IBM

z/VSE Fast Path to Linux on System z - in z/VM or LPAR environment

Ingo Franzki, IBM

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®, 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.

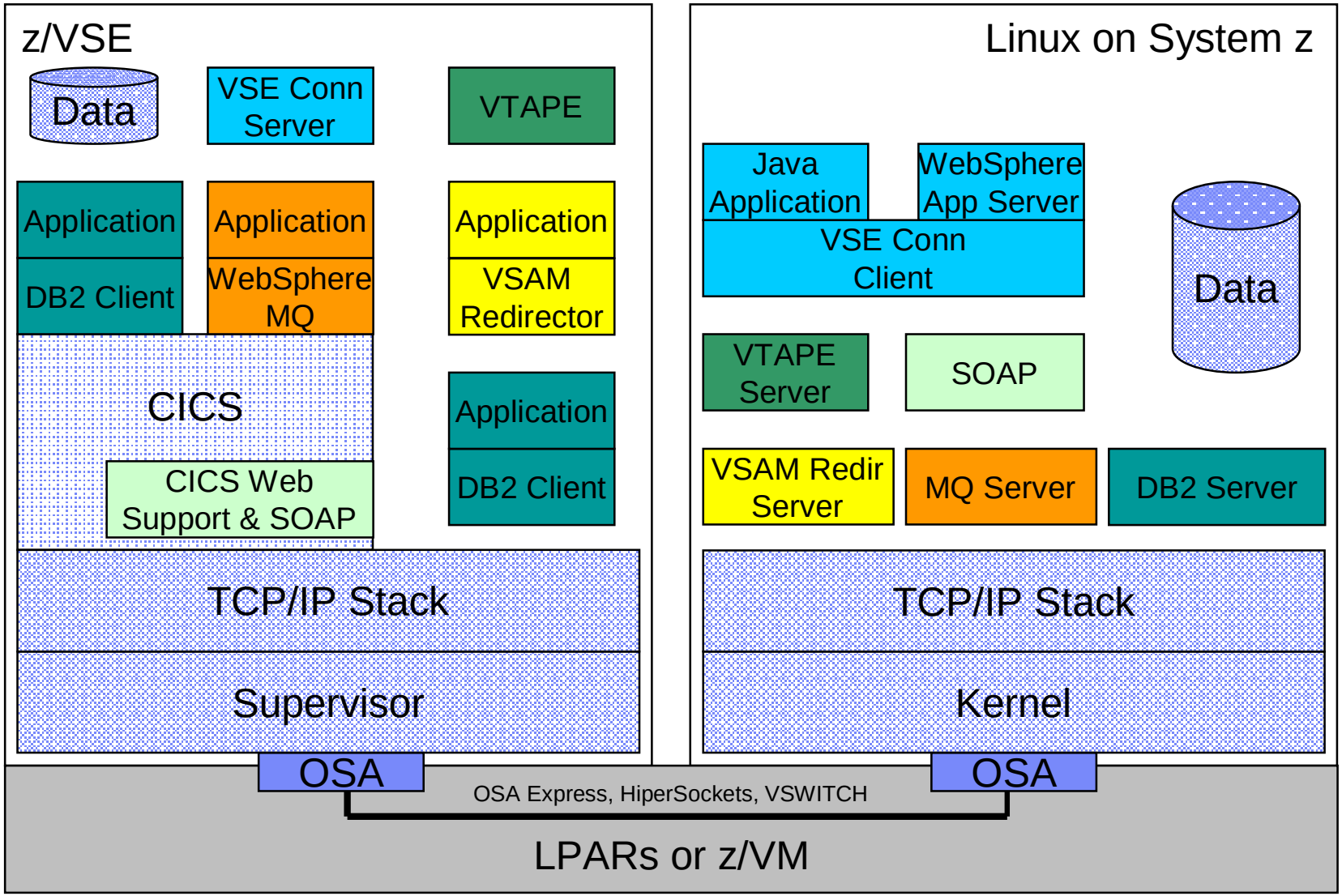
* All other products may be trademarks or registered trademarks of their respective companies.

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.

Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

- Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at http://www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”).
- No other workload processing is authorized for execution on an SE.
- IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

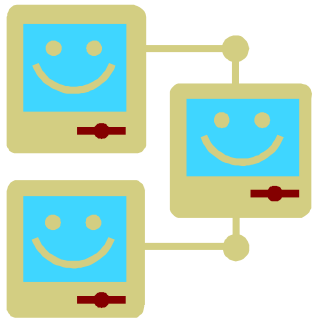
z/VSE Applications communicating with Applications on Linux



z/VSE Applications communicating with Applications on Linux

§ Communication is mostly based on TCP/IP

– Although z/VSE and Linux run on the same box



§ TCP/IP

– Allow reliable communication over a non-reliable network

– Uses sequence numbers, acknowledges, checksums

- To protect against packet loss, duplicate packets, packet sequence errors, damaged or incomplete packets, etc.

à Time consuming processing

§ When z/VSE and Linux run side by side on the same box

– Why do we need all this expensive processing in this case?

– There should be a more direct communication method !

à z/VSE Fast Path to Linux on System z

(for short: Linux Fast Path or just LFP)

Fast Path to Linux on System z (LFP)

- **Allows selected TCP/IP applications to communicate with the TCP/IP stack on Linux without using a TCP/IP stack on z/VSE**
- **All socket requests are transparently forwarded to a Linux on System z system running in the same z/VM**

à Linux Fast Path in a z/VM environment

- Both z/VSE and Linux on System z run as z/VM Guests in the same z/VM-mode LPAR on IBM z10, z114 or z196 servers
- Uses an IUCV connection between z/VSE and Linux

à Linux Fast Path in an LPAR environment

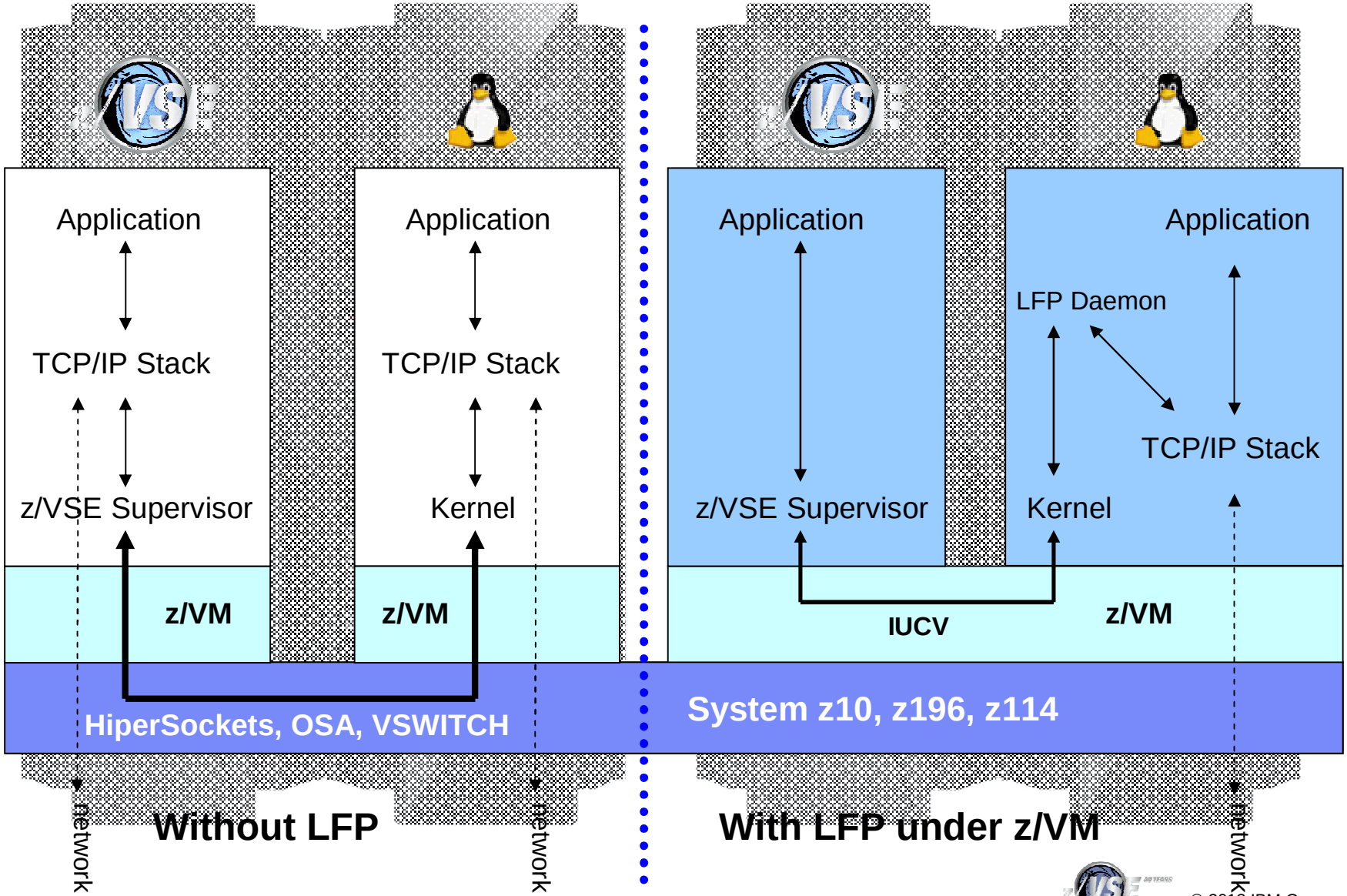
- Both z/VSE and Linux on System z run in their own LPARs on a zEnterprise server
- A HiperSockets connection is used between z/VSE and Linux on System z
- LFP requires the HiperSockets Completion Queue function that is available with a zEnterprise server (z196, z114)

- **The fast path to Linux on System z provides standard TCP/IP socket APIs for programs running on z/VSE**
  - Other than the basic socket API, no other tools are provided
  - Since z/VSE V5.1: LFP supports IPv6 

- **Possible performance increase due to:**
  - Less overhead for TCP/IP processing on z/VSE (TCP, sequence numbers and acknowledging, checksums, resends, etc)
  - More reliable communication method (IUCV) compared to HiperSockets, which is a network device, with all its packet drops, resends, etc.

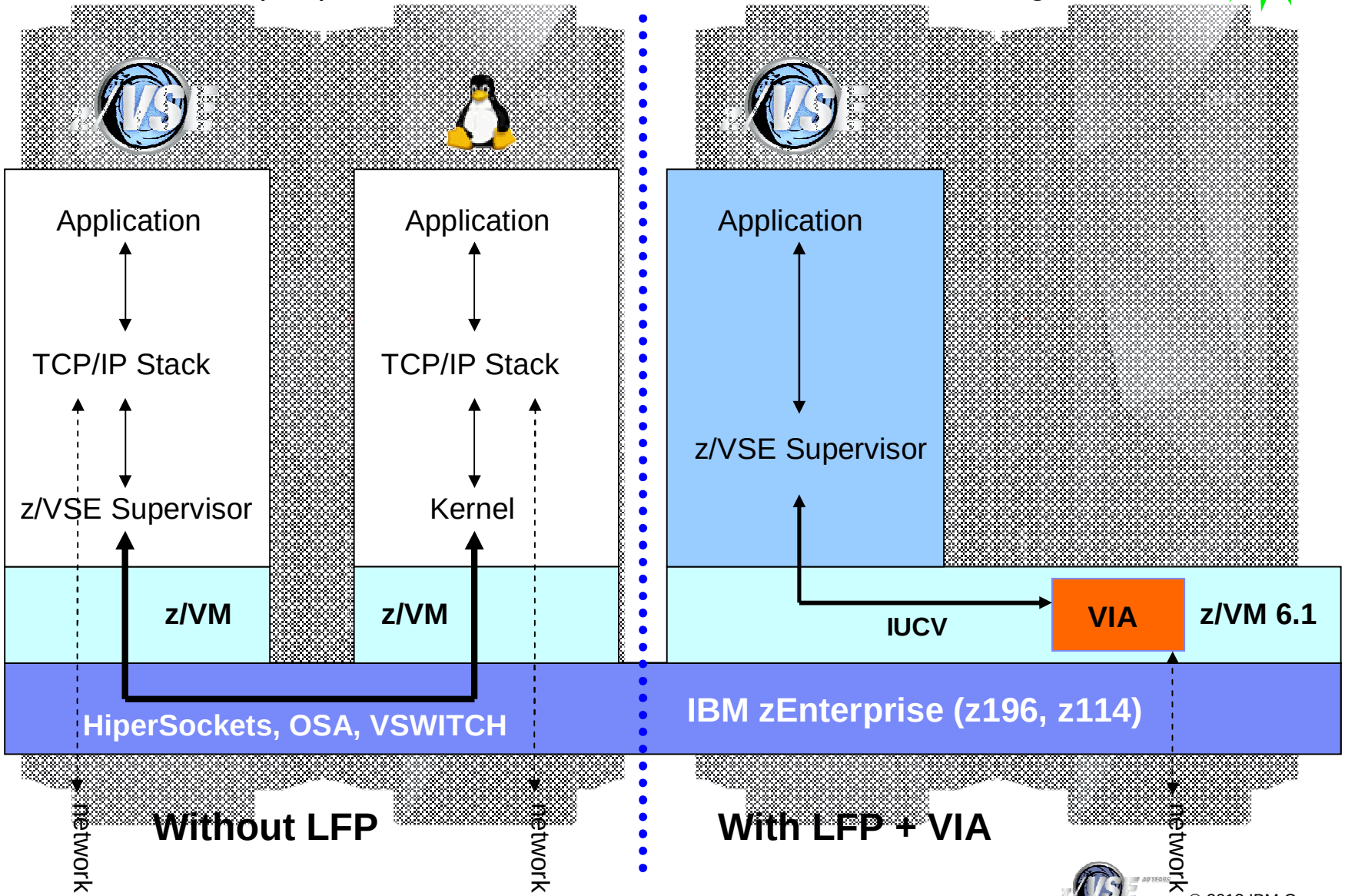
Linux Fast Path in a z/VM environment (z/VSE 4.3 or later)

Faster communication between z/VSE and Linux applications



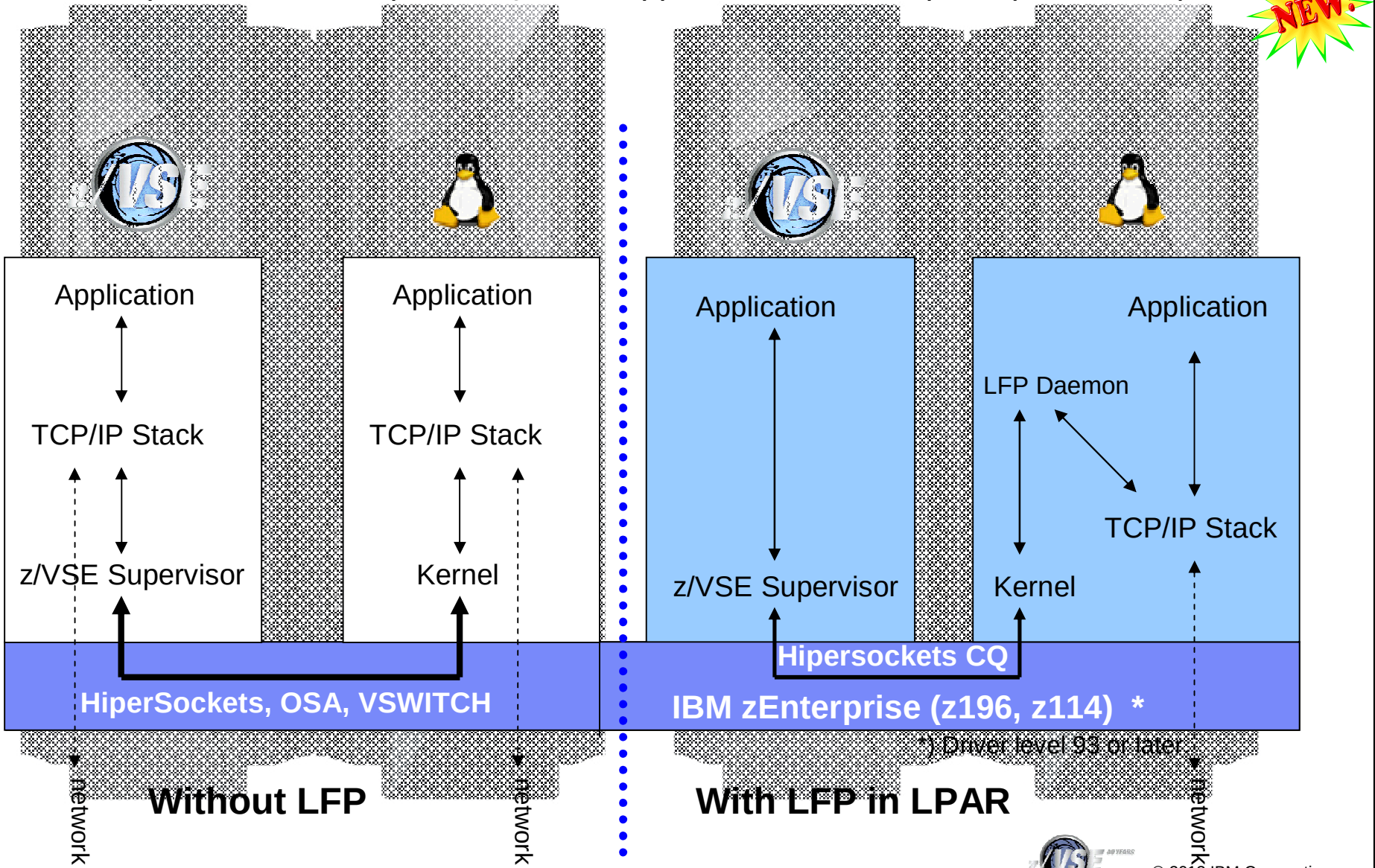
New: z/VSE z/VM IP Assist (VIA) (z/VSE 5.1 + z/VM 6.1)

With z/VM IP Assist (VIA), no Linux is needed to utilize the LFP advantage



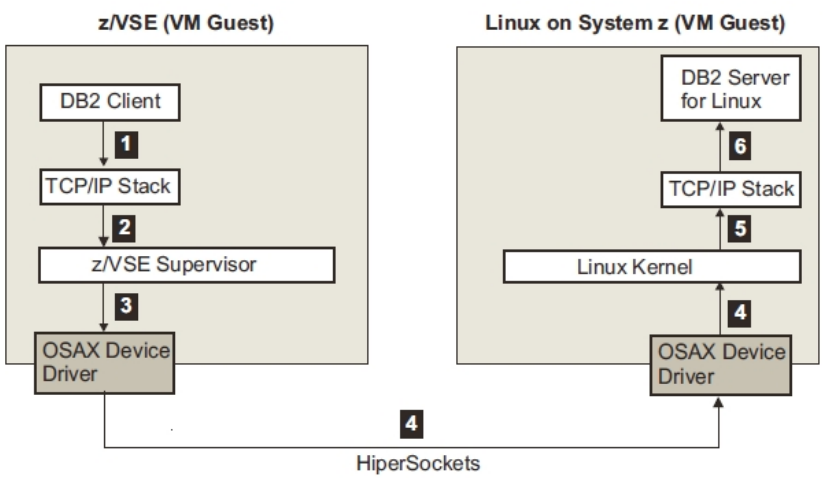
New: Linux Fast Path in an LPAR environment (z/VSE 5.1 + PTFs)

Exploits the HiperSockets Completion-Queue support of IBM zEnterprise (z196, z114)



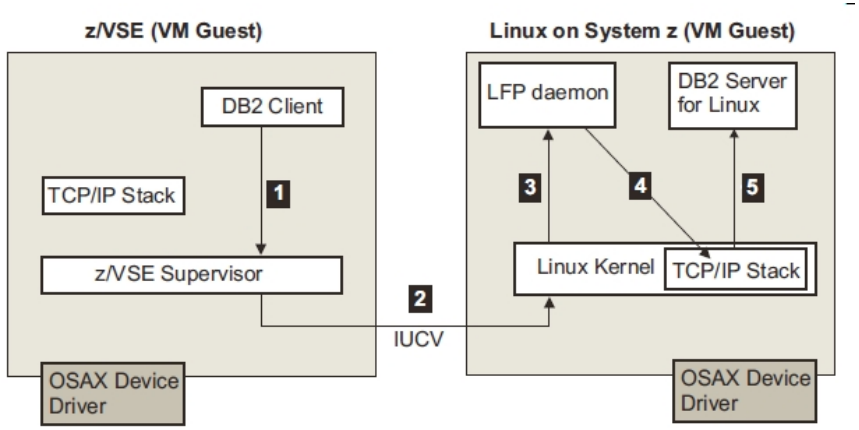
Communication flows when using Linux Fast Path

Using a TCP/IP stack (CSI/BSI):

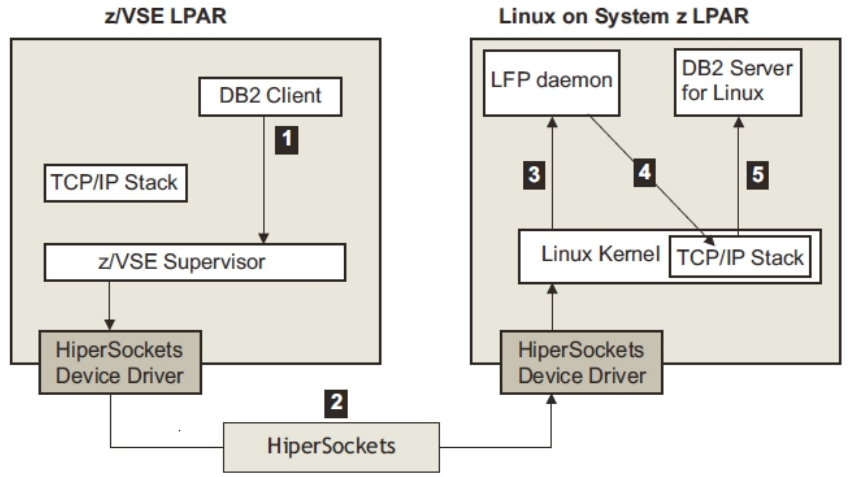


- **Less overhead for TCP/IP processing on z/VSE**
  - Building of IP and TCP packets
  - Sequence numbers and acknowledging
  - Checksums
  - Retransmission of lost packets
- **More reliable communication method compared to a traditional network device**
  - IUCV is a reliable communication method (z/VM)
  - HiperSockets Completion Queue support allows to build a reliable communication path (LPAR)

Using Linux Fast Path in a z/VM environment:



Using Linux Fast Path in an LPAR environment:



Prerequisites for using the Linux Fast Path

In a z/VM environment:

- **Any IBM System z server supported by z/VSE**
- **Any supported z/VM version/release**
- **The use of a z/VM-mode LPAR is recommended**
  - z/VM-Mode LPAR is only available on z10 or later and z/VM 5.4 or later

- **z/VSE 4.3 or later**

- **One of these Linux on System z operating systems:**
  - SUSE Linux Enterprise Server 10 Service Pack 3 together with security update kernel 2.6.16.60-0.57.1
  - SUSE Linux Enterprise Server 11 Service Pack 1
  - Red Hat Enterprise Linux 5 Update 5
  - Red Hat Enterprise Linux 6

- **z/VSE and Linux on System z are configured as z/VM guests within the same LPAR**

- **The IUCV ("Inter-User Communication Vehicle") is configured and enabled in both z/VM guests (z/VSE and Linux on System z)**

Prerequisites for using the Linux Fast Path

In an LPAR environment:

- **A zEnterprise server (z196 or z114) at driver level 93 or later**
  - LFP requires the HiperSockets Completion Queue function, which is only available with a zEnterprise server

- **z/VSE 5.1 + PTFs**

- **One of these Linux on System z operating systems:**
  - SUSE Linux Enterprise Server 11 Service Pack 2
  - Red Hat: IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases

- **z/VSE and Linux on System z both run in LPAR mode**

- **A HiperSockets Connection between z/VSE and Linux systems**

Preparing the system for Linux Fast Path in z/VM environment

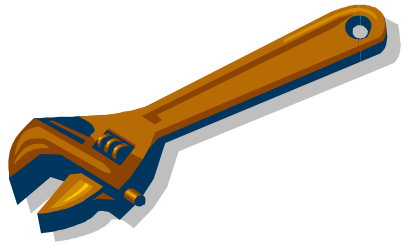
§ Preparing the LPAR

– For use with LFP in z/VM environment, the Linux on System z and z/VSE must run under the same z/VM system

– The use of a z/VM Mode-LPAR is recommended

- Allows you to mix CPs and IFL in one z/VM Installation
- Linux runs on IFLs
- z/VSE runs on CPs

– Change the LPAR Mode to z/VM-Mode and add the IFLs to it



§ Preparing z/VM

– LFP uses IUCV as the underlying communication vehicle. Therefore the z/VSE and the Linux on System z guests on the z/VM system need to be configured for IUCV.

– The following z/VM parameters for the guest systems are relevant:

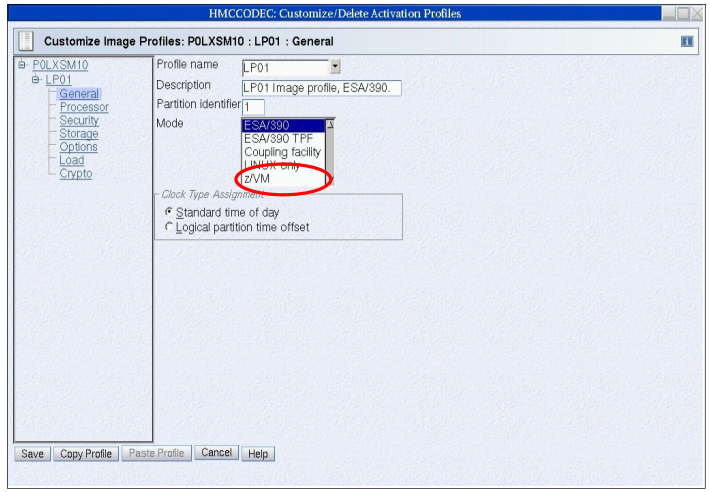
- IUCV ALLOW
- IUCV ANY
- IUCV MSGLIMIT
- OPTION MAXCONN *maxno*

– For details about the parameters check the z/VM documentation.

z/VM-Mode LPAR

LPAR Modes on z10 and later:

Logical partition mode	PU type	Operating systems	PUs usage
ESA/390	CPs	z/Architecture operating systems ESA/390 operating systems Linux	CPs DED *or* CPs SHR
	CPs *and* zAAPs *or* zIIPs	z/OS z/VM (V5.3 and later, for guest exploitation)	CPs DED *and* zAAPs DED, *and/or* zIIPs DED *or* CPs SHR *and* zAAPs SHR *and/or* zIIPs SHR
ESA/390 TPF	CPs	TPF z/TPF	CPs DED *or* CPs SHR
Coupling facility	ICFs *or* CPs	CFCC	ICFs DED *or* ICFs SHR, *or* CPs DED *or* CPs SHR
Linux only	IFLs or CPs	Linux z/VM	IFLs DED *or* IFLs SHR, *or* CPs DED *or* CPs SHR
z/VM	CPs, IFLs, zAAPs, zIIPs ICFs	z/VM V5.4 or later	All PUs must be either SHR *or* DED



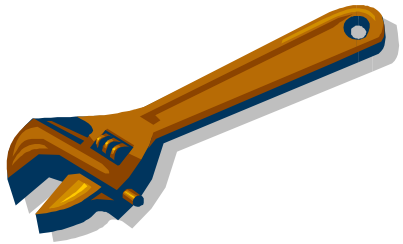
Preparing to use Linux Fast Path

§ Preparing Linux on System z

- Download and install the LFP Daemon
  - Part of the “z/VSE Connector Workstation Code“ component 5686-CF8-38 / 51P
  - Member IJBLFPLX.W from PRD2.PROD or download from Internet
  - This ZIP file contains an RPM (RPM Package Manager) that can be used to install the LFPD
- Configure one or multiple LFPD Instances
  - Textual configuration files in /etc/opt/ibm/vselfpd/confs-available and /etc/opt/ibm/vselfpd/confs-enabled
- It is recommended to use separate (virtual) network adapters or at least separate IP addresses for each LFPD Instance (give each VSE its own IP address)
- Start LFP daemon using lfpd-ctl or automatically at boot via init.d start script

§ Preparing z/VSE

- The LFP code is part of the z/VSE system, no installation step needed
- Start and configure an LFP Instance
  - Textual configuration statements in LIBR member or SYSIPT of start job
  - LFP Instance operation via IJBLFPOP tool
- LFP does not require a partition to run
- Every LFP Instance is identified by a 2 digit number (System ID)
  - Same concept as used by TCP/IP stacks



Sample configuration on z/VSE

For LFP in z/VM Environment:

```
* $$ JOB JNM=LFPSTART,CLASS=0,DISP=L
// JOB LFPSTART
// EXEC IJBLFPOP,PARM='START DD:SYSIPT LOGALL'
ID = 01
MTU = 8192
IucvMsgLimit = 1024
InitialBufferSpace = 512K
MaxBufferSpace = 4M
IucvSrcAppName = TESTV
IucvDestAppName = LINR02
IucvDestVMId = LINLFP
WindowSize = 65535
WindowThreshold = 25
/*
/&
* $$ EOJ
```

IJBLFPOP will read input from SYSIPT

IUCV Name of LFP on z/VSE

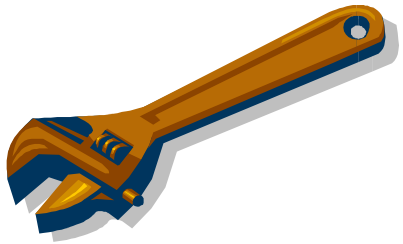
IUCV Name of LFPD on Linux

Guest name where Linux runs

z/VSE Skeletons for use with LFP

§ The following skeletons are available in ICCF library 59 for use with LFP:

Skeleton	Description
SKLFPSTA	Start an LFP Instance
SKLFPSTO	Stop an LFP Instance
SKLFPLST	List all active LFP Instances
SKLFPINF	Query information about an active LFP Instance
SKLFPACT	Contains control statements to activate LFP you many need to include into the JCL of your applications



Operating an Linux Fast Path on z/VSE

§ List active LFP Instances

- // EXEC IJBLFPOP,PARM='LIST'
- LFPB025I ACTIVE LFP INSTANCES: 1
  INSTANCE 01 HAS 3 ACTIVE TASKS
  LFPB026I END OF ACTIVE LFP INSTANCES LIST

§ Display information about an active instance

- // EXEC IJBLFPOP,PARM='INFO <INSTID> [SHOWTASKS] [LOGALL]'
- LFPB023I INFO ABOUT LFP INSTANCE '01':

```
*** INSTANCE ***
  STATUS ........................ : UP
  WINDOW SIZE ................... : 65,535
  ...
*** DEVICE ***
  DEVICE STATUS ................. : ACTIVE
  PACKETS WAITING FOR MSG COMPLETE : 0
  MAXIMUM PACKETS USED .......... : 37
  ...
*** TASKS ***
  ACTIVE TASK COUNT ............. : 3
 -- TASK #1 -
  TASK ID (PARTITION ID)......... : 2E (Z1)
  SOCKET COUNT .................. : 1
  L2 SOCKET LIST COUNT .......... : 1
  ...
LFPB024I END OF INFO ABOUT LFP INSTANCE '01'.
```

Sample configuration on Linux on System z

lfpd-LINR02.conf:

```
# lfpd configuration file
IUCV_SRC_APPNAME = LINR02
# ensure that only TESTV from VSER05 can connect
PEER_IUCV_VMID = VSER05
PEER_IUCV_APPNAME = TESTV
IUCV_MSGLIMIT = 1024
MTU_SIZE = 8192
MAX_SOCKETS = 1024
INITIAL_IO_BUFS = 128
WINDOW_SIZE = 65535
WINDOW_THRESHOLD = 25
VSE_CODEPAGE = EBCDIC-US
VSE_HOSTID = 10.0.0.1
RESTRICT_TO_HOSTID = yes
LOG_INFO_MSG = no
```

IUCV Name of LFPD on Linux

Guest name where z/VSE runs

IUCV Name of LFP on z/VSE

This is the IP address VSE will appear under

Note: The configuration file must be named “lfpd-XXX”, where XXX is the IUCV_SRC_APPNAME specified in the configuration file !
The XXX characters in the filename must be specified in uppercase !

Operating an Linux Fast Path on Linux on System z

§ Display LFP daemon status

– **lfpd-admin <--iucv_appname|-i appname> <--status|-s>**

```
Status:
-------
  z/VSE instance is connected.
  Peer VM ID ......... : VSER05
  Peer IUCV Appl. name : TESTV
  Applied host id .... : 10.0.0.1
  Applied host name .. : linlfp
  Allocated I/O buffers ....... : 128
  ...
  Number of active z/VSE tasks : 1
  Number of active sockets : 1
Trace Status:
-------------
  Running in daemon mode
  No trace is running
Configuration:
--------------
  LOCAL_IUCV_APPNAME = LINR02
  PEER_IUCV_VMID = VSER05
  PEER_IUCV_APPNAME = TESTV
  MAX_VSE_TASKS = 512
  MTU_SIZE = 8192
  MAX_SOCKETS = 1024
  INITIAL_IO_BUFS = 128
  WINDOW_SIZE = 65536
  WINDOW_THRESHOLD = 25% (16384 bytes)
  ...
-----------------
```

Socket API Support of Linux Fast Path

§ Linux Fast Path supports the following Socket APIs

- LE/C Socket API
- EZA SOCKET and EZASMI
- CSI's SOCKET Macro (limited support)

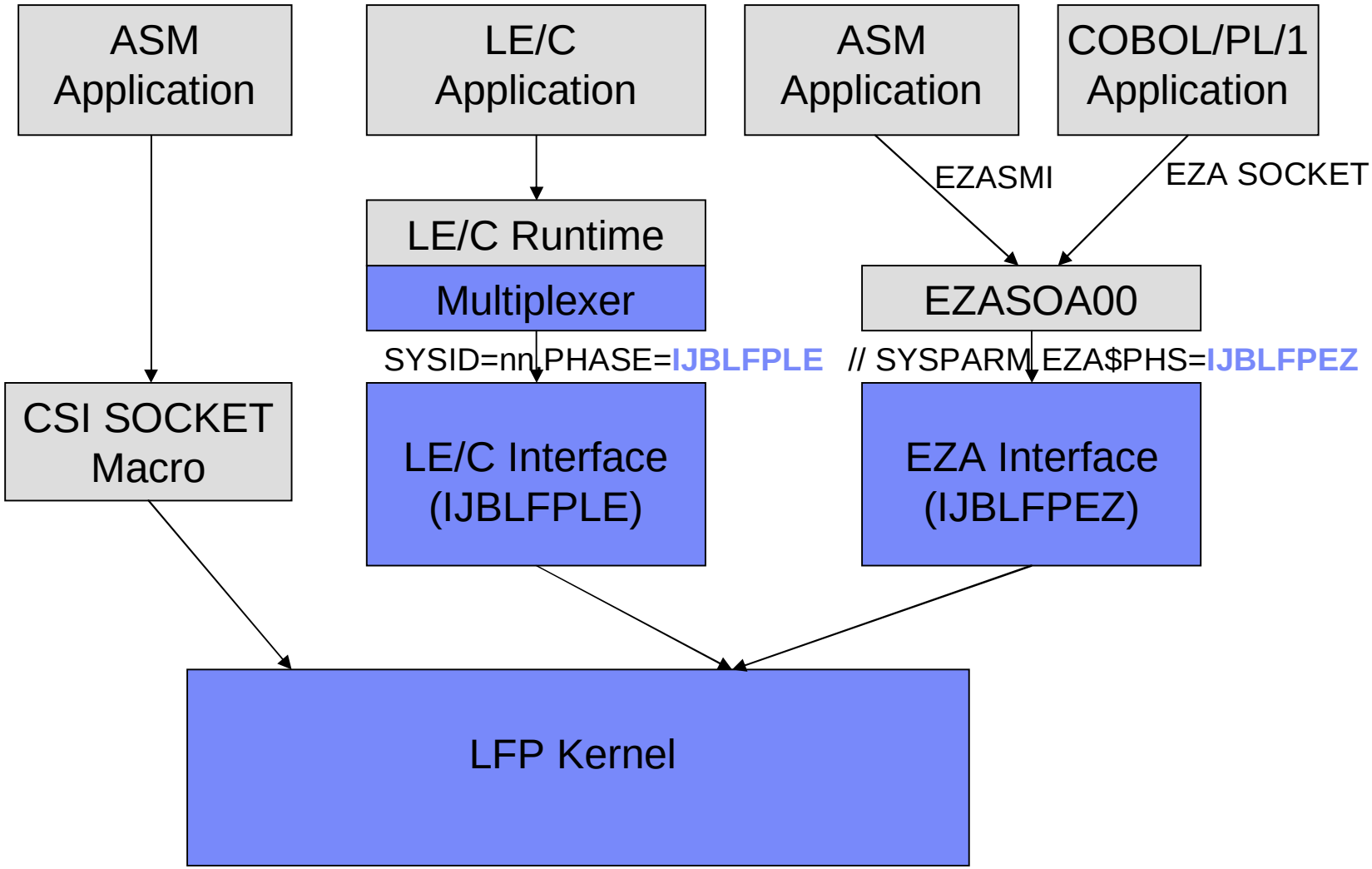
§ LE/C Socket API considerations

- The LE/C interface phase for LFP is shipped as IJBLFPLE.PHASE in IJSYSRS.SYSLIB
- You must configure the LE/C TCP/IP Socket API Multiplexer to use the LFP LE/C TCP/IP interface phase IJBLFPLE for the IDs of all LFP instances that are running
- To configure the multiplexer, use skeleton EDCTCPMC in ICCF library 62
- You can add entries for all your LFP instances with the following statement:
  - EDCTCPME SYSID='01',PHASE='IJBLFPLE'

§ EZA SOCKET and EZASMI considerations

- With the EZA socket and EZASMI interfaces you can specify which interface module is to be used
- For LFP, you must use the EZA interface module IJBLFPEZ
- You must set the JCL parameter "EZA$PHA" in all your jobs that you want to use LFP
- To do so use the following statement in your jobs:
  - // SETPARM [SYSTEM] EZA$PHA=IJBLFPEZ
- If you are using the EZA SOCKET or EZASMI interface under CICS, you need to activate the EZA 'TASK-RELATED-USER-EXIT' (TRUE)

Socket API Support of Linux Fast Path



LE/C Socket API Multiplexer

- **Different Stacks use different Interface routines**
  - TCP/IP for VSE (CSI/IBM): $EDCTCPV
  - Linux Fast Path: IJBLFPLE
  - IPv6/VSE (BSI/IBM): BSTTTCPV

- **Avoid complicated setup using specific LIBDEFs for different stacks**
- **Interface phase is selected by System ID**
- **Use skeleton EDCTCPMC in ICCF library 62**

```
// EXEC ASMA90,SIZE=(ASMA90,64K),PARM='EXIT(LIBEXIT(EDECKXIT)),SIZE(MAXC
                -200K,ABOVE)'
EDCTCPMC CSECT
EDCTCPMC AMODE ANY
EDCTCPMC RMODE ANY
*
         EDCTCPME SYSID='00',PHASE='$EDCTCPV'
         EDCTCPME SYSID='01',PHASE='IJBLFPLE'
         EDCTCPME SYSID='02',PHASE='BSTTTCP6'
*
         END
/*
```

Specifying the System ID (Instance ID)

- **Using the System ID, you specify which Stack or LFP Instance an application will use**
- **The following table shows how to specify instance IDs and where they can be applied**
  - The settings are checked from top to bottom as listed in the table

	LE/C Socket API	EZA SOCKET and EZASMI APIs	CSI SOCKET Macro
'LFP$ID' (environment variable)	X		
// SETPARM [SYSTEM] LFP$ID=NN	X	X	
'SYSID' (environment variable)	X		
IDENT.TCPNAME passed to INITAPI call		X	
ID parameter on SOCKET macro			X
// OPTION SYSPARM='NN'	X	X	X
Default '00'	X	X	X

CICS task isolation options

- **LFP isolates CICS tasks from each other**
  - This means that sockets that are allocated by one CICS task, can **not** be used by another CICS task
    - except the socket is passed to the other CICS task via GIVESOCKET/TAKESOCKET calls
  - When a CICS task ends, all sockets allocated by this task will be closed (terminated) automatically
    - Except it has been given to another task prior to task termination

- **Some programs rely on passing sockets from one CICS task to another without the use of GIVESOCKET/TAKESOCKET**
  - For example, DB2 (client or server) application requestor, requires socket sharing, if running under CICS
- **To allow such programs to work with LFP, you need to specify the following JCL statement for the program:**
  - // SETPARM [SYSTEM] LFP$CIC=SHARE
  - This setting applies to the LE/C socket interface as well as the EZA interfaces

- **If socket sharing is active, the applications are responsible to close sockets that are no longer needed**
  - No automatic cleanup will be performed at end of CICS task
  - If the applications miss to do proper cleanup, dead sockets may be left over

CSI SOCKET macro considerations

- **For the CSI SOCKET macro, the Linux Fast Path only supports the following connection types:**
  - TCP
  - UDP
  - CONTROL
  - Other connection types (such as CLIENT, TELNET, FTP, RAW, and so on) are not supported and will be rejected if used with the Linux Fast Path.
- **For CONTROL type connections, the only commands supported are:**
  - GETHOSTBYNAME
  - GETHOSTBYADDR
  - GETHOSTNAME
  - GETHOSTID
  - For details, refer to the individual macro descriptions in the "TCP/IP for VSE V1R5F Programmers Guide" manual.
- **For CONTROL type connections, these commands (from Barnard Software, Incorporated) are also supported:**
  - NTOP
  - PTON
  - GETVENDORINFO
  - For details, refer to the “IPv6/VSE Programming Guide” manual

z/VSE Fast Path to Linux on System z (LFP)

§ Most existing applications run unchanged with Linux Fast Path

- Provided they use one of the supported Socket API (LE/C, EZA or ASM SOCKET)
  - And they do not use any CSI or BSI specific interface, features or functions
  - Since z/VSE V5.1: LFP supports IPv6 NEW!

§ IBM Applications supporting Linux Fast Path

- VSE Connector Server
- CICS Web Support
- VSE Web Services (SOAP) support (client and server)
- CICS Listener
- DB2/VSE Server and Client
- WebSphere MQ Server and Client
- VSAM Redirector
- VSE VTAPE
- VSE LDAP Support
- VSE Script Client
- POWER PNET
- TCP/IP-TOOLS included in IPv6/VSE product (e.g. FTP Server/Client)

§ Customer applications should run unchanged:

- Provided they use one of the supported Socket API (LE/C, EZA or ASM SOCKET)

Performance measurements using Linux Fast Path

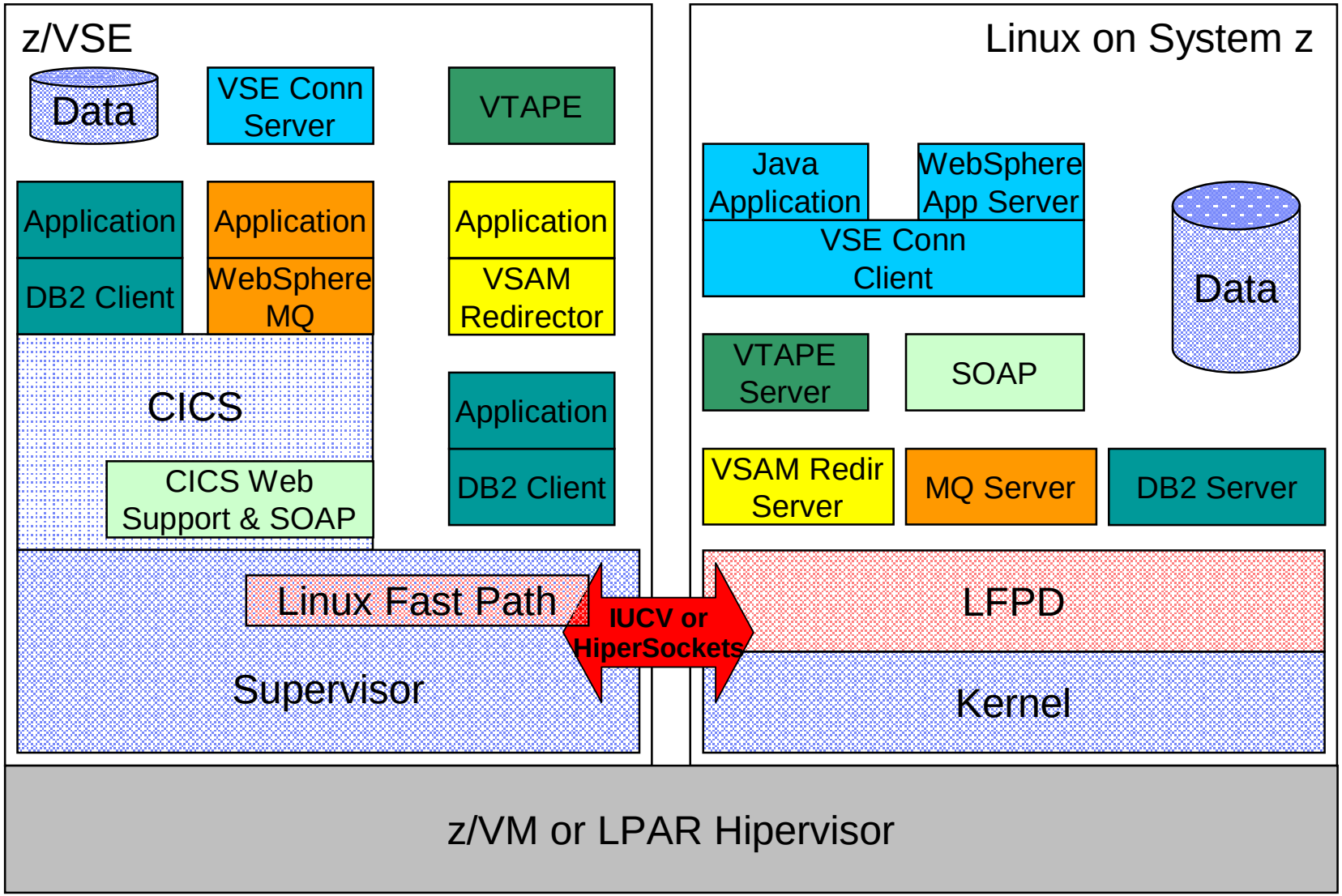
Comparison TCP/IP for VSE versus Linux Fast Path (z/VM Environment):

Workload	TCP/IP for VSE	Linux Fast Path (LFP)	Difference
FTP (BSI FTP server)			
§VSE **à** Linux (1GB) (NULL file, no I/O)	19 MB/sec 29% CPU (5% App + 24% TCPIP)	72 MB/sec 20% CPU (App)	3.7 times faster 9% less CPU
§Linux **à** VSE (1GB) (NULL file, no I/O)	21 MB/sec 55% CPU (11% App + 44% TCPIP)	70 MB/sec 20% CPU (App)	3.3 times faster 35% less CPU
Socket Application (running 3 times)			
§VSE **à** Linux (100MB) §Linux **à** VSE (100MB)	4.6 MB/sec (*3 = 13.8 MB/sec) 9.7 MB/sec (*3 = 29.1 MB/sec) 26% CPU (3*1% App + 23% TCP/IP)	14.6 MB/sec (*3 = 43.8 MB/sec) 16.2 MB/sec (*3 = 48.6 MB/sec) 9 % CPU (3*3% App)	3.2 times faster 1,7 times faster 17% less CPU

Environment: IBM System z10 EC (2097-722). TCP/IP connection via shared OSA adapter.

à Significant benefits in transfer rate as well as CPU usage
à Reduced Sub Capacity Cost

z/VSE Applications communicating with Applications on Linux



Questions ?

THANK YOU