 MYLAN

# vmcp detach nic 0700
NIC 0700 is destroyed; devices 0700-0702 detached

# vmcp detach lan mylan owner r3515038
LAN R3515038 MYLAN is destroyed
```

© 2012 IBM Corporation

Virtual Switch - VSWITCH



© 2012 IBM Corporation

Virtual Switch

- connects to Guest LAN
- up to 8 associated OSA connections
- connects to external network using OSA-EXPRESS
- redundant OSA-Express configuration
- single subnet for physical and virtual network
- no router necessary
- failover and Link Aggregation
- port isolation

© 2012 IBM Corporation

LAN Channel Station - LCS

- OSA Express in non-QDIO mode
  - HighSpeed TokenRing
  - ATM running Ethernet LAN Emulation
- May be preferred instead of QETH for security reasons
  - administrator defines OSA Address Table ⇒ restricted access
  - for QETH each Linux registers its own IP address
- performance not as good as QETH

© 2012 IBM Corporation

Basic Network Setup

- QETH manual setup
  - general
  - SLES11
  - SLES10
  - RHEL6
  - SLES5
- QETH YaST2 setup in SLES11

© 2012 IBM Corporation

QDIO Device Manual

```
# modprobe vmcp
# vmcp 'attach 8027-8029 *h'

# vmcp q osa
OSA  8027 ATTACHED TO R3515038 8027 DEVTYPE HIPER       CHPID FB IQD
OSA  8028 ATTACHED TO R3515038 8028 DEVTYPE HIPER       CHPID FB IQD
OSA  8029 ATTACHED TO R3515038 8029 DEVTYPE HIPER       CHPID FB IQD
```

© 2012 IBM Corporation

QDIO Device Manual

```
# modprobe qeth
# echo 0.0.8027,0.0.8028,0.0.8029 > /sys/bus/ccwgroup/drivers/qeth/group
# lsqeth
Device name                   :
--------------------------------------------
        card_type             : HiperSockets
        cdev0                 : 0.0.8027
        cdev1                 : 0.0.8028
        cdev2                 : 0.0.8029
        chpid                 : FB
        online                : 0
        portname              : no portname required
        portno                : 0
        route4                : n/a
        route6                : n/a
        state                 : DOWN
        priority_queueing     : always queue 2
        buffer_count          : 128
        layer2                : -1
        isolation             : none
```

© 2012 IBM Corporation

QDIO Device Manual

```
# echo 1 > /sys/devices/qeth/0.0.8027/online
# ifconfig hsi0
hsi0      Link encap:Ethernet  HWaddr 06:00:FB:0F:00:29
          BROADCAST NOARP MULTICAST  MTU:8192  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)
```

© 2012 IBM Corporation

QDIO Device Manual

```
# ifconfig hsi0 10.40.33.38
# ifconfig hsi0
hsi0      Link encap:Ethernet  HWaddr 06:00:FB:0F:00:29
          inet addr:10.40.33.38  Bcast:10.255.255.255  Mask:255.0.0.0
          inet6 addr: fe80::400:fbff:fe0f:29/64 Scope:Link
          UP BROADCAST RUNNING NOARP MULTICAST  MTU:8192  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:1 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 b)  TX bytes:70 (70.0 b)

# ping 10.40.33.39
PING 10.40.33.39 (10.40.33.39) 56(84) bytes of data.
64 bytes from 10.40.33.39: icmp_seq=1 ttl=64 time=0.124 ms
64 bytes from 10.40.33.39: icmp_seq=2 ttl=64 time=0.074 ms
^C
--- 10.40.33.39 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 999ms
rtt min/avg/max/mdev = 0.074/0.099/0.124/0.025 ms
```

© 2012 IBM Corporation

QDIO Device Manual

```
# ifconfig hsi0 10.40.33.38 netmask 255.255.255.0
# ifconfig hsi0
hsi0      Link encap:Ethernet  HWaddr 06:00:FB:0F:00:29
          inet addr:10.40.33.38  Bcast:10.40.33.255  Mask:255.255.255.0
          inet6 addr: fe80::400:fbff:fe0f:29/64 Scope:Link
          UP BROADCAST RUNNING NOARP MULTICAST  MTU:8192  Metric:1
          RX packets:4 errors:0 dropped:0 overruns:0 frame:0
          TX packets:7 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:336 (336.0 b)  TX bytes:602 (602.0 b)
```

© 2012 IBM Corporation

QDIO Device Manual

- network configuration lost after reboot
- devices still attached to z/VM guest

```
# vmcp q osa
OSA   8027 ATTACHED TO R3515038 8027 DEVTYPE HIPER       CHPID FB IQD
OSA   8028 ATTACHED TO R3515038 8028 DEVTYPE HIPER       CHPID FB IQD
OSA   8029 ATTACHED TO R3515038 8029 DEVTYPE HIPER       CHPID FB IQD
```

© 2012 IBM Corporation

QDIO Device Manual - SLES11

- making it persistent

```
qeth_configure 0.0.8027 0.0.8028 0.0.8029 1
```

- adds the udev rules in
  - /etc/udev/rules.d/51-hsi-0.0.8027.rules
  - /etc/udev/rules.d/70-persistent-net.rules
- create /etc/sysconfig/network/ifcfg-hsi0
  - can copy ifcfg-hsi0 and modify

```
BOOTPROTO='static'
IPADDR='10.40.33.38'
BROADCAST='10.40.33.255'
STARTMODE='auto'
NETMASK='255.255.255.0'
```

© 2012 IBM Corporation

QDIO Device Manual - SLES11

```
file:/etc/udev/rules.d/51-hsi-0.0.8027.rules
# Configure hsi device at 0.0.8027/0.0.8028/0.0.8029
ACTION=="add", SUBSYSTEM=="drivers", KERNEL=="qeth", IMPORT{program}="collect 0.
0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="add", SUBSYSTEM=="ccw", KERNEL=="0.0.8027", IMPORT{program}="collect 0.
0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="add", SUBSYSTEM=="ccw", KERNEL=="0.0.8028", IMPORT{program}="collect 0.
0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="add", SUBSYSTEM=="ccw", KERNEL=="0.0.8029", IMPORT{program}="collect 0.
0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="remove", SUBSYSTEM=="drivers", KERNEL=="qeth", IMPORT{program}="collect
 --remove 0.0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="remove", SUBSYSTEM=="ccw", KERNEL=="0.0.8027", IMPORT{program}="collect
 --remove 0.0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="remove", SUBSYSTEM=="ccw", KERNEL=="0.0.8028", IMPORT{program}="collect
 --remove 0.0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
ACTION=="remove", SUBSYSTEM=="ccw", KERNEL=="0.0.8029", IMPORT{program}="collect
 --remove 0.0.8027 %k 0.0.8027 0.0.8028 0.0.8029 qeth"
TEST=="[ccwgroup/0.0.8027]", GOTO="hsi-0.0.8027-end"
ACTION=="add", SUBSYSTEM=="ccw", ENV{COLLECT_0.0.8027}=="0", ATTR{[drivers/ccwgr
oup:qeth]group}="0.0.8027,0.0.8028,0.0.8029"
ACTION=="add", SUBSYSTEM=="drivers", KERNEL=="qeth", ENV{COLLECT_0.0.8027}=="0",
 ATTR{[drivers/ccwgroup:qeth]group}="0.0.8027,0.0.8028,0.0.8029"
LABEL="hsi-0.0.8027-end"
ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.8027", ATTR{layer2}="0"
ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.8027", ATTR{online}="1"
```

© 2012 IBM Corporation

QDIO Device Manual - SLES11

- /etc/udev/rules.d/70-persistent-net.rules

```
file:/etc/udev/rules.d/70-persistent-net.rules
# S/390 qeth device at 0.0.8027
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="qeth", KERNELS=="0.0.8027", ATTR{type
}=="1", KERNEL=="hsi*", NAME="hsi0"
```

© 2012 IBM Corporation

QDIO Device Manual - SLES 10

```
# modprobe qeth
# echo 0.0.8027,0.0.8028,0.0.8029 > /sys/bus/ccwgroup/drivers/qeth/group
# echo 1 > /sys/devices/qeth/0.0.8027/online
# ifconfig hsi0 10.40.33.38 netmask 255.255.255.0
```

- making it persistent
- create hwcfg-qeth-bus-ccw-0.0.8027 - copy from OSA file

```
#file:/etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.8027
STARTMODE='auto'
MODULE='qeth'
MODULE_OPTIONS=''
MODULE_UNLOAD='yes'
SCRIPTUP='hwup-ccw'
SCRIPTUP_ccw='hwup-ccw'
SCRIPTUP_ccwgroup='hwup-qeth'
SCRIPTDOWN='hwdown-ccw'
CCW_CHAN_IDS='0.0.8027 0.0.8028 0.0.8029'
CCW_CHAN_NUM='3'
CCW_CHAN_MODE='OSAPORT'
```

© 2012 IBM Corporation

QDIO Device Manual - SLES 10

- create ifcfg-qeth-bus-ccw-0.0.8027 - copy from OSA file

```
#file:/etc/sysconfig/network/ifcfg-qeth-bus-ccw-0.0.8027
BOOTPROTO='static'
UNIQUE=''
STARTMODE='onboot'
IPADDR='10.40.33.38'
NETMASK='255.255.255.0'

# hwup qeth-bus-ccw-0.0.8027
# ifup qeth-bus-ccw-0.0.8027

file:/etc/sysconfig/network/ifcfg.template
```

© 2012 IBM Corporation

QDIO Device Manual RHEL 6

- making it persistent

```
# modprobe qeth
# echo 0.0.8027,0.0.8028,0.0.8029 > /sys/bus/ccwgroup/drivers/qeth/group
# echo 1 > /sys/devices/qeth/0.0.8027/online
# ifconfig hsi0 10.40.33.38 netmask 255.255.255.0
```

- create ifcfg-hsi0 - copy from OSA file

```
#file:/etc/sysconfig/network-scripts/ifcfg-hsi0
DEVICE=hsi0
USERCTL=no
ONBOOT=yes
BOOTPROTO=static
NETMASK=255.255.255.0
IPADDR=10.40.33.38
PORTNAME=OSAPORT
SUBCHANNELS=0.0.8027,0.0.8028,0.0.8029
NETTYPE=qeth
TYPE=Ethernet
OPTIONS="layer2=0"
```

© 2012 IBM Corporation

QDIO Device Manual RHEL 6

- RHEL 6 generates the udev rule by magic once the device has been set online manually

```
file:/etc/udev/rules.d/70-persistent-net.rules
# S/390 qeth device at 0.0.8027
SUBSYSTEM=="net", ACTION=="add", DRIVERS=="qeth", KERNELS=="0.0.8027", ATTR{type}=="1"
```

© 2012 IBM Corporation

QDIO Device Manual - RHEL 5

- making it persistent

```
# modprobe qeth
# echo 0.0.8027,0.0.8028,0.0.8029 > /sys/bus/ccwgroup/drivers/qeth/group
# echo 1 > /sys/devices/qeth/0.0.8027/online
# ifconfig hsi0 10.40.33.38 netmask 255.255.255.0
```

```
#file:/etc/modprobe.conf
options dasd_mod dasd=6370
alias eth0 qeth
alias hsi0 qeth
```

© 2012 IBM Corporation

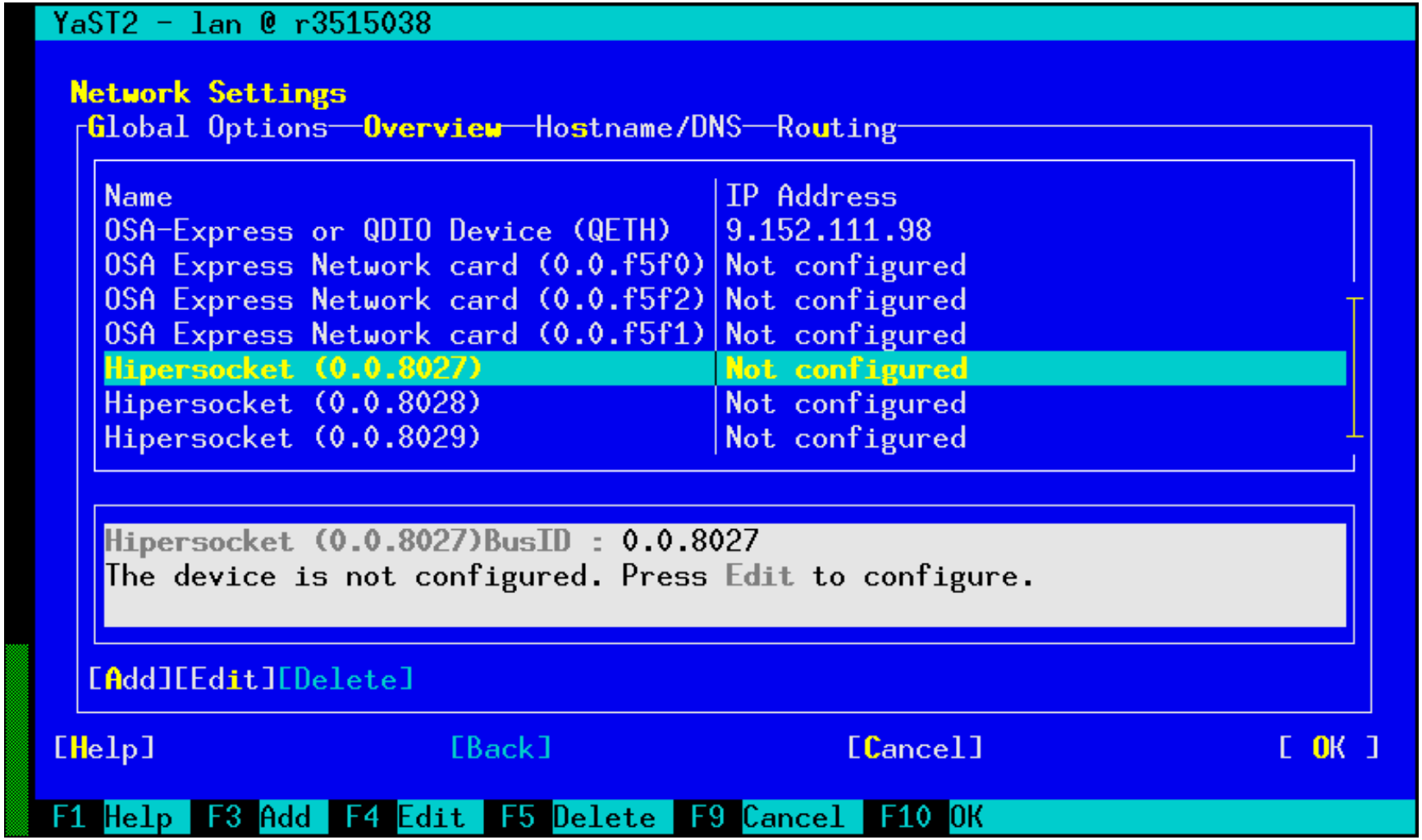
QDIO Device Manual - RHEL 5

- create ifcfg-hsi0 - copy from OSA file

```
#file:/etc/sysconfig/network-scripts/ifcfg-hsi0
DEVICE=hsi0
USERCTL=no
ONBOOT=yes
BOOTPROTO=static
NETMASK=255.255.252.0
IPADDR=10.40.33.38
PORTNAME=OSAPORT
SUBCHANNELS=0.0.8027,0.0.8028,0.0.8029
NETTYPE=qeth
TYPE=Ethernet
```

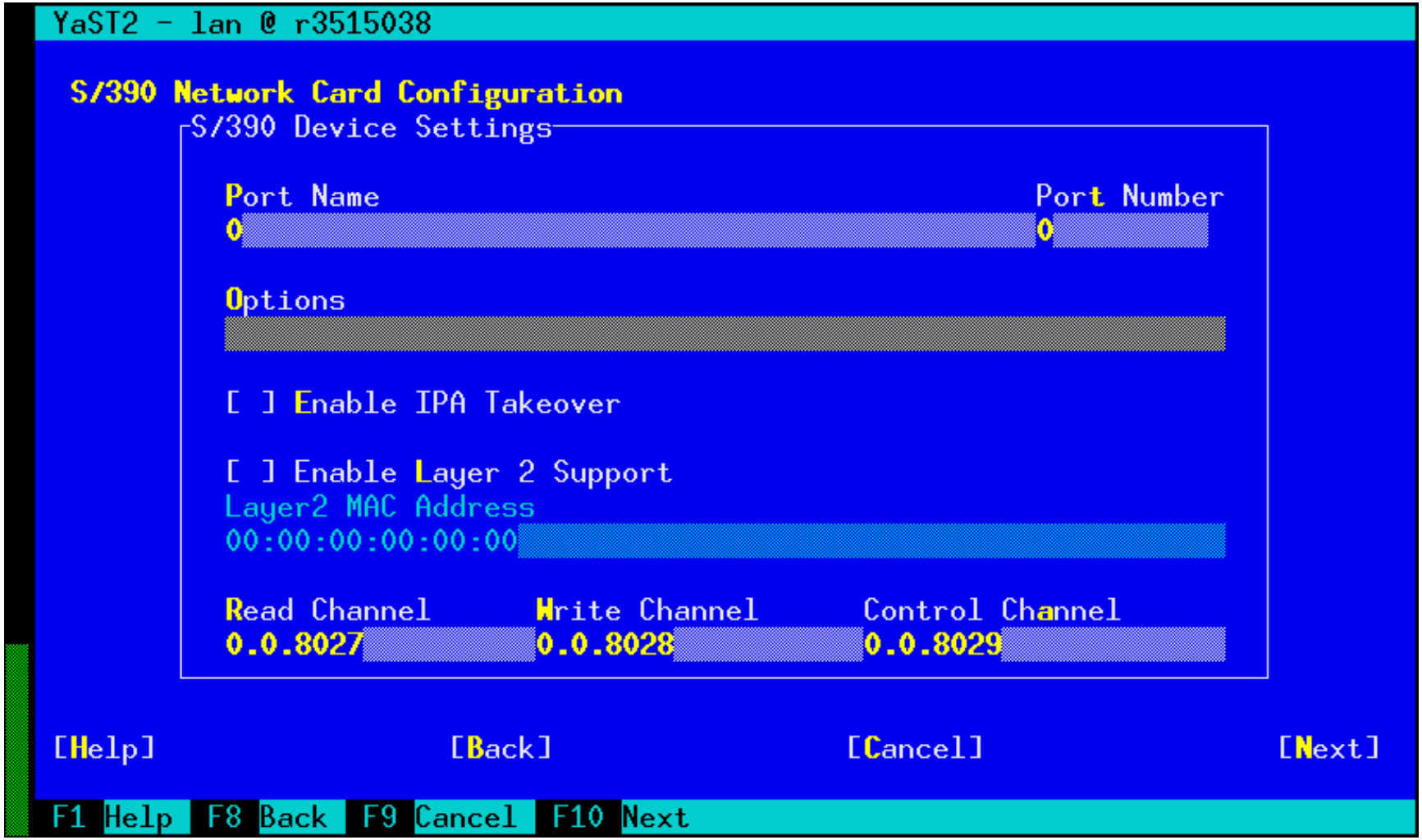
© 2012 IBM Corporation

QDIO Device YaST2 - SLES 11



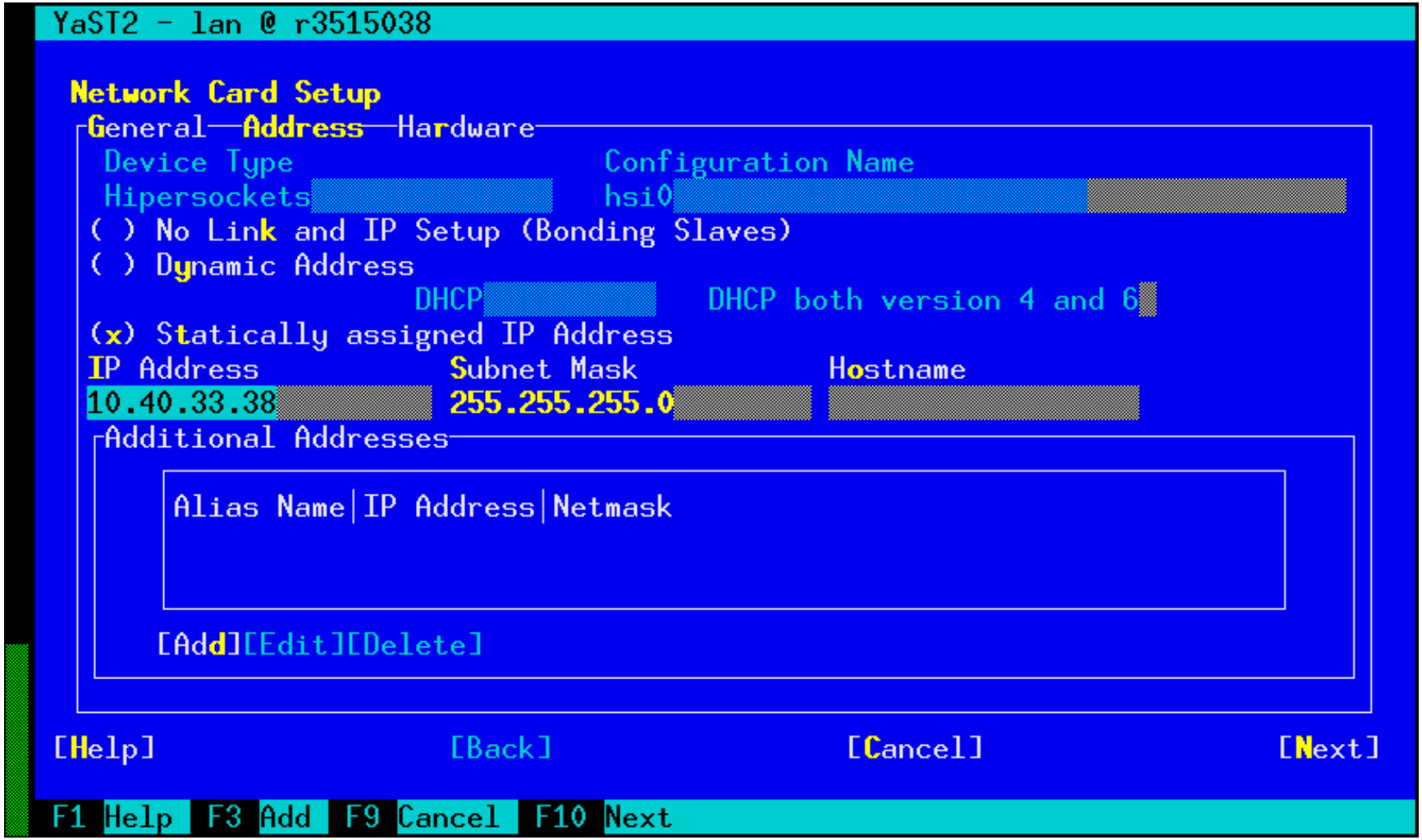
© 2012 IBM Corporation

QDIO Device YaST2 - SLES 11



© 2012 IBM Corporation

QDIO Device YaST2 - SLES 11



© 2012 IBM Corporation

QDIO Device YaST2 - SLES 10

- I would recommend the manual procedure
- failed for me on SLES10 SP3 and SP4
- don’t know why
- maybe fix missing

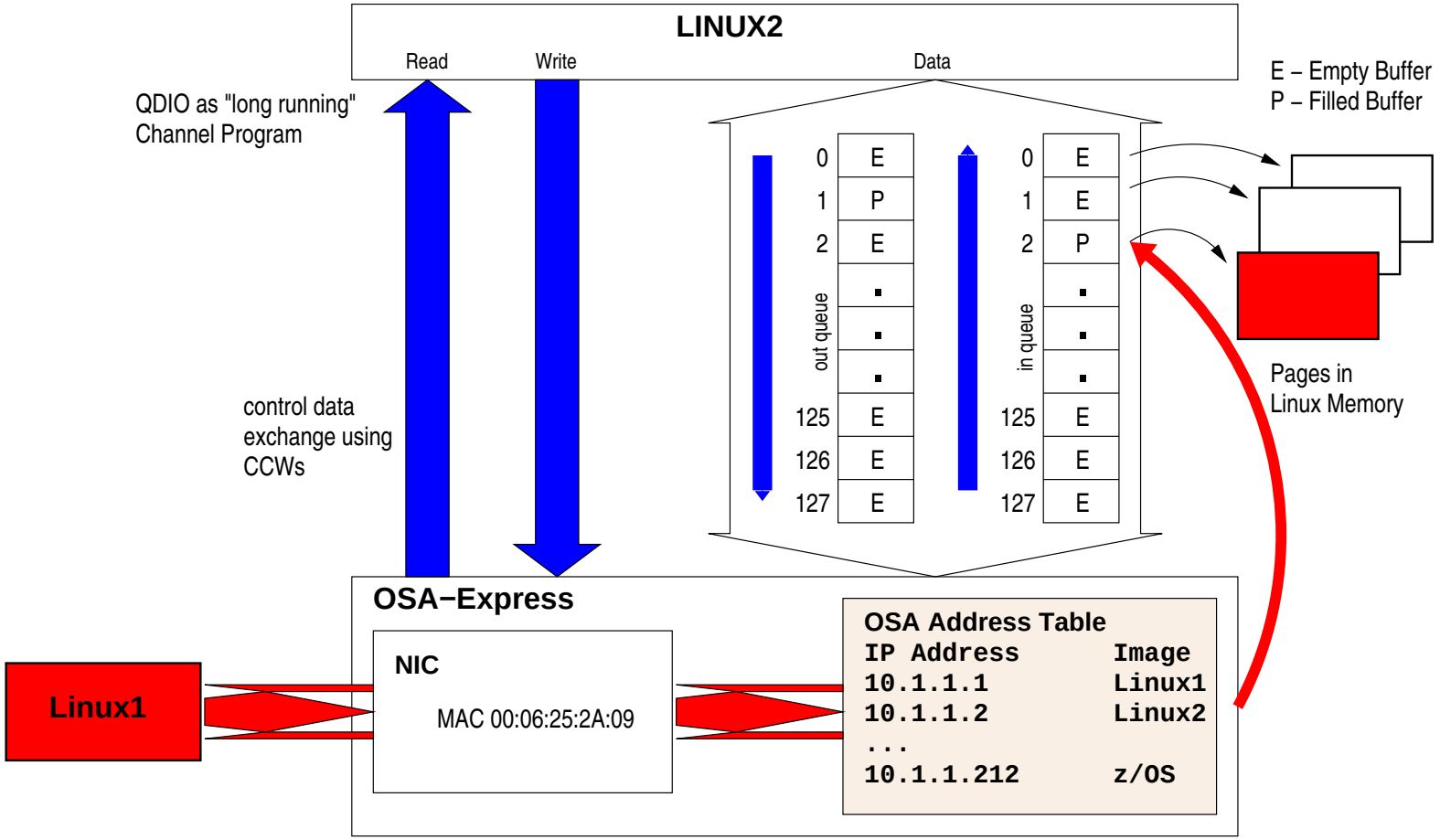
© 2012 IBM Corporation

QDIO - Queued Direct I/O

- OSA
- HiperSockets
- FCP

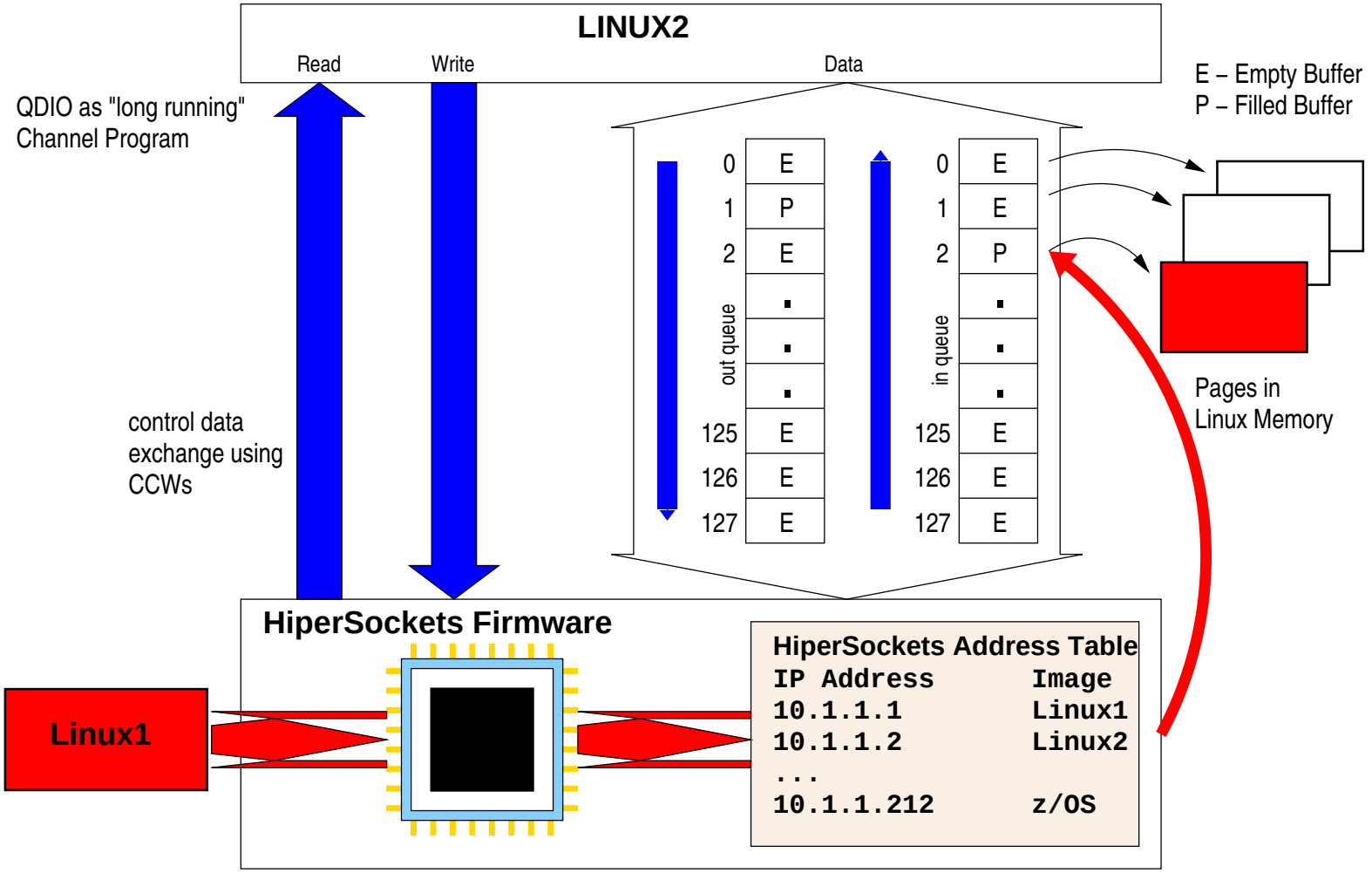
© 2012 IBM Corporation

QDIO Architecture



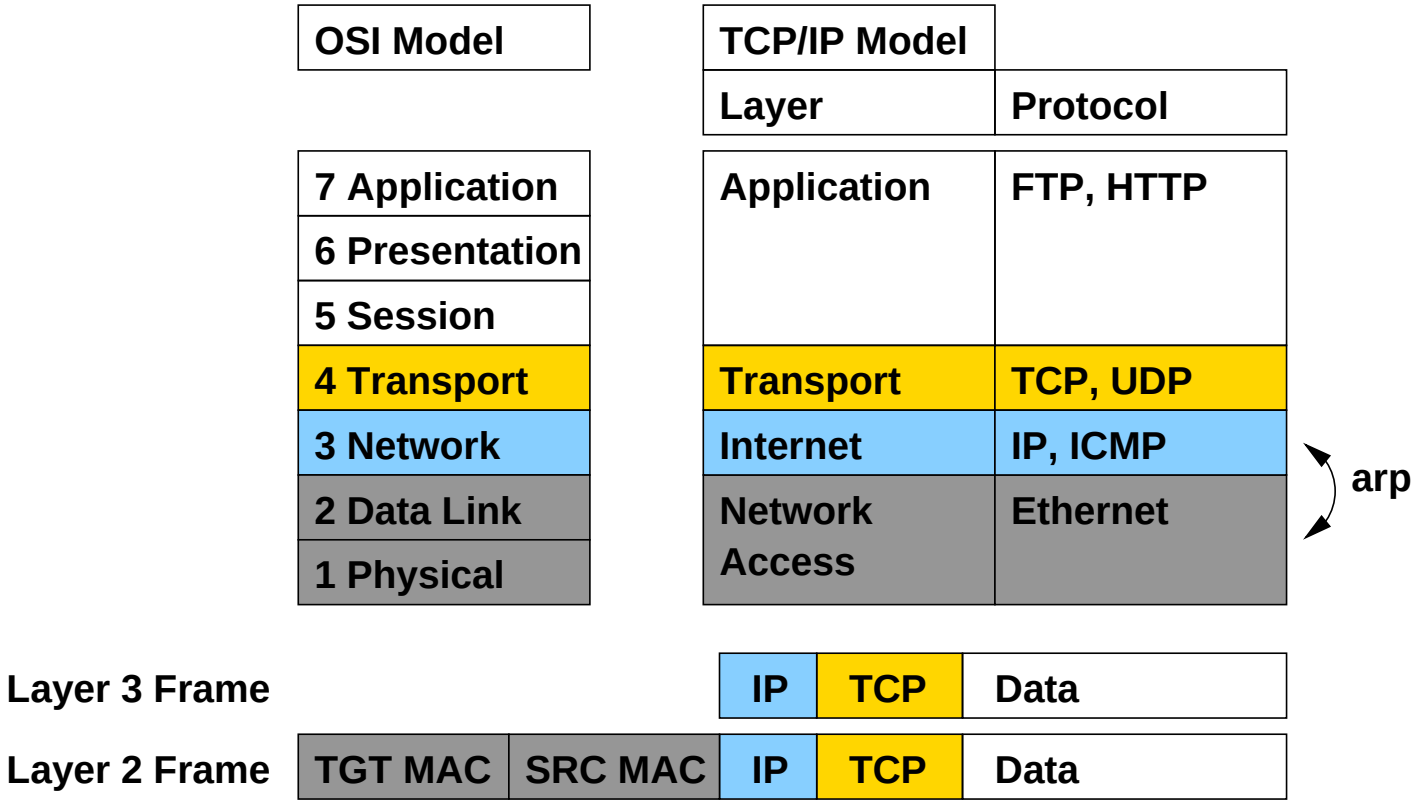
© 2012 IBM Corporation

QDIO Architecture

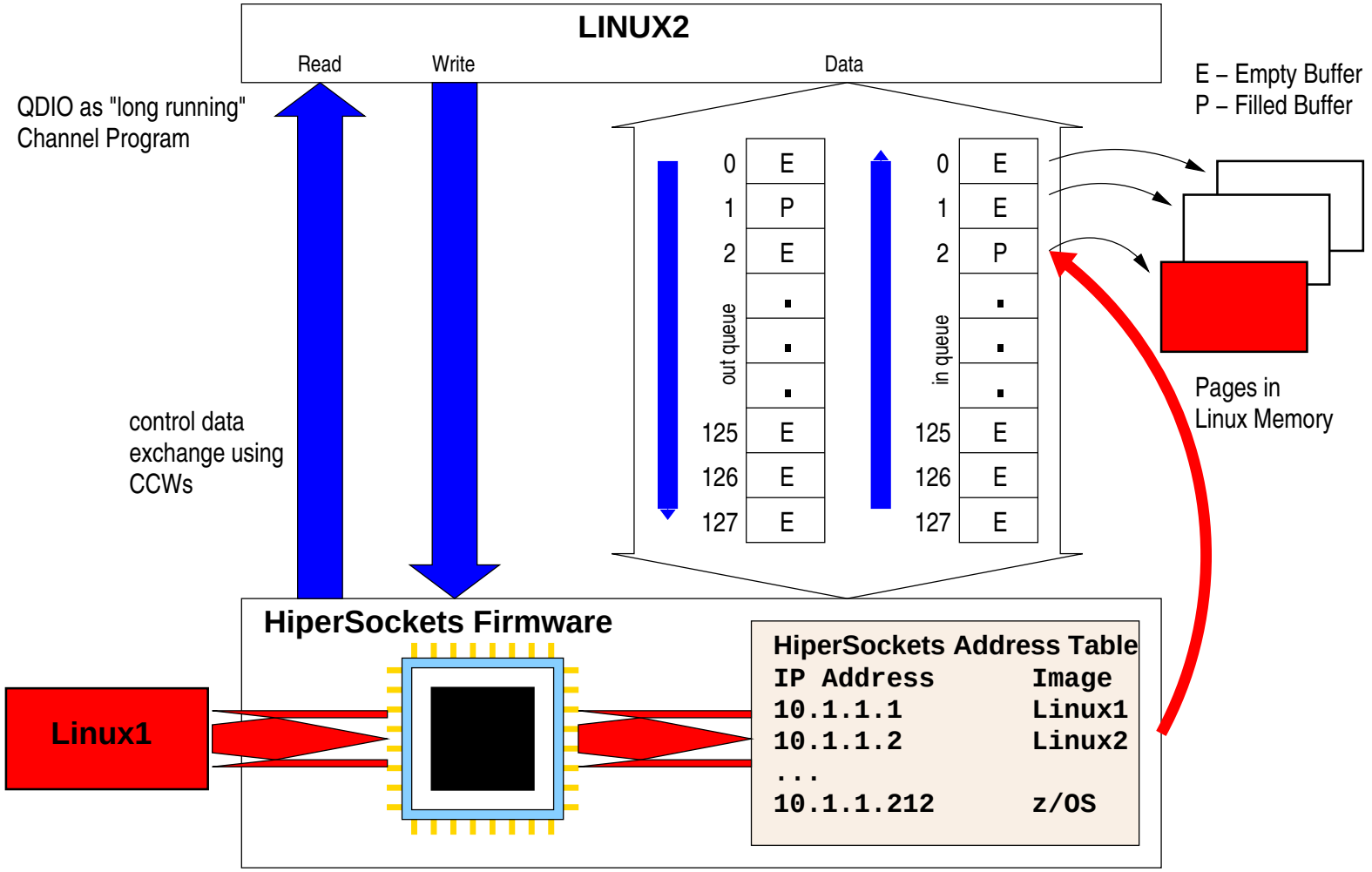


© 2012 IBM Corporation

Layer 2 Mode

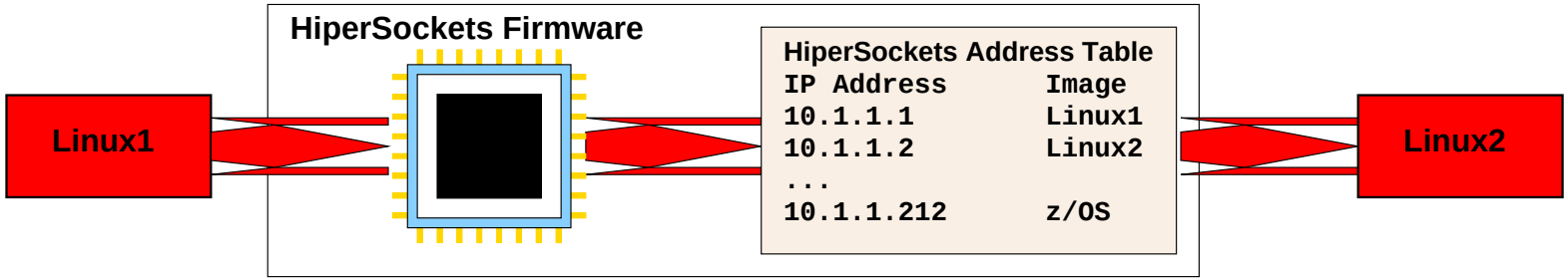

OSI Model
TCP/IP Model
Layer
Protocol
7 Application
6 Presentation
5 Session
4 Transport
3 Network
2 Data Link
1 Physical
Application
FTP, HTTP
Transport
TCP, UDP
Internet
IP, ICMP
Network Access
Ethernet
arp
Layer 3 Frame
IP
TCP
Data
Layer 2 Frame
TGT MAC
SRC MAC
IP
TCP
Data

Layer 2 Mode



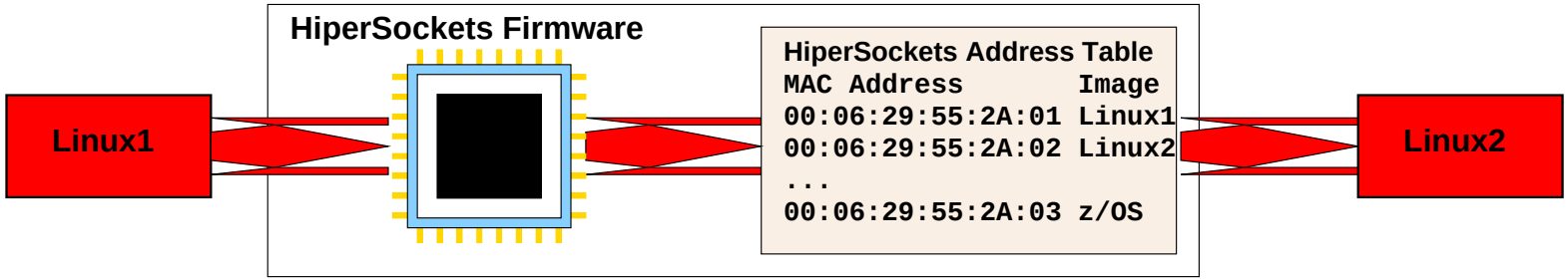
© 2012 IBM Corporation

Layer 2 Mode



© 2012 IBM Corporation

Layer 2 Mode



© 2012 IBM Corporation

Layer 2 Mode

- HiperSockets
  - Layer 2 and Layer 3 seperated
- Guest LAN
- VSWITCH

© 2012 IBM Corporation

Statistics

```
# cd /sys/kernel/debug/qdio/0.0.8029
# echo 1 > statistics

# cat statistics
Assumed adapter interrupts:      121424
           QDIO interrupts:      0
            Requested PCIs:      0
      Inbound tasklet runs:      0
   Inbound tasklet resched:      0
  Inbound tasklet resched2:      0
     Outbound tasklet runs:      15135
                 SIGA read:      1891
                SIGA write:      121030
                 SIGA sync:      0
              Inbound calls:     1891
            Inbound handler:     0
       Inbound stop_polling:     121029
         Inbound queue full:     0
             Outbound calls:     121030
           Outbound handler:     15135
        Outbound queue full:     0
     Outbound fast_requeue:      0
      Outbound target_full:      0
                QEBSM eqbs:      378223
          QEBSM eqbs partial:    19
                QEBSM sqbs:      243951
          QEBSM sqbs partial:    49
       Discarded interrupts:     395
```

© 2012 IBM Corporation

QDIO Assist

```
# vmcp q v osa
OSA   8027 ON OSA    8027 SUBCHANNEL = 0015
      8027 DEVTYPE HIPER       VIRTUAL CHPID FB IQD REAL CHPID FB
      8027 QDIO-ELIGIBLE       QIOASSIST-ELIGIBLE
OSA   8028 ON OSA    8028 SUBCHANNEL = 0016
      8028 DEVTYPE HIPER       VIRTUAL CHPID FB IQD REAL CHPID FB
      8028 QDIO-ELIGIBLE       QIOASSIST-ELIGIBLE
OSA   8029 ON OSA    8029 SUBCHANNEL = 0017
      8029                TOKEN      = 0000000EB3084B00
      8029 DEVTYPE HIPER       VIRTUAL CHPID FB IQD REAL CHPID FB
      8029 QDIO ACTIVE         QIOASSIST-ELIGIBLE      QEBSM
      8029
      8029 INP + 01 IOCNT = 00000000  ADP = 127 PROG = 000 UNAVAIL = 001
      8029          BYTES = 0000000000000000
      8029 OUT + 01 IOCNT = 00000000  ADP = 000 PROG = 000 UNAVAIL = 128
      8029          BYTES = 0000000000000000
      8029 OUT + 02 IOCNT = 00000000  ADP = 000 PROG = 000 UNAVAIL = 128
      8029          BYTES = 0000000000000000
      8029 OUT + 03 IOCNT = 00000000  ADP = 000 PROG = 000 UNAVAIL = 128
      8029          BYTES = 0000000000000000
      8029 OUT + 04 IOCNT = 00000000  ADP = 000 PROG = 000 UNAVAIL = 128
      8029          BYTES = 0000000000000000
```

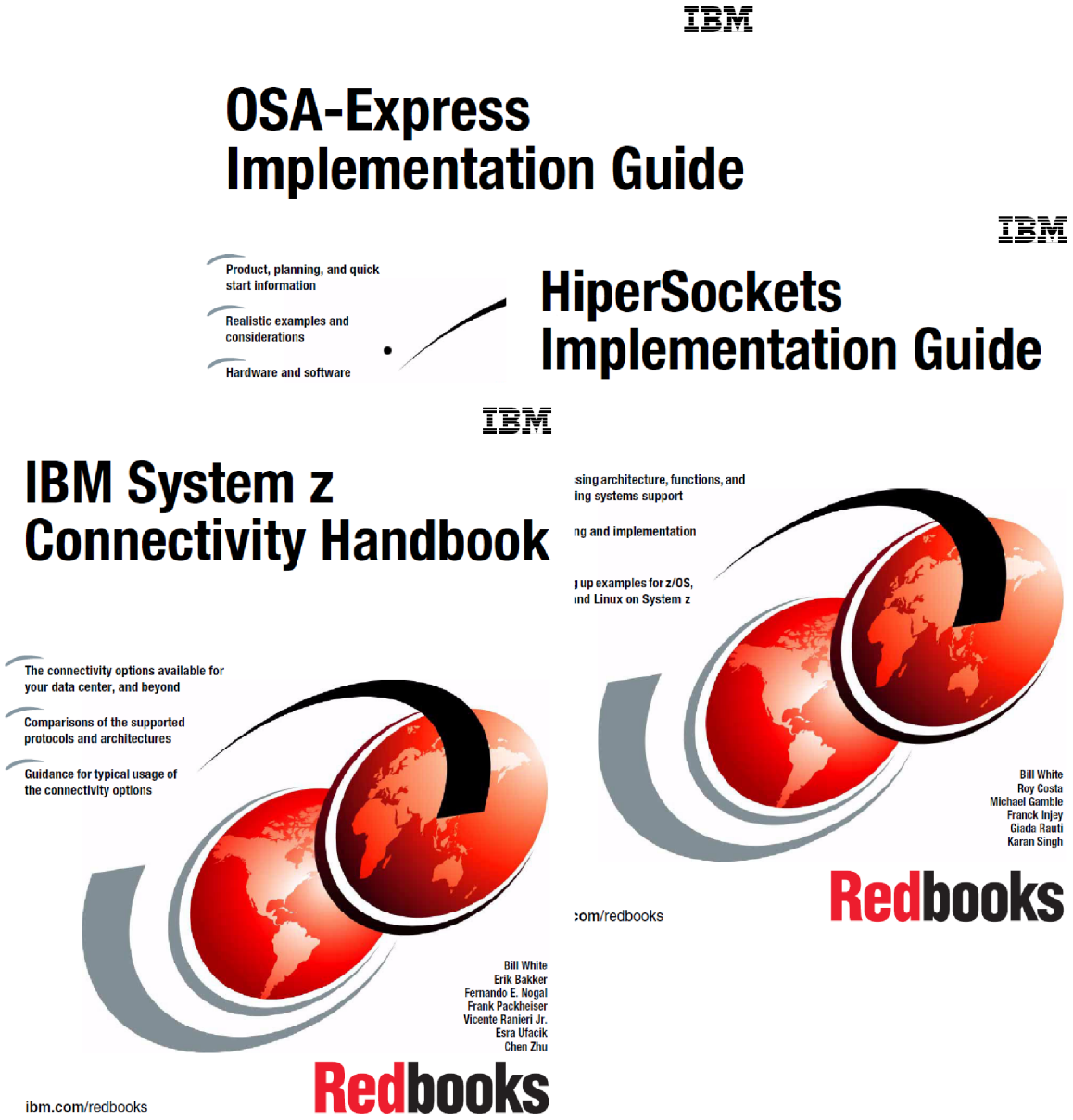
© 2012 IBM Corporation

QDIO Assist

- avoids interception to z/VM
- SQBS and EQBS are used to synchronize buffer states with z/VM storage
  - SQBS to write buffer state
  - EQBS to read buffer state
- big performance improvement because CPU always in guest level

© 2012 IBM Corporation

Redbooks



© 2012 IBM Corporation

Links

- Linux on System z - Tuning Hints & Tips
  `http://www.ibm.com/developerworks/linux/linux390/perf/index.html`
- developerWorks
  `http://www.ibm.com/developerworks/linux/linux390`
- IBM Redbooks
  `http://www.redbooks.ibm.com`

© 2012 IBM Corporation

Thank You !

For starting out with their very good presentations

- Susanne Wintenberger
- Mario Held

© 2012 IBM Corporation

Questions ?

Dr. Stefan Reimbold
Diplom-Physiker

Linux on System z Service

Schoenaicher Strasse 220
D-71032 Boeblingen
Mail: Postfach 1380
D-71003 Boeblingen

Phone +49-7031-16-2368
Stefan.Reimbold@de.ibm.com

© 2012 IBM Corporation