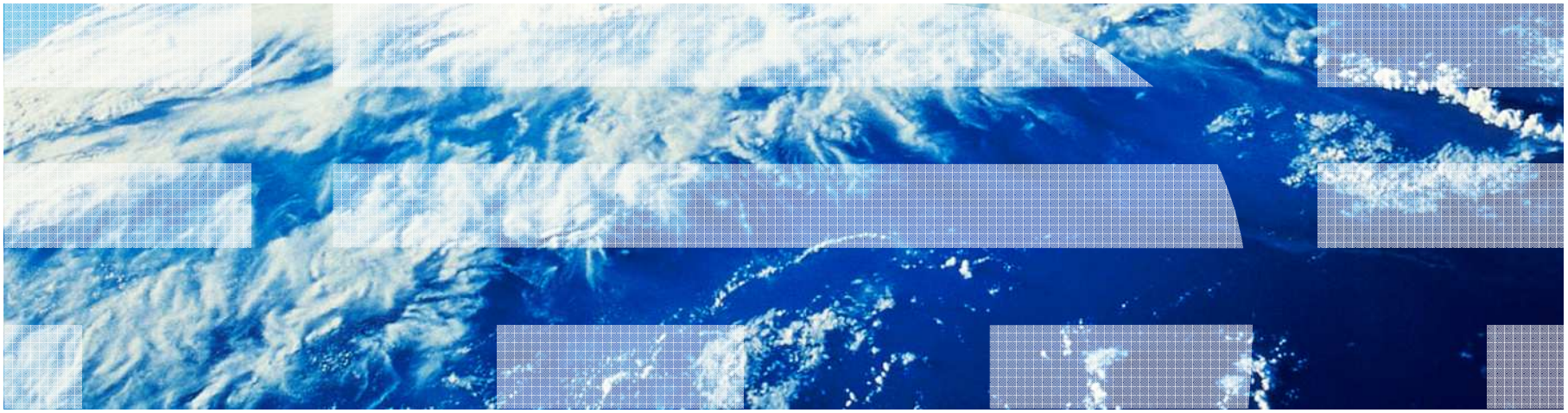
G08 – Aktuelles zu Netzwerkoptionen mit z/VSE, z/VM und Linux

Ingo Franzki & Dr. Manfred Gnirss



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.

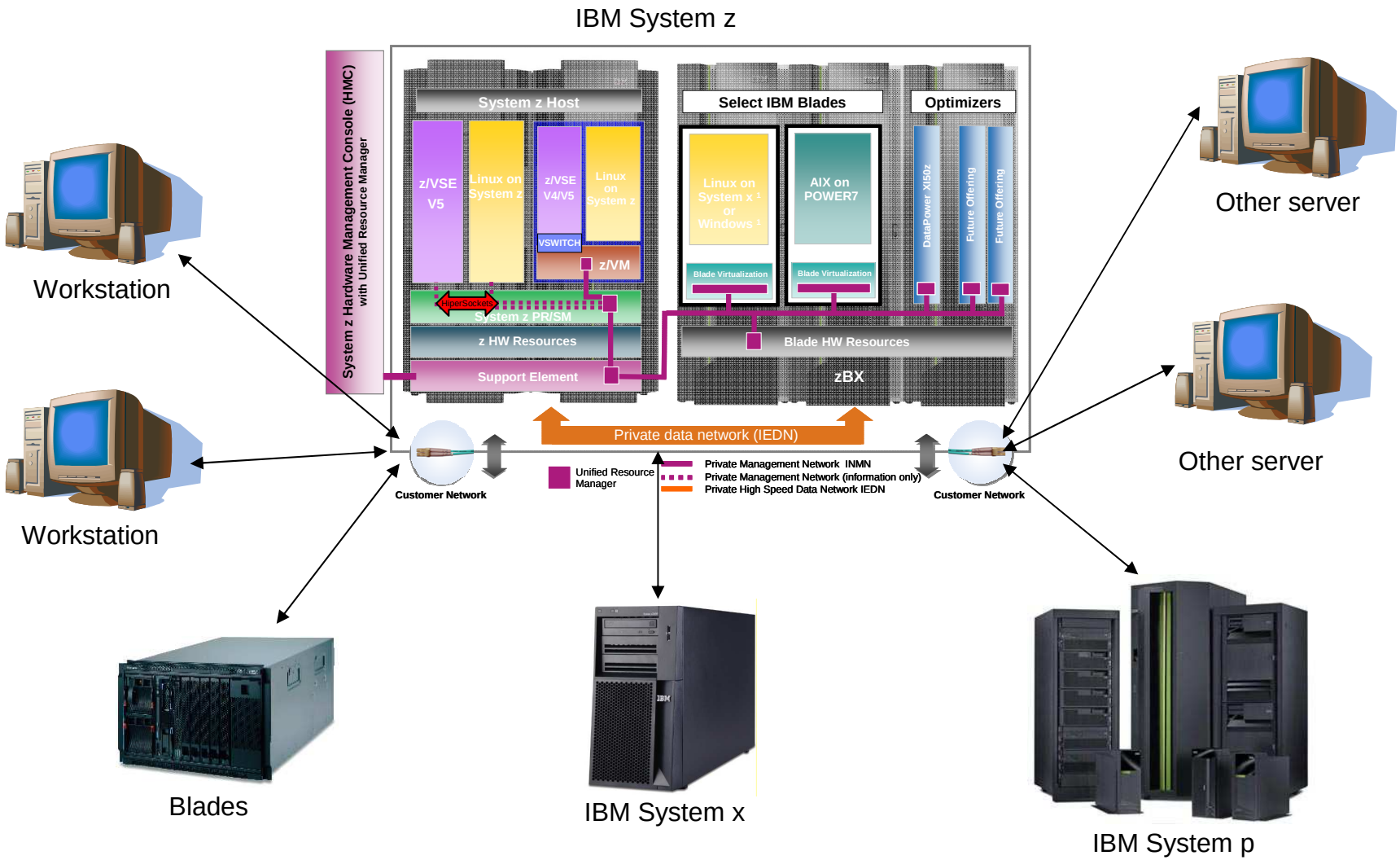
Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both 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.



Networking with z/VSE, z/VM and Linux - Overview



z/VSE TCP/IP Products

- **IPv6/VSE V1.1** (licensed from Barnard Software, Inc)
  - IPv6/VSE provides:
    - An **IPv6 TCP/IP stack**
    - IPv6 application programming interfaces (APIs)
    - IPv6-enabled applications
  - The IPv6 TCP/IP stack of IPv6/VSE can be run concurrently with an IPv4 TCP/IP stack within one z/VSE system
  - The IPv6/VSE product also includes
    - A **full-function IPv4 TCP/IP stack**
    - IPv4 application programming interfaces
    - IPv4 applications.
  - The IPv4 TCP/IP stack does not require the IPv6 TCP/IP stack to be active.
  - With **z/VSE V5.1** IPv6/VSE became a **base product**. With z/VSE V4.3 it is an optional product
  - Supports Layer 2 and 3 mode (z/VSE V5.1)
  - Supports Virtual LAN (VLAN) (z/VSE V5.1)

- **TCP/IP for VSE/ESA V1.5** (licensed from CSI International)
  - Supports IPv4 only
  - Layer 3 mode only

- **Fast Path to Linux on System z** (part of z/VSE V4.3 or later)

OSA Express 4s, OSA Express 3, OSA Express 2

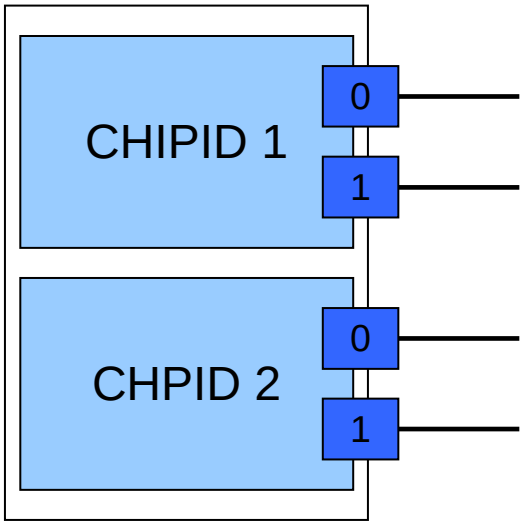
- **CHPID types**
  - **OSC** OSA-ICC (for emulation of TN3270E and non-SNA DFT 3270)
  - **OSD** Queue Direct Input/Output (QDIO) architecture
  - **OSE** non-QDIO Mode (OSA-2, for SNA/APPN connections)
  - **OSN** OSA-Express for NCP: Appears to z/VSE as a device-supporting channel data link control (CDLC) protocol.
  - **OSX** OSA-Express for zBX. Provides connectivity and access control to the Intra-Ensemble Data Network (IEDN) from z196 and z114 to Unified Resource Manager functions

- **For an OSA Express adapter in QDIO mode, you need 3 devices**
  - A read device
  - A write device
  - A datapath device

- **Add the devices in the IPL procedure as device type OSAX:**
  - `ADD cuu1-cuu3,OSAX`

- **In TCP/IP for VSE define a LINK:**
  -
```
DEFINE LINK,ID=...,TYPE=OSAX,
     DEV=cuu1 (or DEV=(cuu1,cuu2)),
     DATAPATH=cuu3,
     IPADDR=addr,
     ...
```
- **In IPv6/VSE define a DEVICE:**
  - `DEVICE device_name OSAX cuu1 portname cuu3`

OSA Express Multi-Port support



- **OSA Express 3 or later provides 2 ports per CHPID for selected features**
  - Default is port 0
  - To use port 1, you must specify this at the DEFINE LINK or DEVICE/LINK statement:
    - **TCP/IP for VSE:**

```
DEFINE LINK,ID=...,TYPE=OSAX,
       DEV=cuu1 (or DEV=(cuu1,cuu2)),
       DATAPATH=cuu3,
       OSAPORT=1,
       ...
```

    - **IPv6/VSE:**

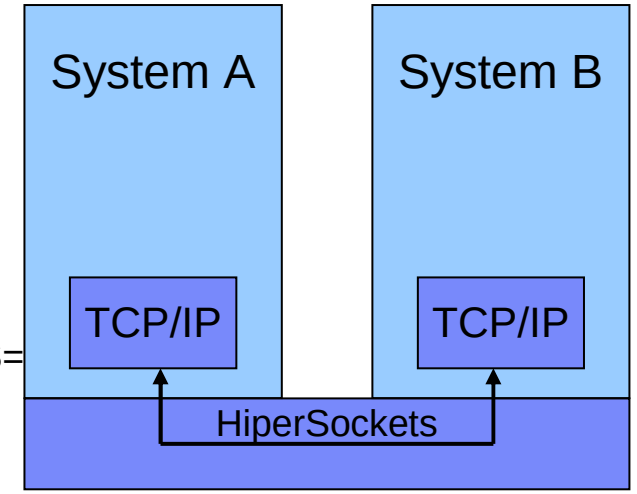
```
DEVICE device_name OSAX cuu1 portname cuu3
LINK device_name adapter_no IPv6_addr netmask mtu
```

  - For CHIPID type OSE (non-QDIO mode) you must use OSA/SF to select the OSA port

HiperSockets

- **“Network within the box” functionality**
  - allows high speed any-to-any connectivity among operating systems
  - without requiring any physical cabling

- **CHPID type IQD**
  - Uses the QDIO (Queue Direct I/O) architecture
  - For an HiperSockets adapter, you need 3 devices
    - A read device
    - A write device
    - A datapath device
  - Add the devices in the IPL procedure as device type OSAX with mode 01:
    - **`ADD cuu1-cuu3,OSAX,01`**
  - Frame size is defined via CHPARM parameter (formerly OS=
    - CHPARM=00 (default): 16K (MTU=8K)
    - CHPARM=40 24K (MTU=16K)
    - CHPARM=80 40K (MTU=32K)
    - CHPARM=C0 64K (MTU=56K)



Layer 2 vs. Layer 3 Mode

- **Layer 2:**
  - TCP/IP stack passes a **frame** to the network card
  - Addressing uses **MAC addresses**
  - TCP/IP stack must perform ARP to translate IP to MAC
- **Layer 3:**
  - TCP/IP Stack passes an (IP) **packet** or **datagram** to the network card
  - Addressing uses IP addresses (IPv4 or IPv6)
  - The network card performs ARP to translate IPv4 to MAC

OSI Model:

Data	**7. Application Layer**	Application
	6. Presentation Layer	representation encryption
	5. Session Layer	Inter host comm.
Segment	**4. Transport Layer**	Flow control
Packet/ Datagram	**3. Network Layer**	Logical addressing
Frame	**2. Data Link Layer**	Physical addressing
Bit	**1. Physical Layer**	Media

Layer 2 vs. Layer 3 Mode (continued)

- **Layer 2:**
  - Supported by **IPv6/VSE** product (BSI) with **IPv6** OSA Express adapter (OSD, OSX) only, no HiperSockets

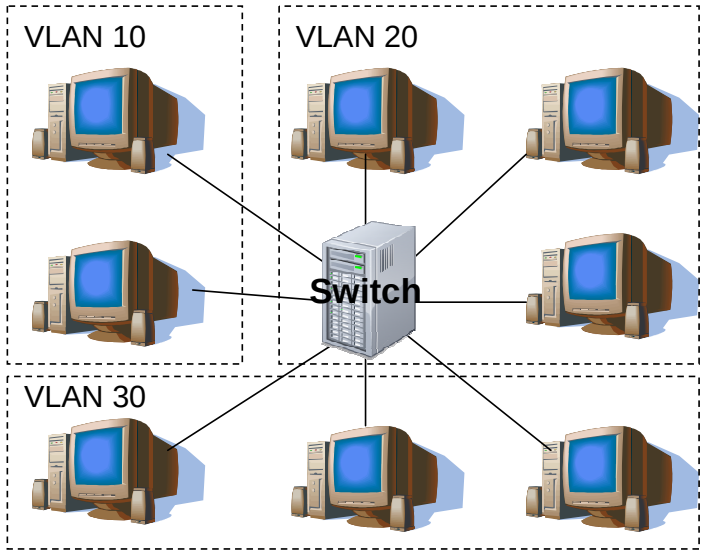
- **Layer 3:**
  - Supported by **IPv6/VSE product** (BSI) with **IPv4 and IPv6**
  - Supported by **TCP/IP for VSE** product (CSI) with **IPv4**

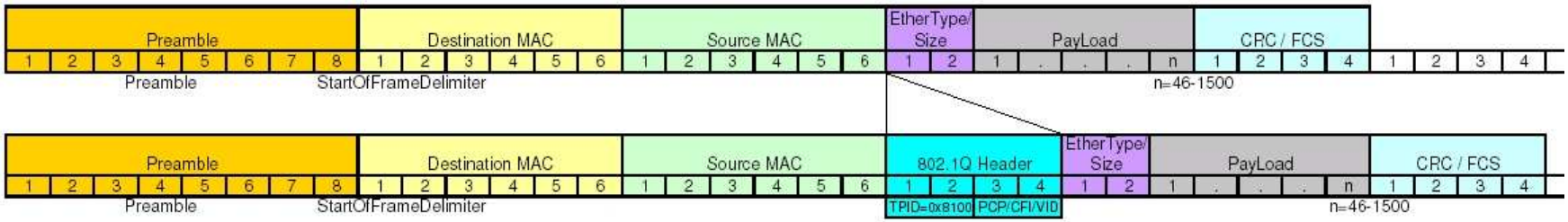
- **VSWITCH:**
  - z/VM allows to define VSWITCH in Layer 2 or layer 3 mode
  - z/VSE V4.2 and 4.3:
    - Supports Layer 3 VSWITCH (IPv4 only)
  - z/VSE V5.1:
    - Supports Layer 2 VSWITCH (IPv4 and IPv6)
    - Supports Layer 3 VSWITCH (IPv4 only)

→ Be carefully when connecting z/VSE systems to already existing VSWITCHes

Virtual LAN (VLAN)

- **VLAN allows a physical network to be divided administratively into separate logical network**
- **These logical networks operate as if they are physically independent of each other**
- **A VLAN tag is inserted into the Link Layer Header**
  - **3 bit priority:** an be used to prioritize different classes of traffic (voice, video, data)
  - **12 bit VLAN ID:** specifies the VLAN to which the frame belongs

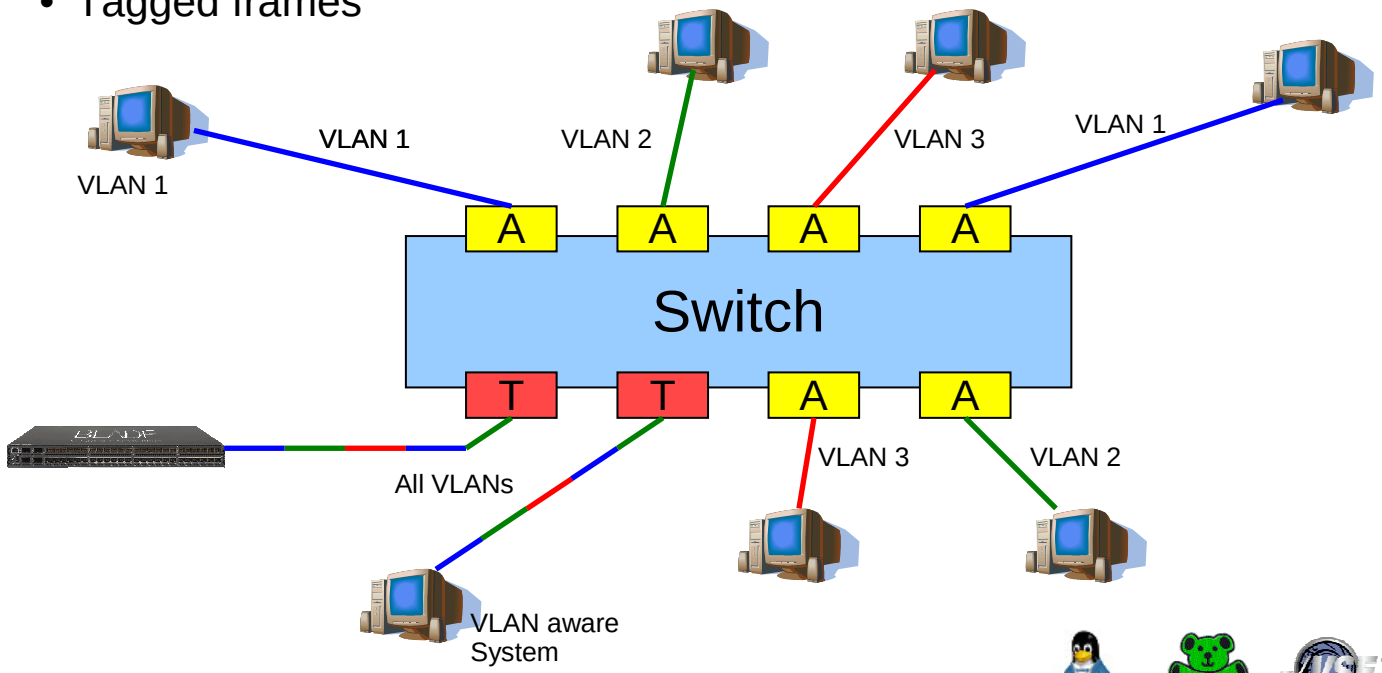




Source: Wikipedia: http://en.wikipedia.org/wiki/File:TCPIP_802.1Q.jpg

Virtual LAN (VLAN) – Trunc Port / Access Port

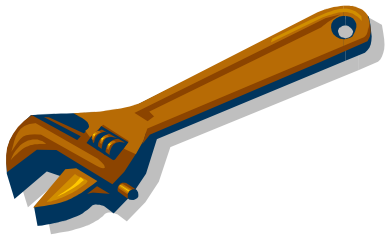
- **Switches have different types of ports**
  - **Access Port**
    - Not VLAN-aware
    - Un-tagged frames
  - **Trunc Port**
    - VLAN-aware
    - Tagged frames



Virtual LAN (VLAN) – z/VSE support

- **z/VSE provides VLAN support for OSA Express (CHPID type OSD and OSX) and HiperSockets devices**
  - In a **Layer 3** configuration, VLANs can be **transparently** used by **IPv6/VSE** and **TCP/IP for VSE/ESA**
  - If you wish to configure VLANs for OSA-Express (CHPID type OSD and OSX) devices in a **Layer 2** configuration that carries **IPv6 traffic**, you require the **IPv6/VSE** product

- **You can use one of the following two ways to configure your system to use VLAN:**
  1. **Configure** one or more VLANs **in the TCP/IP stack** of **IPv6/VSE**
     - For details of IPv6/VSE commands, refer to IPv6/VSE Installation Guide
  2. **Generate** and catalog phase **IJBOCONF** containing the **Global VLANs** to be used with your OSAX devices
     - z/VSE provides skeleton SKOSACFG to generate phase IJBOCONF
     - The VLANs contained in IJBOCONF can be **transparently** used for **Layer 3** links by I**Pv6/VSE** and **TCP/IP for VSE/ESA**

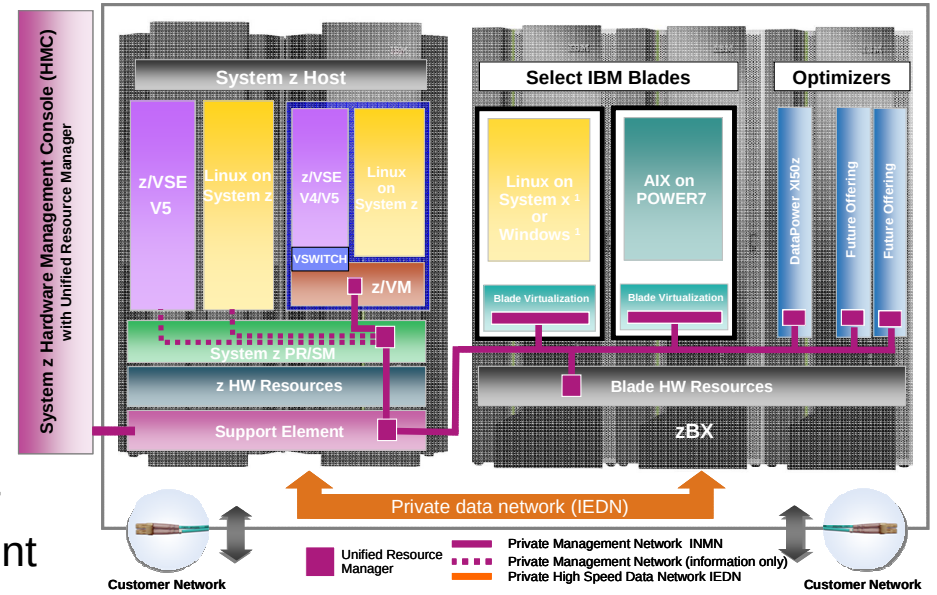


Intra-Ensemble Data Network (IEDN) support

- **OSA-Express for zBX (CHPID type OSX)**
  - Provides connectivity and access control to the Intra-Ensemble Data Network (IEDN) from zEnterprise 196 and 114 to Unified Resource Manager functions

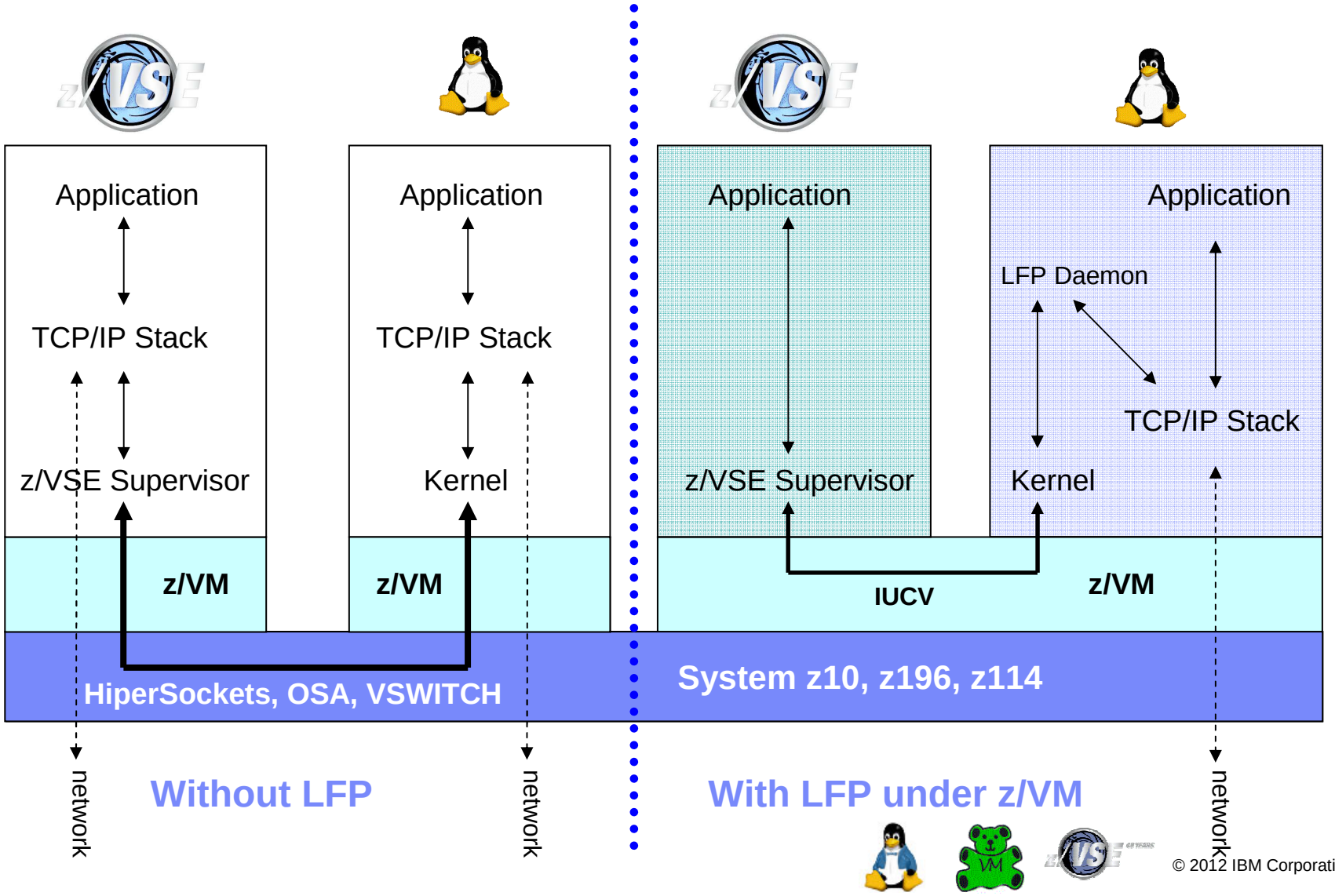
- **An Intra-Ensemble Data Network (IEDN) provides connectivity between:**
  - A zEnterprise CEC (Central Electrical Complex) and System z Blade Center Extensions (zBXs)
  - Two or more zEnterprise CECs

- **z/VSE supports the IEDN network of a zEnterprise 196 or 114**
  - **z/VSE V4.2, V4.3 and V5.1:**
    - z/VM VSWITCH and **OSDSIM** mode in a **z/VM 6.1** guest environment
  - **z/VSE V5.1:**
    - **OSA Express for zBX** devices either in an **LPAR** or **z/VM** guest environment with **dedicated** OSAX devices
    - This requires **VLAN** support



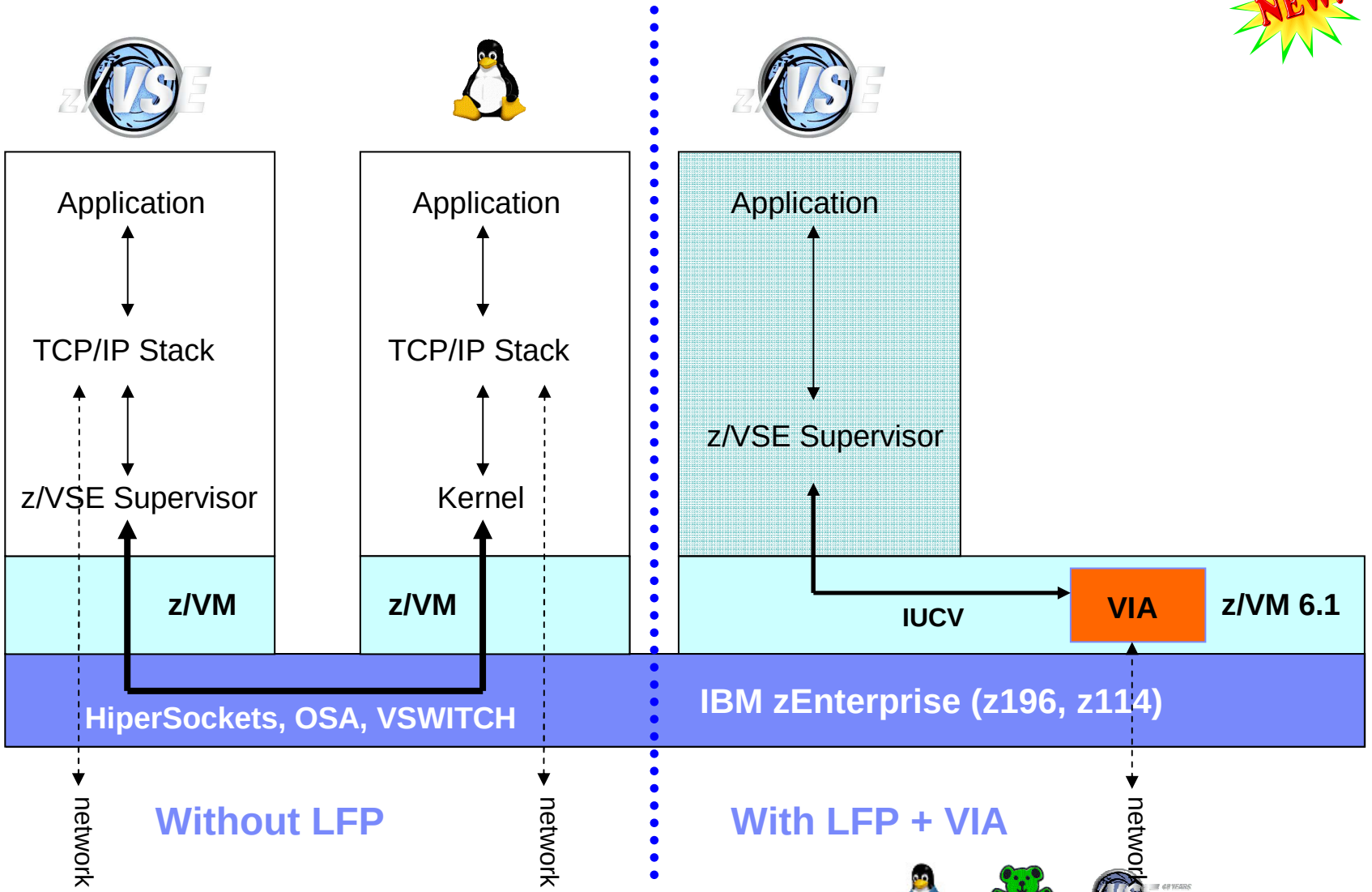
Linux Fast Path in a z/VM environment (z/VSE 4.3 or later)

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

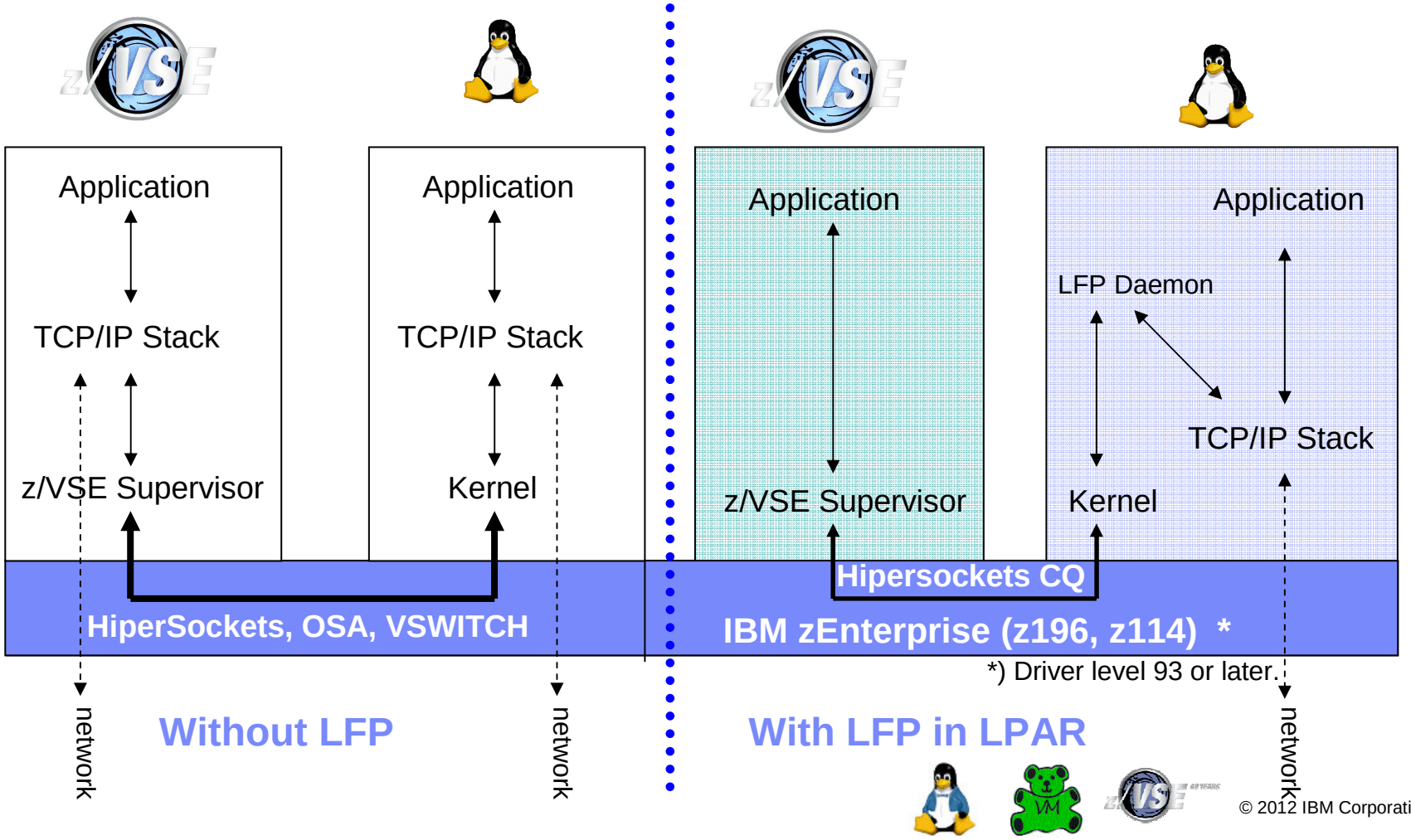


New: Linux Fast Path in an LPAR environment (z/VSE 5.1 + PTFs)

Faster communication between z/VSE and Linux applications

NEW!

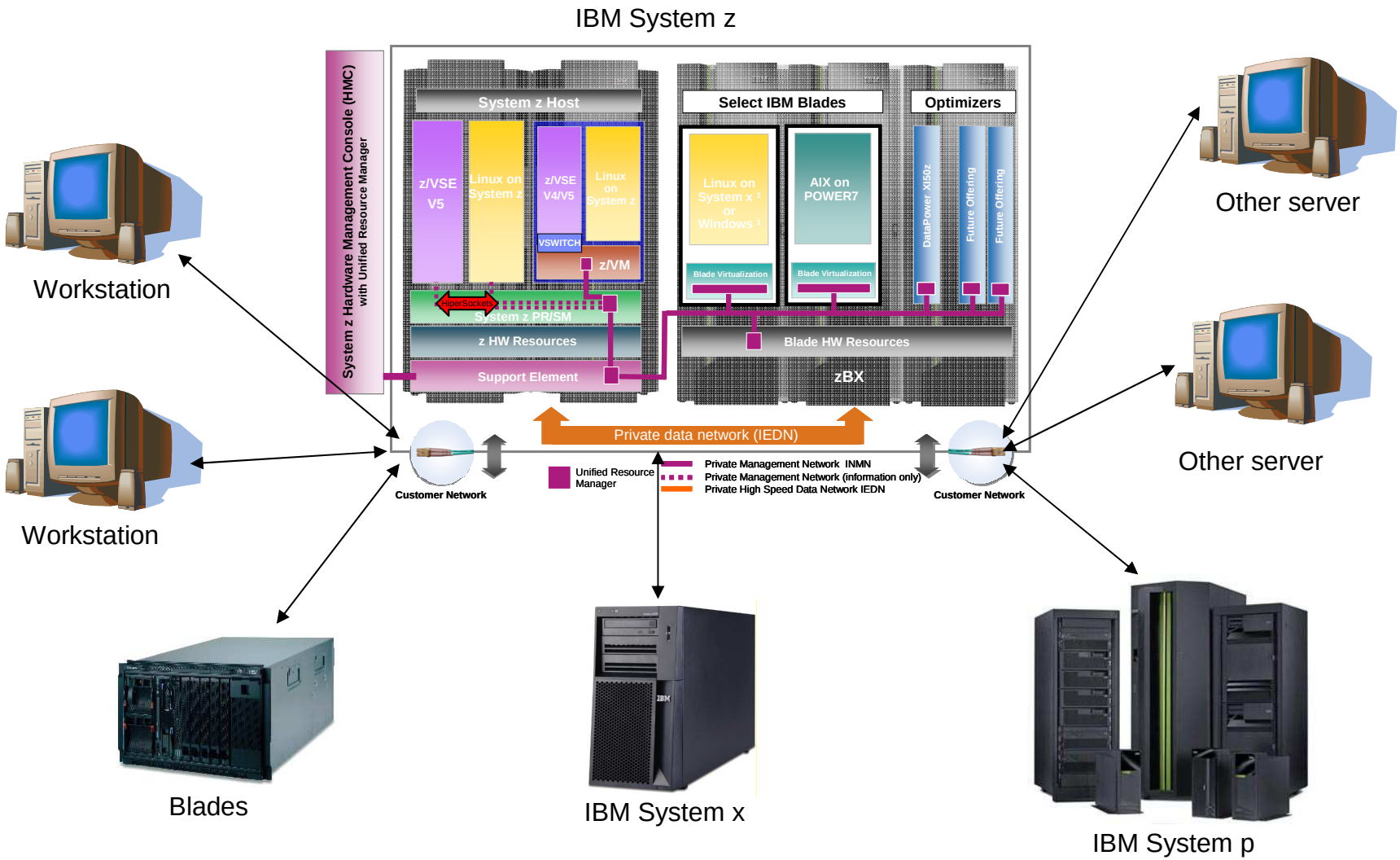
→ Exploits the HiperSockets Completion-Queue support of IBM zEnterprise (z196, z114)



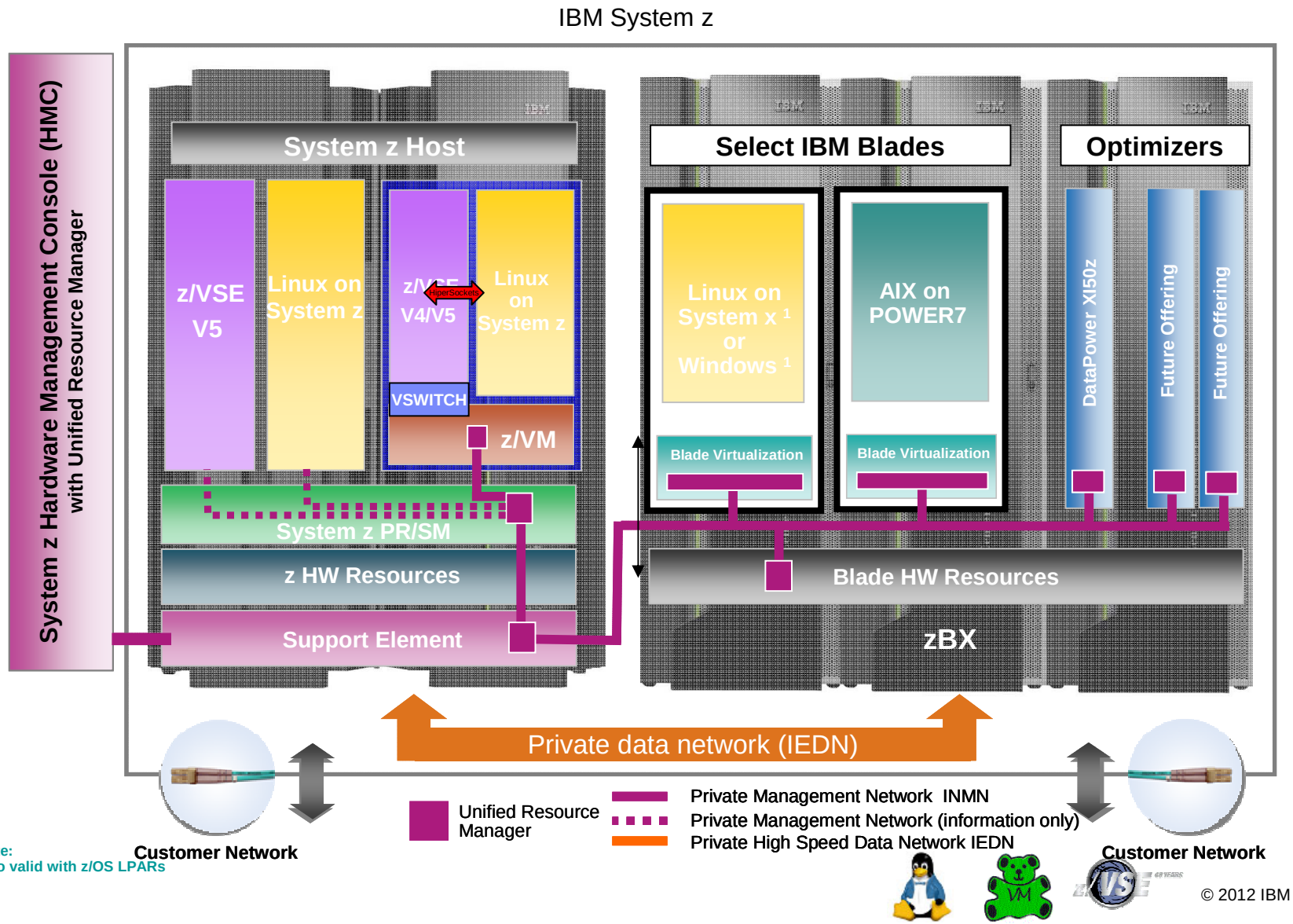
Part 2
Extended Connectivity

- Recent Enhancements
- Connecting the External Network to the IEDN at the TOR

Networking - Overview



Networking – Overview . . .



Note:
also valid with z/OS LPARs

HiperSockets Completion Queues

- Transfer HiperSockets messages asynchronously
- Used whenever traditional synchronous queues are full
- Automatic enablement; no z/VM configuration required
- Helpful when traffic is “bursty”
- Exploitation by CP VSWITCH only; no guest simulation

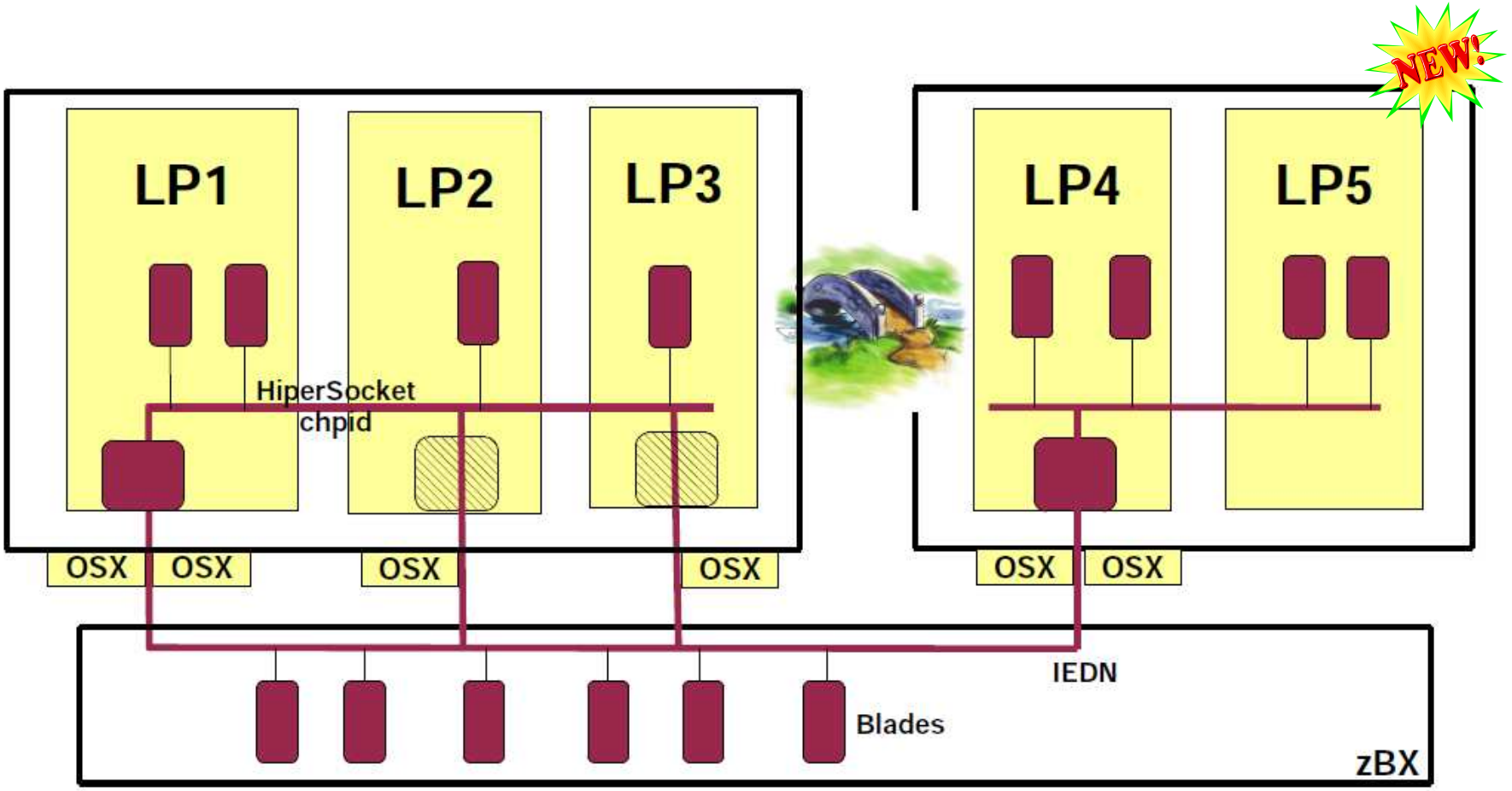
- Requires Driver 93 with current MCLs

- Statement of Direction July 12, 2011:

  IBM plans to support transferring HiperSockets messages asynchronously, in addition to the current synchronous manner on z196 and z114. This could be especially helpful in burst situations. The Completion Queue function is designed to allow HiperSockets to transfer data synchronously if possible and asynchronously if necessary, thus combining ultra-low latency with more tolerance for traffic peaks. HiperSockets Completion Queue is planned to be supported in the z/VM and z/VSE environments in a future deliverable

- Operating System Support
  - z/OS V1.13 (Toleration, no exploitation)
  - Linux on System z distributions
    - Red Hat Enterprise Linux (RHEL) 6.2
    - Novell SUSE Linux Enterprise Server (SLES) 11 SP2
  - z/VSE 5.1 plus PTFs for LFP in LPAR

HiperSocket VSWITCH Integration with zEnterprise IEDN



- z/VM guest only
- Built-in failover and failback
- Special IOCP definition will be required

- Same or different LPAR
- One active bridge per CEC
- PMTU simulation

HiperSocket VSWITCH Integration with zEnterprise IEDN . . .

- Virtual Switch bridge between Ethernet LAN and HiperSockets
  - zEnterprise IEDN (OSX) connections
  - Guests can use simulated OSA or dedicated HiperSockets
  - VLAN aware
  - One HiperSocket chpid only
- Full redundancy
  - Up to 5 bridges per CEC
  - One bridge per LPAR
  - Automatic takeover
  - Optionally designate one “primary”
    - Primary will perform “takeback” when it comes up
  - Each bridge can have more than one OSA uplink

- Requirements:
  - Hardware
    - zEnterprise 196 or 114 system, driver 93, with bundle 22z (or higher) applied.
  - APARs/PTFs
    - VM - APAR VM65042/PTF UM33691
    - TCP/IP - APAR PM46988/PTF UK77220
    - PerfKit - APAR VM65044/PTF VM33693
  - Guest Operating Systems
    - Linux RHEL 5.8 (GA-level)
    - Linux RHEL 6.2 (GA-level)
    - Linux SLES 10 SP4 update (kernel 2.6.16.60-0.95.1)
    - Linux SLES 11 SP2 (GA-level)
    - Note: IBM is working with SUSE to provide an appropriate SLES 11 SP1 kernel update.

Statement of Direction July 12, 2011: Within a zEnterprise environment, it is planned for HiperSockets to be integrated with the intraensemble data network (IEDN), extending the reach of the HiperSockets network outside of the central processor complex (CPC) to the entire ensemble, appearing as a single Layer 2 network. HiperSockets integration with the IEDN is planned to be supported in z/OS V1.13 and z/VM in a future deliverable.

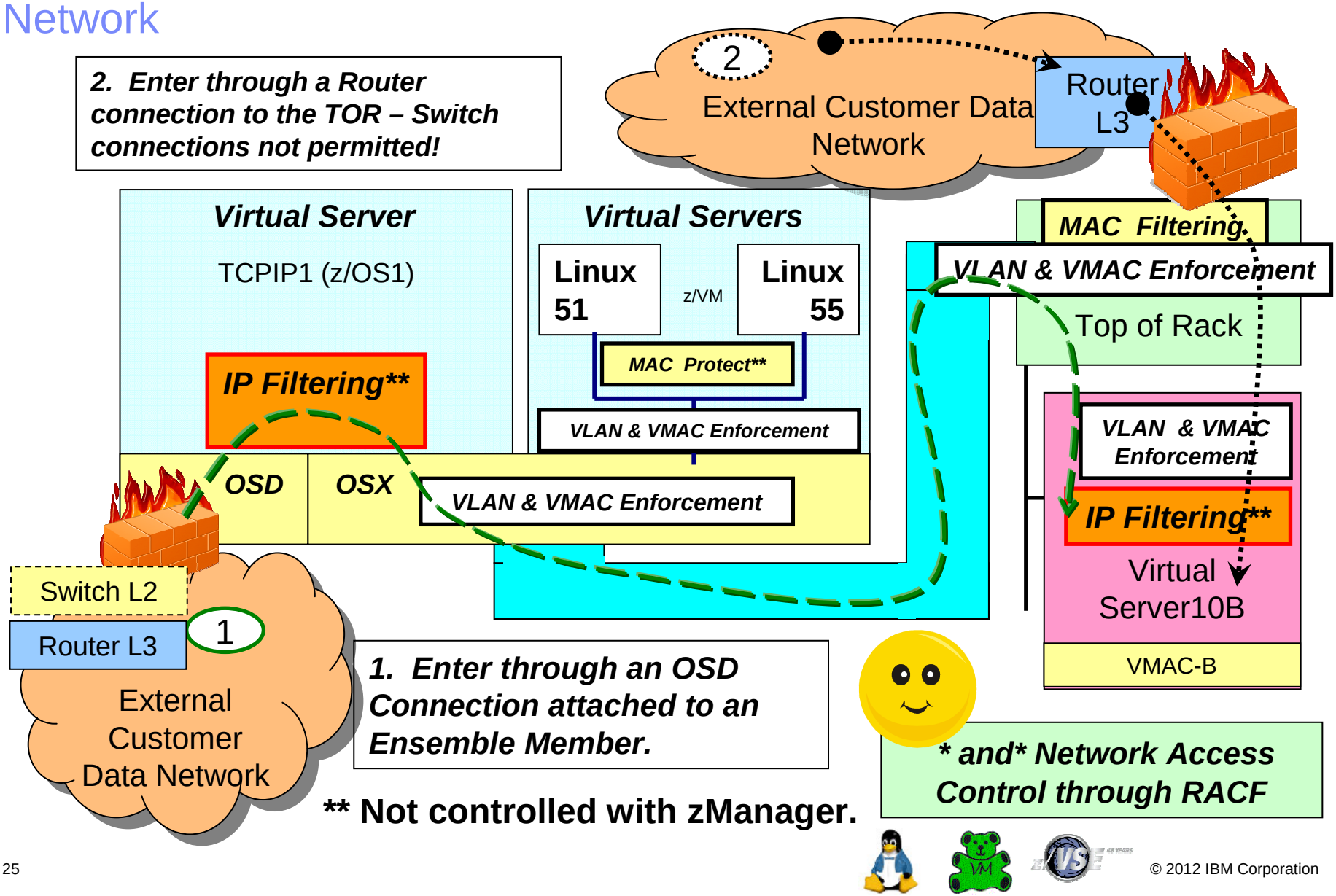
Networking with zEnterprise

- *Connecting the External Network to the IEDN at the TOR: Avoiding the BIG Mistake”*

Abstract

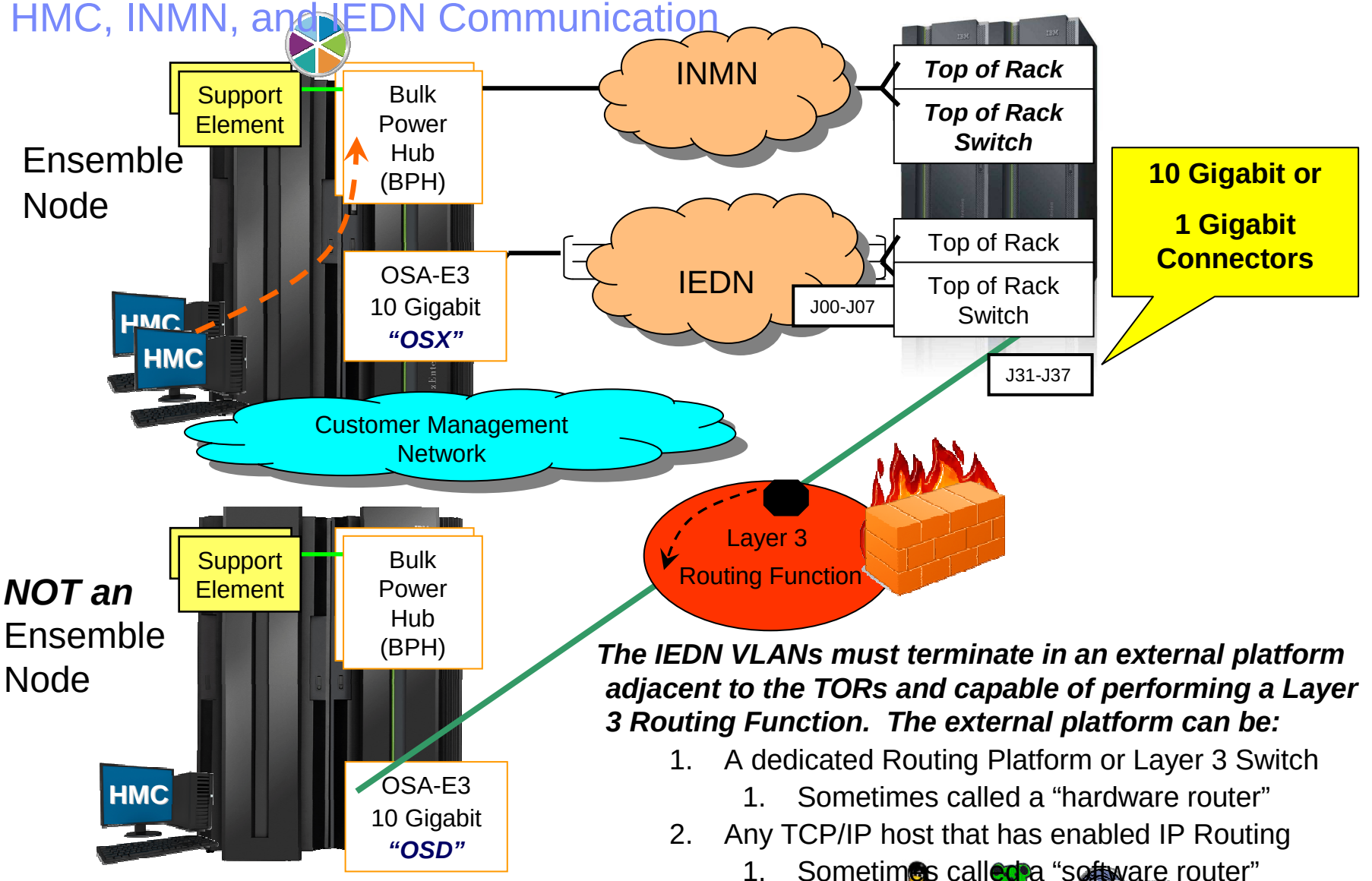
- Connecting the External Customer Network to the Ensemble by attaching to ports in the zBX Top-of-Rack (TOR) switch is simple to do, but it is not the same as connecting any external device to just any Layer 2 switch. The connection must be a ROUTED connection and not a SWITCHED connection. The biggest mistake you can make is to attempt a connection that relies on Layer 2 switching protocols.
- Such attempts are likely to fail because the zBX TOR has been configured to expect Layer 3 routed connectivity and is incompatible with typical Layer 2 switching protocols.
- For further information about networking with the zEnterprise System Ensemble, consult the extensive information in **IBM® zEnterprise™ System Network Virtualization, Management, and Security (Parts 1 and 2 - Overview and Details)** at:
  - **http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS4160**
- **For other assistance,**
  - **Please work with your IBM representative, who may open a TechXpress request to consult with the IBM Advanced Technical Support zEnterprise Communications Server Networking Team:**
    - http://techsales4.austin.ibm.com/tsna/techxpress.nsf/request.html

Connecting the Customer Data Network to the intraensemble Data Network



External Connection to the Top of Rack Switches of the zBX

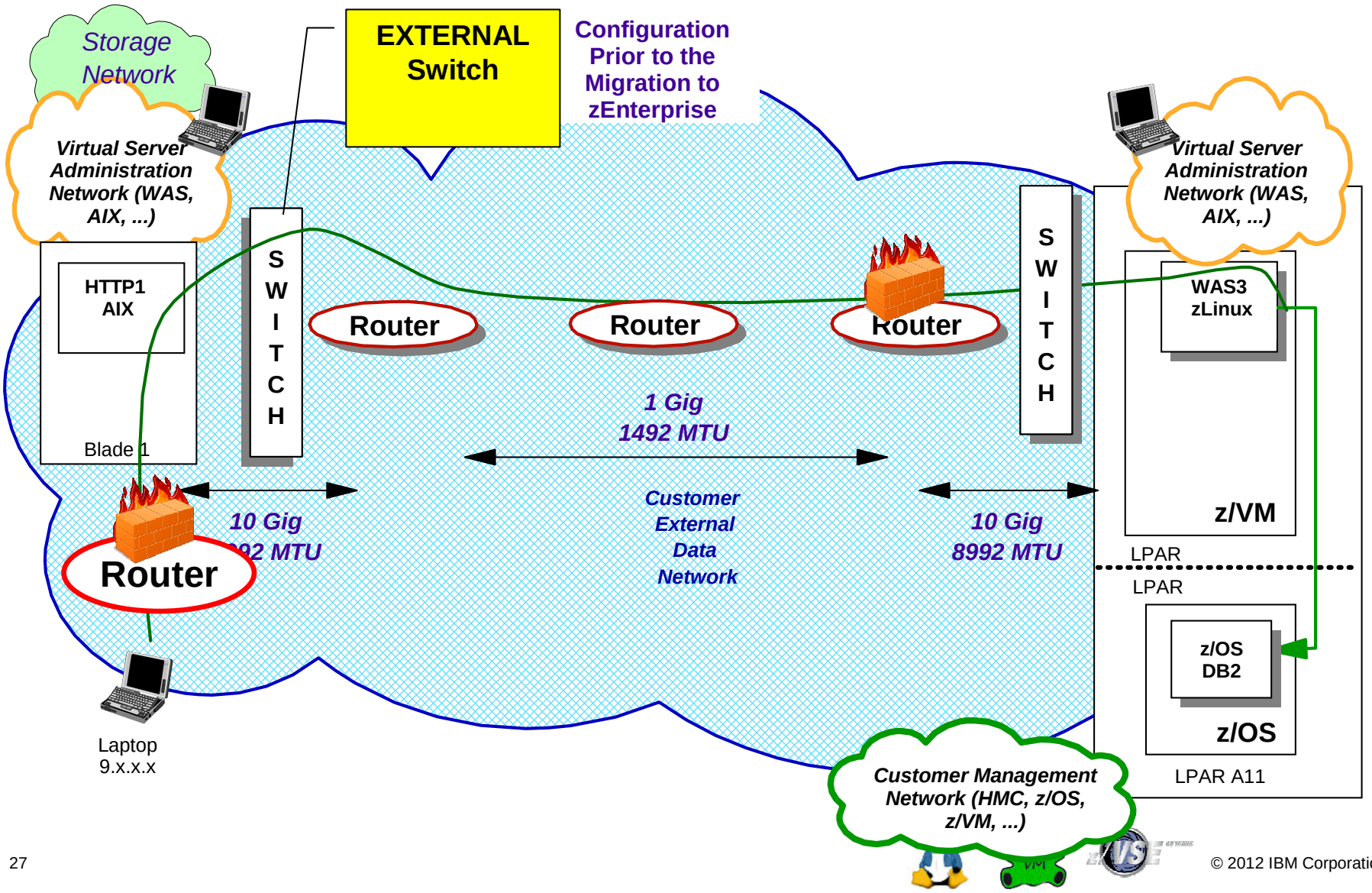
HMC, INMN, and IEDN Communication



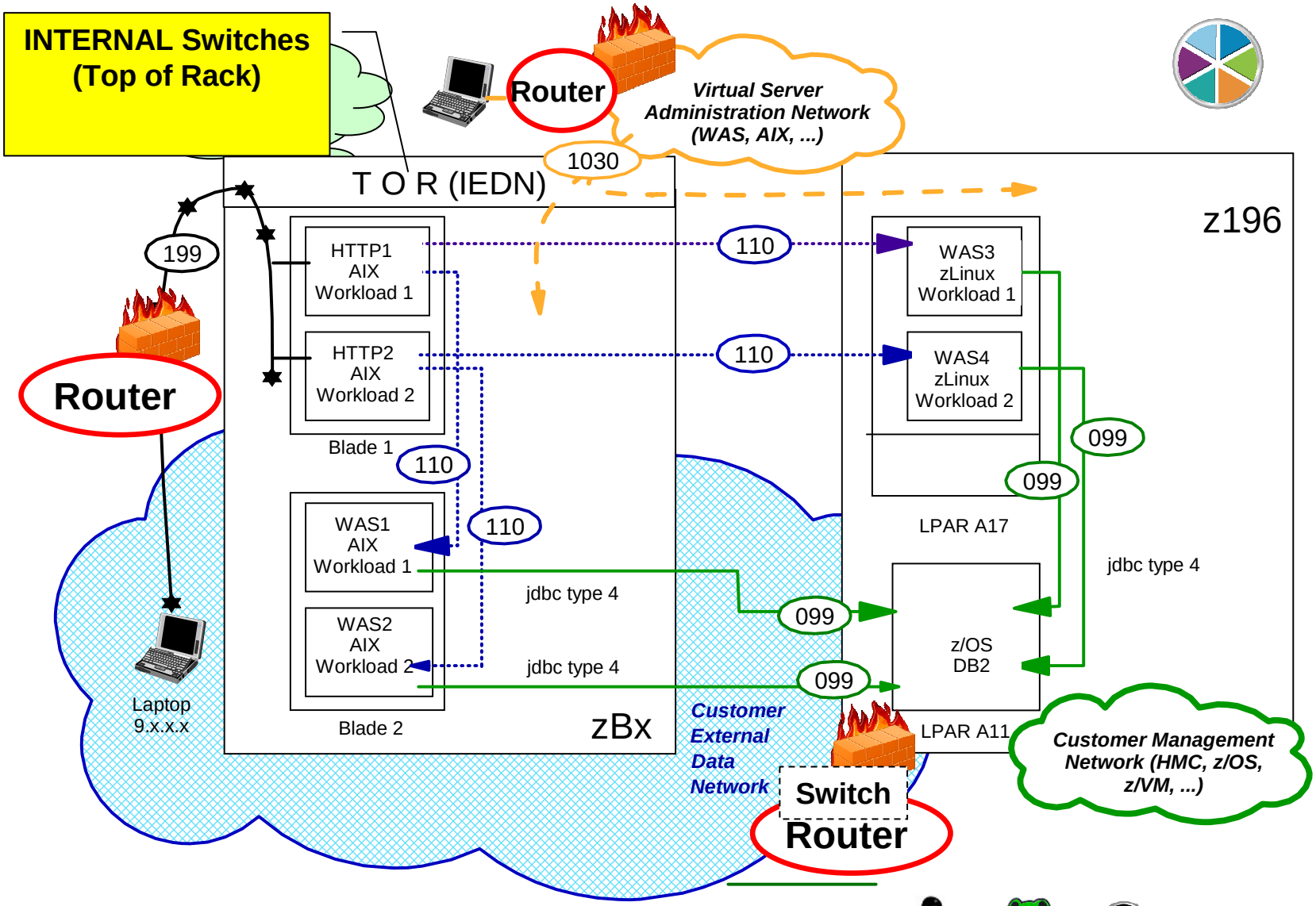
The IEDN VLANs must terminate in an external platform adjacent to the TORs and capable of performing a Layer 3 Routing Function. The external platform can be:

1. A dedicated Routing Platform or Layer 3 Switch
   1. Sometimes called a "hardware router"
2. Any TCP/IP host that has enabled IP Routing
   1. Sometimes called a "software router"

Conventional Network Design (without Ensemble): External Switch

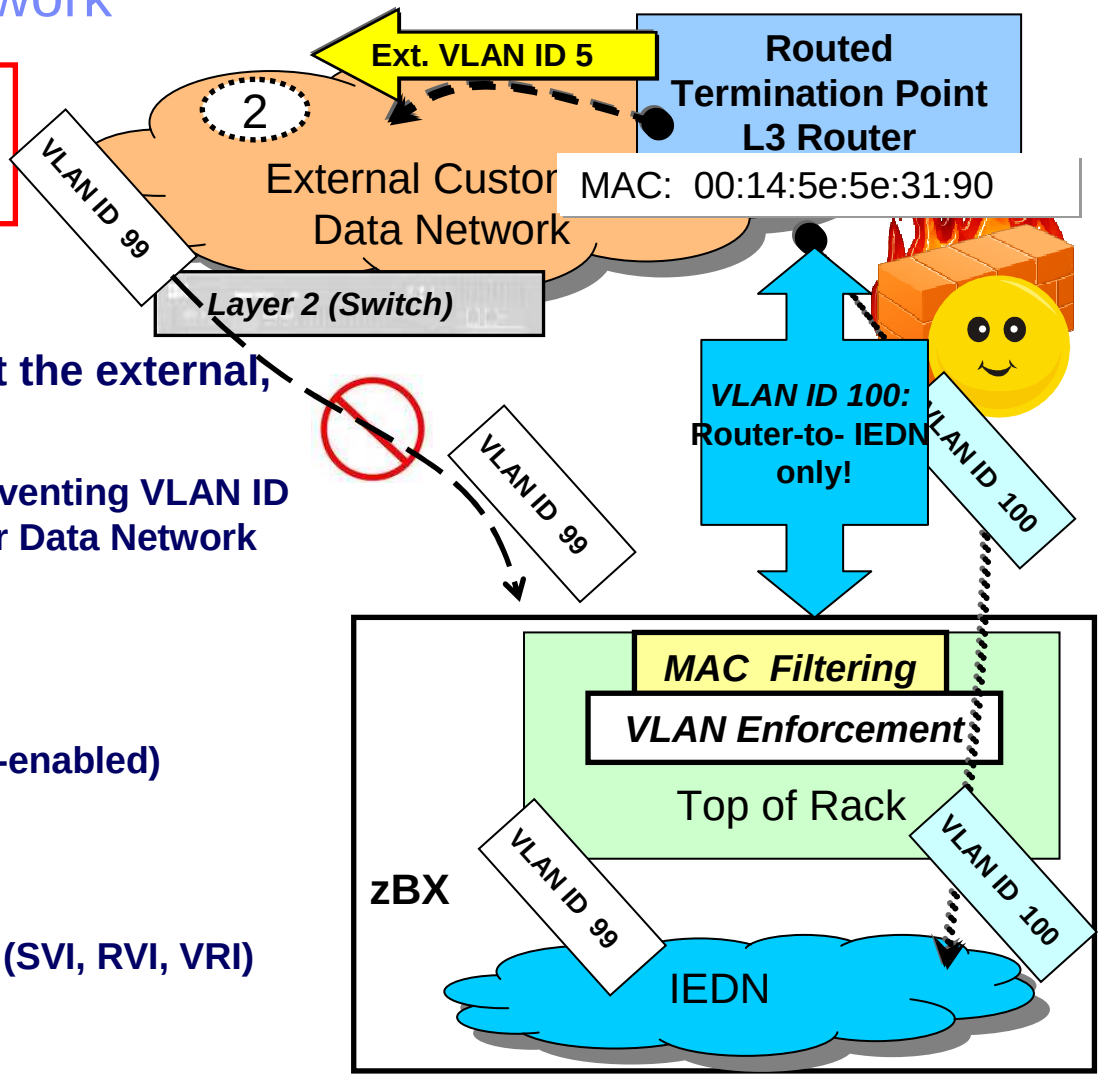


Sample Virtual Network Design (with Ensemble): Top of Rack Switch



A Closer Look at the IEDN TOR and its Secure Connections to the External Customer Data Network

2. Enter through a Router connection to the TOR – Switch connections not permitted!



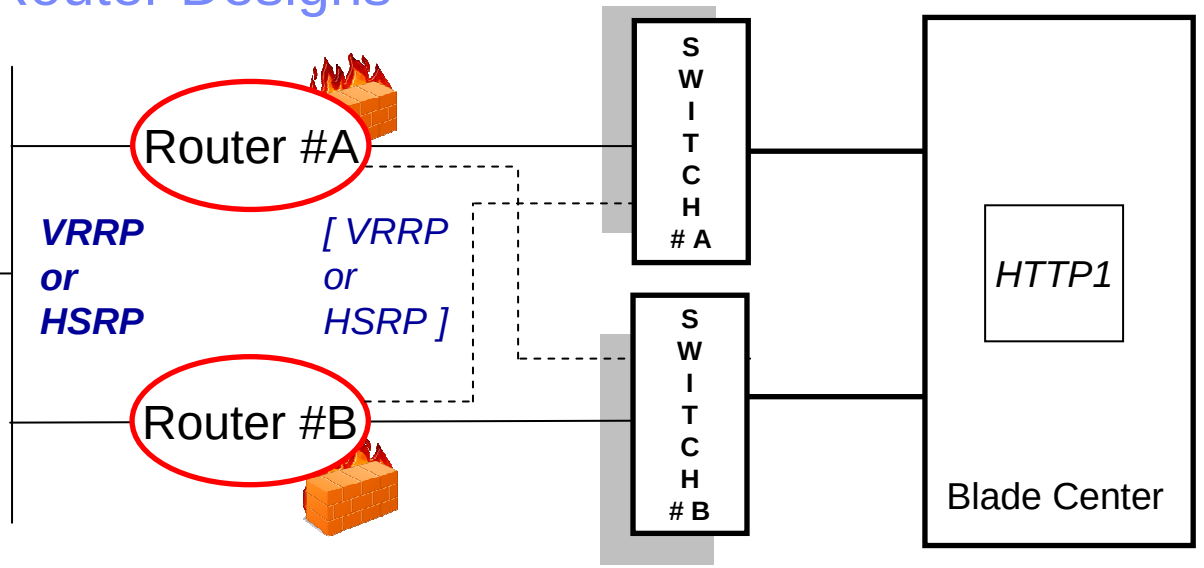
❖**Selected IEDN VLANs terminate at the external, Layer 3 Routing platform.**

- **Integrity of the IEDN is preserved by preventing VLAN ID collisions between the external Customer Data Network and the IEDN!**

- **Routed termination points ONLY**
  - ✓**Dedicated Router Platform**
  - ✓**Operating System Platform (routing-enabled)**
  - ✓**L2/L3 Switch with**
    - ❑**Routed Interface or**
    - ❑**Sub-interface definitions**
    - ❑***With Caution*: Virtual Interfaces (SVI, RVI, VRI)**
- **No external Layer 2 Switch!**
  - ✓**No Layer 2 Messages to TOR**
  - ✓**No STP messages**
  - ✓**No BPDUs, etc.**

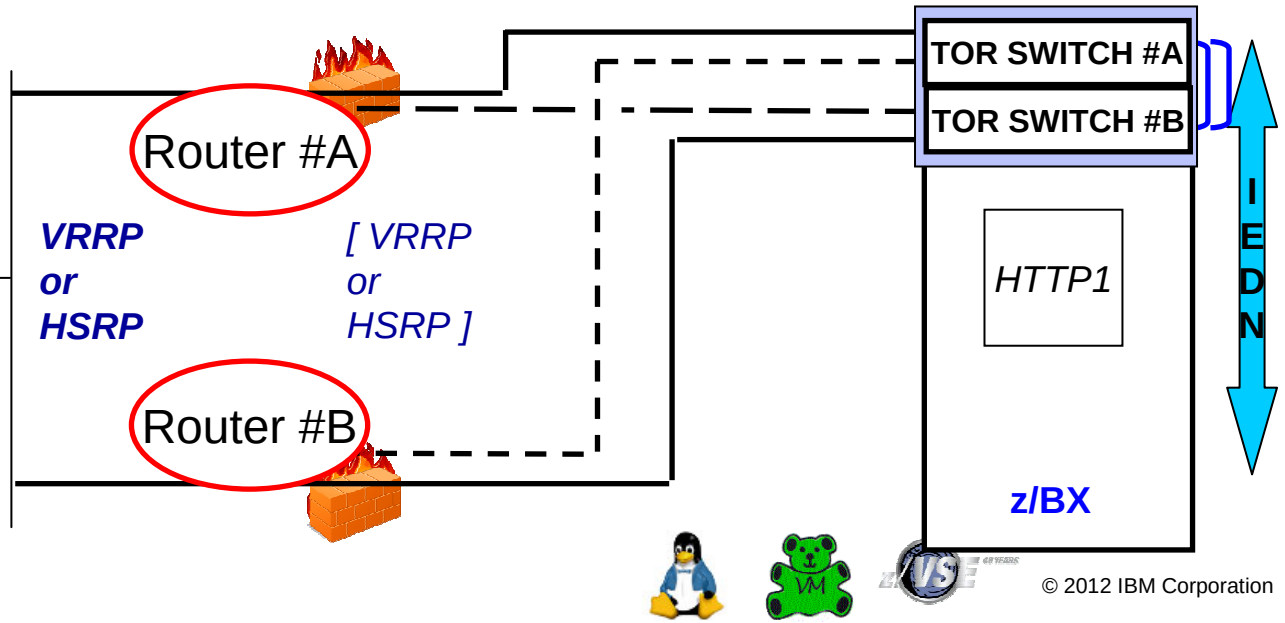
Sample Redundant Router Designs

Sample Conventional Design for Router Attachment



Sample IEDN Design for External Router Attachment to Ensemble:

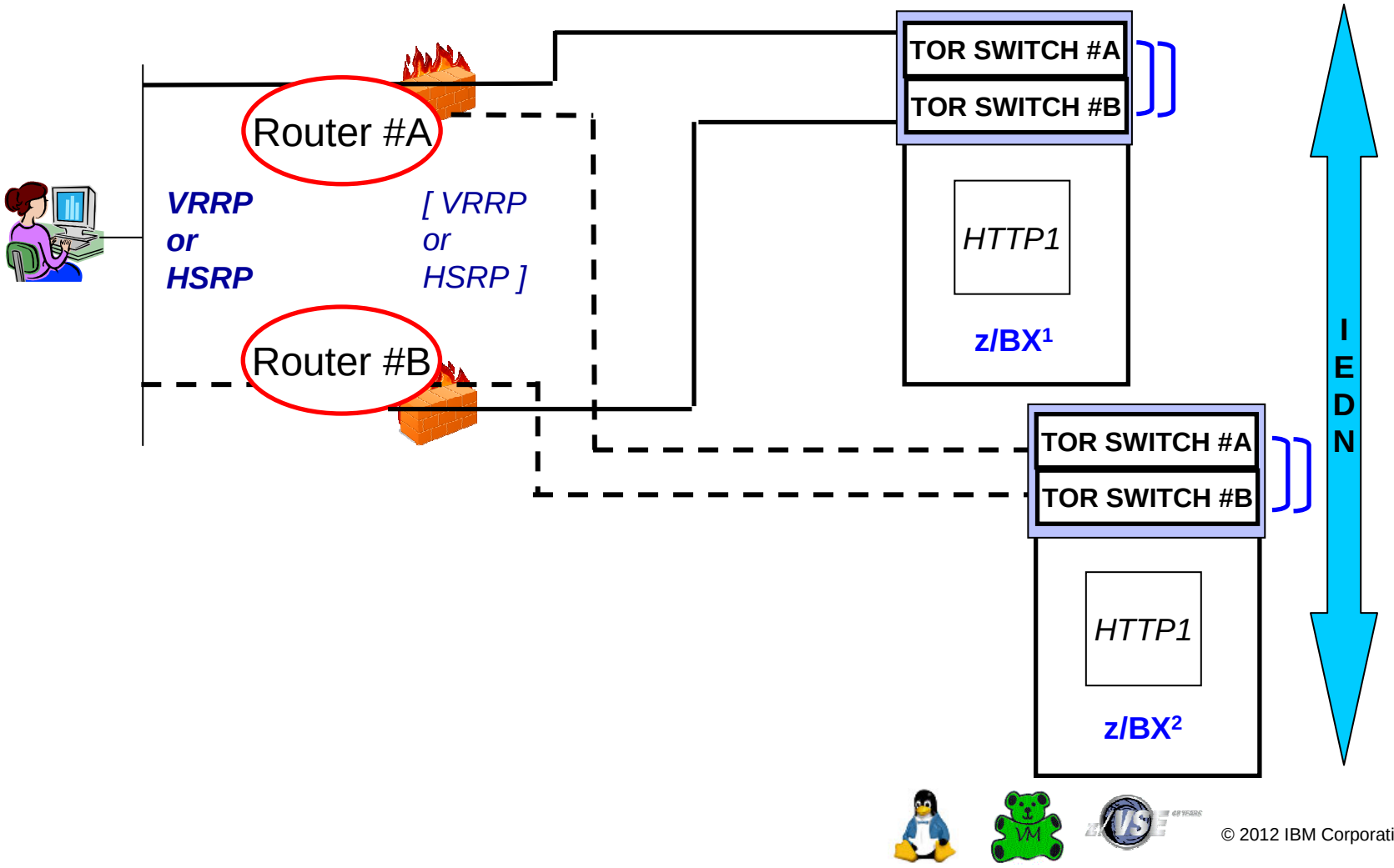
Each Router Connects to Both TORs



Sample Redundant zBX Design for High Availability

Sample IEDN Design for External Router Attachment to Ensemble:

Each Router Connects to TOR A in one zBX and TOR B in a 2nd zBX



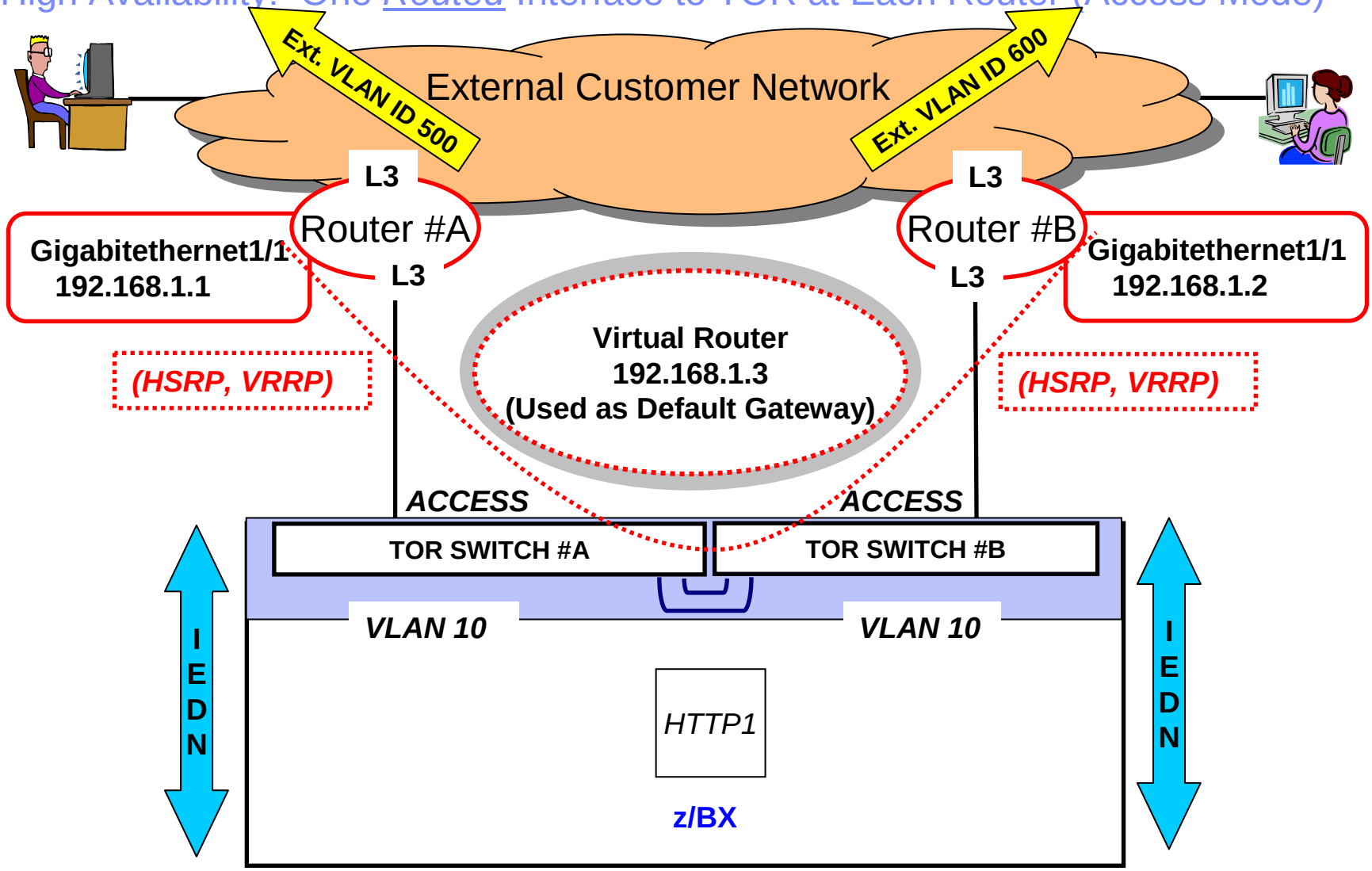
Why We Recommend External Network Layer 3 Connectivity

"Layer 3 Connectivity (Routed Connectivity):

Recommended and Supported"

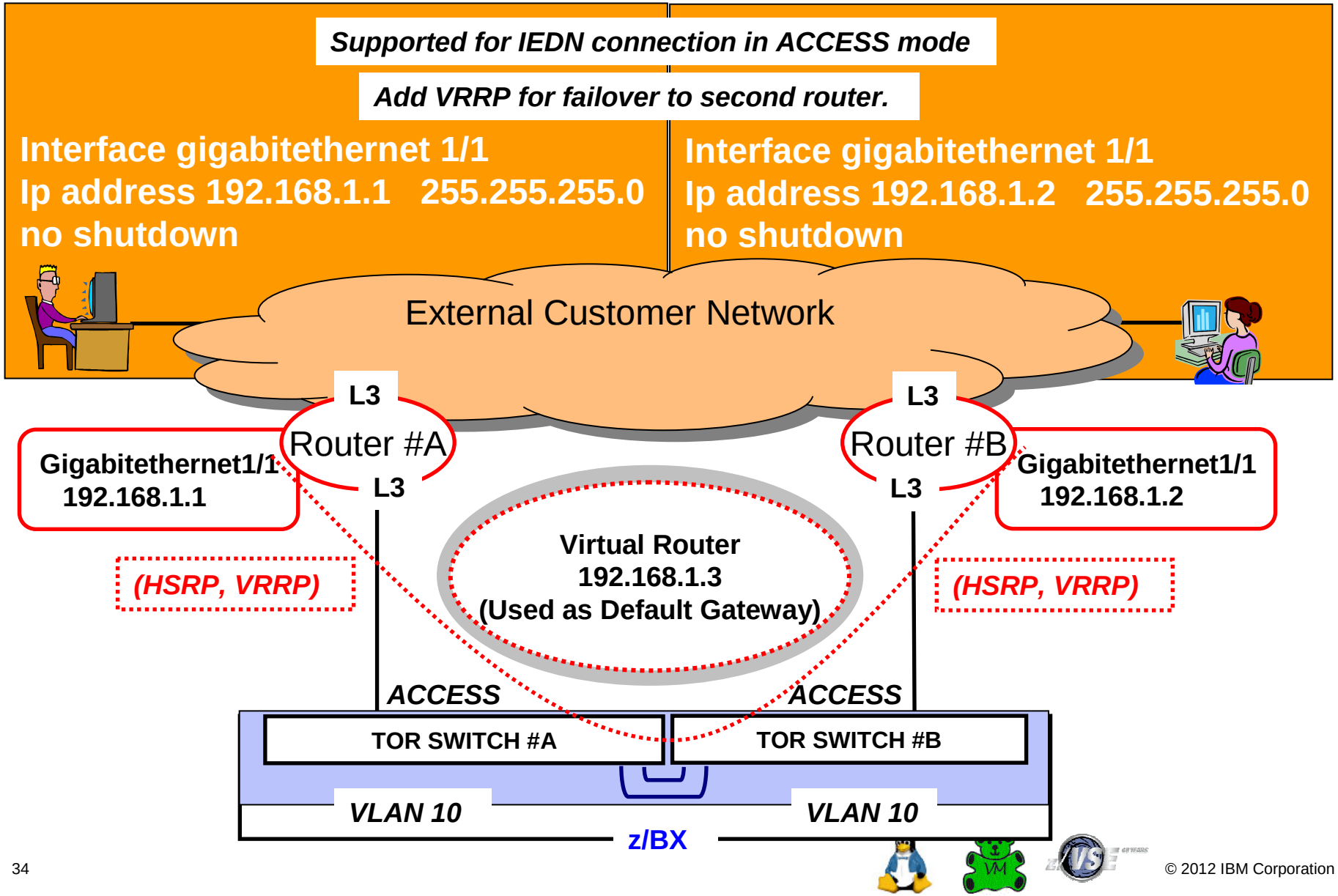
- Intra-Ensemble Data Network is a 'closed' flat Layer 2 network
  - 'Closing' the network is accomplished by decreeing that entry into the network is achieved through Layer 3 IP routing (either an LPAR or external physical router) into the zBX
  - 'Best practice' from a security and administrative perspective is to place a firewall/router prior to entry into the IEDN. This approach provides:
    - Secure isolation/logging/auditing that are typical security requirements when crossing security zones
    - Distinct network administration responsibilities and boundaries (VLANs, VMACs, access controls, etc.)
  - In the closed network environment zManager takes total responsibility for:
    - Network fabric configuration, monitoring and management
    - ensuring no virtual MAC or VLAN conflicts (collisions) can occur (and can not be spoofed)
    - preventing STP packets from passing through the TOR into the IEDN layer 2 LAN segment
    - Identifying, authorizing access and ensuring all virtual servers within the ensemble can successfully communicate with each other
    - Assuring network high availability is provided (eliminating single points of failures)
    - Single point of RAS (network diagnostic responsibilities)

High Availability: One *Routed* Interface to TOR at Each Router (Access Mode)

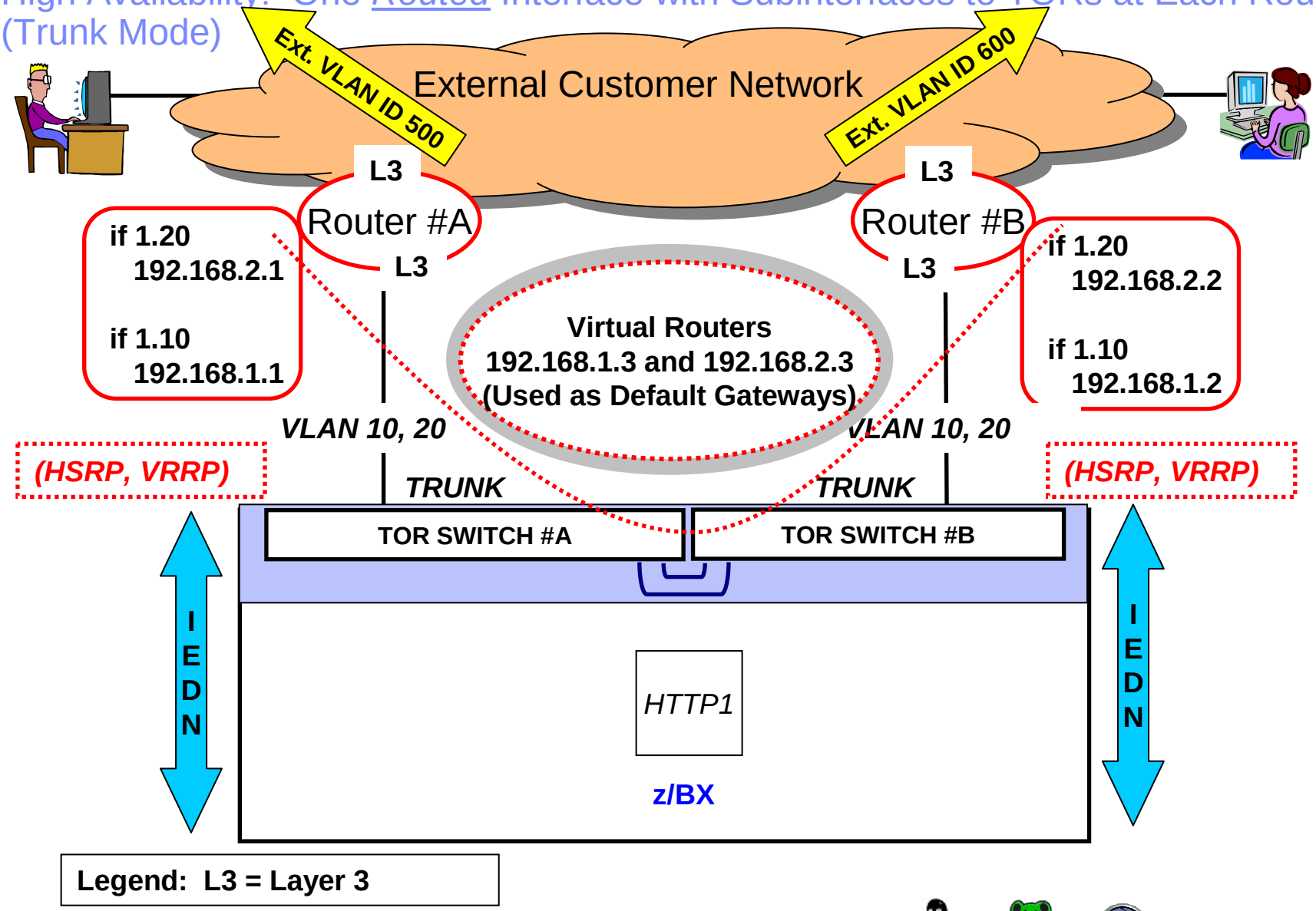


Legend: L3 = Layer 3

Cisco IOS CLI: Sample *Routed* Access Mode Configuration



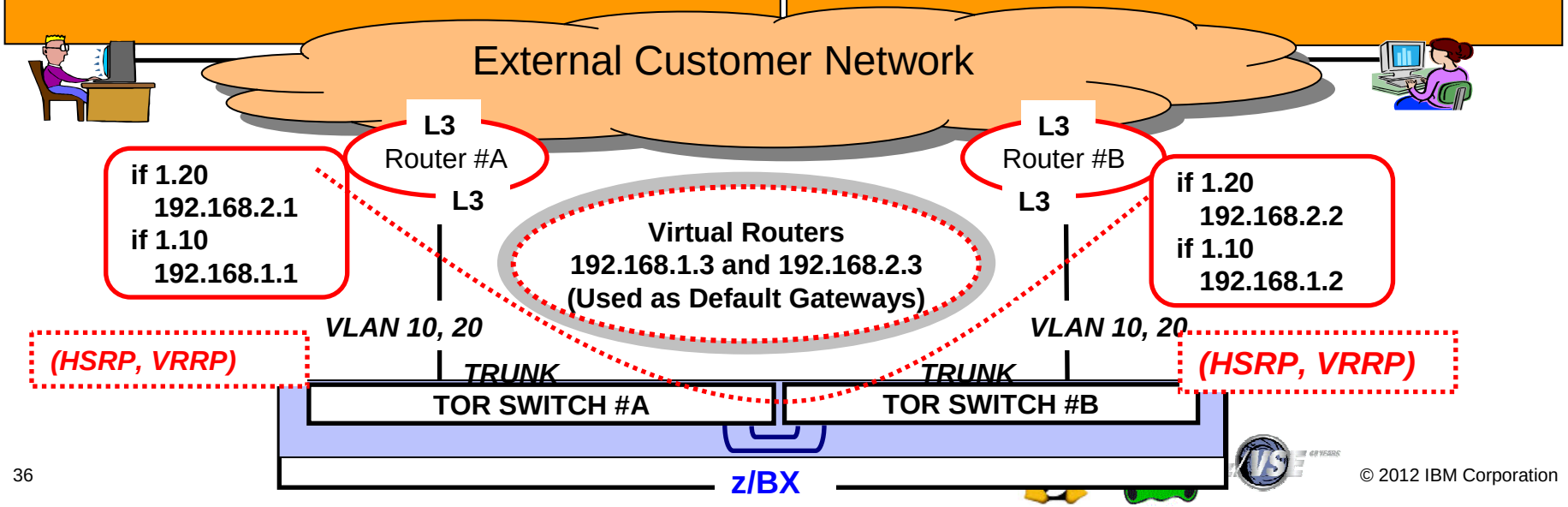
High Availability: One *Routed* Interface with Subinterfaces to TORs at Each Router (Trunk Mode)



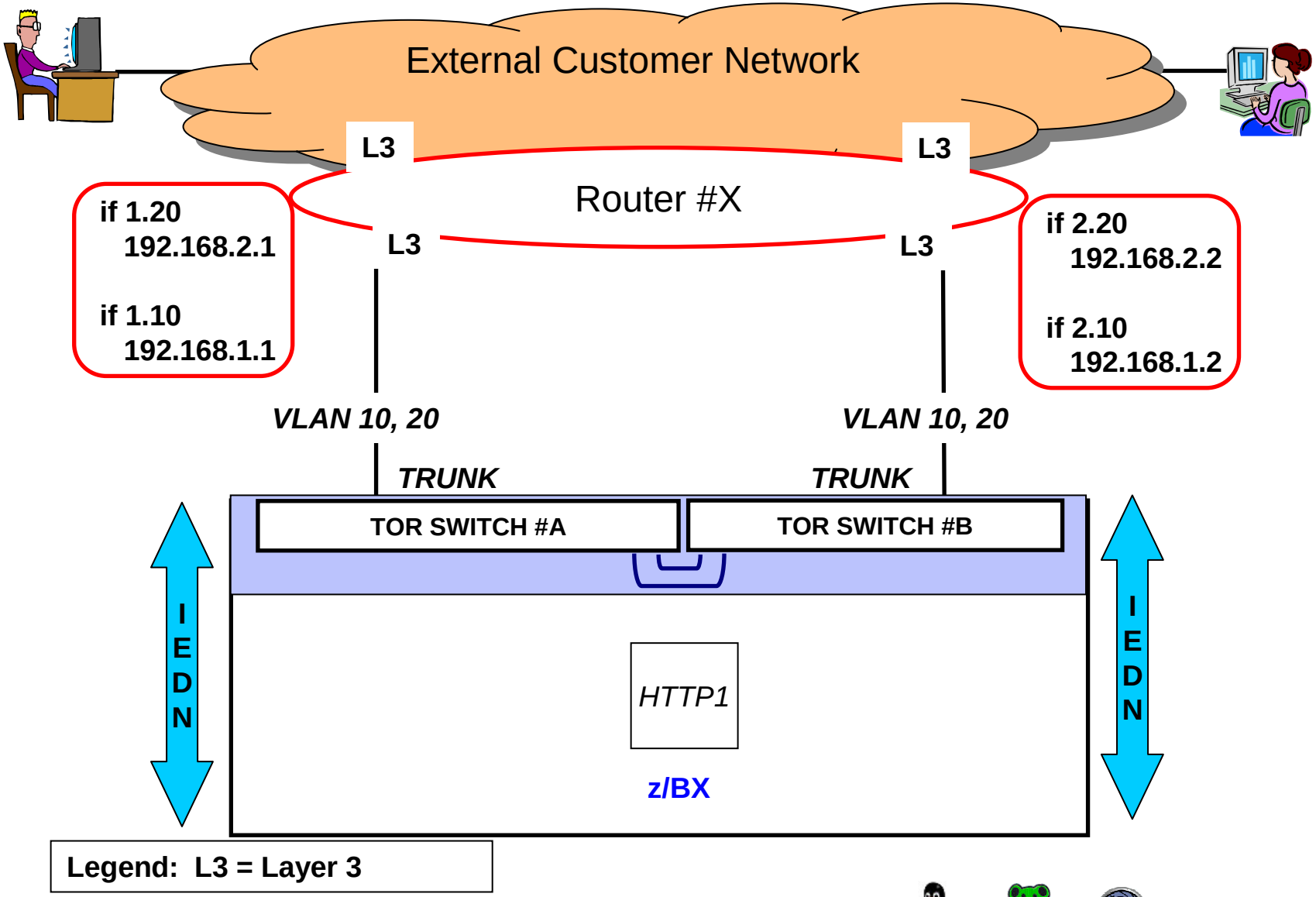
Cisco IOS CLI: Sample *Routed* Trunk Mode Configuration (Two Routers)

```
Interface gigabitethernet 1/1
No switchport
No shutdown

Interface gigabitethernet1/1.10
Encapsulation dot1Q 10
Ip address 192.168.1.1   255.255.255.0
no shutdown

Interface gigabitethernet1/1.20
Encapsulation dot1Q 20
Ip address 192.168.2.1   255.255.255.0
no shutdown
```

```
Interface gigabitethernet 1/1
No switchport
No shutdown

Interface gigabitethernet1/1.10
Encapsulation dot1Q 10
Ip address 192.168.1.2   255.255.255.0
no shutdown

Interface gigabitethernet1/1.20
Encapsulation dot1Q 20
Ip address 192.168.2.2   255.255.255.0
no shutdown
```

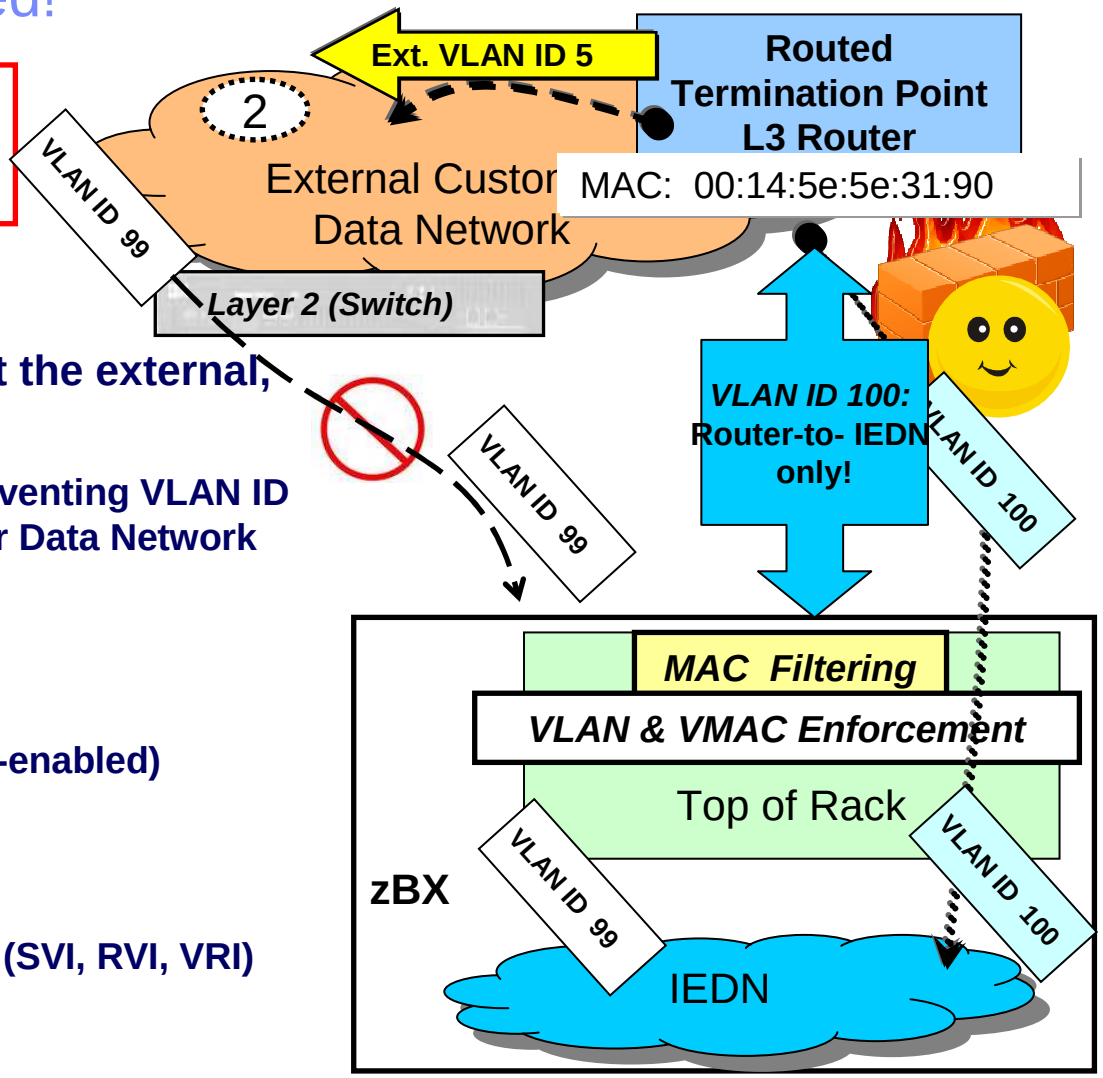


High Availability: Duplicate *Routed* Interfaces with Sub-interfaces to TORs at 1 Router



What is the BIG Mistake You Must Avoid? No Switching Protocol (Layer 2) Messages Permitted!

2. Enter through a Router connection to the TOR – Switch connections not permitted!



❖Selected IEDN VLANs terminate at the external, Layer 3 Routing platform.

•Integrity of the IEDN is preserved by preventing VLAN ID collisions between the external Customer Data Network and the IEDN!

•Routed termination points ONLY
- **✓Dedicated Router Platform**
- **✓Operating System Platform (routing-enabled)**
- **✓L2/L3 Switch with**
  - **❑Routed Interface or**
  - **❑Sub-interface definitions**
  - **❑*With Caution*: Virtual Interfaces (SVI, RVI, VRI)**

•No external Layer 2 Switch!
- **✓No Layer 2 Messages to TOR**
- **✓No STP messages**
- **✓No BPDUs, etc.**

Acknowledgement

Our very best thanks belong to

Gwen Dente
IBM Advanced Technical Support, Gaithersburg, MD (USA)

and

Friedrich Michael Welter
IBM STG Systems Software Development, Boeblingen, Germany

for their input and support to this presentation

Herzlichen Dank für Ihre Aufmerksamkeit

Ingo Franzki

Senior IT Specialist
z/VSE Development & Service

IBM Deutschland Research
& Development GmbH
Schoenaicher Strasse 220
71032 Boeblingen, Germany

Phone +49 7031 16-4648
ifranzki@de.ibm.com

Dr. Manfred Gnirss

Senior IT Specialist
Technical Sales Support
Global Client Center
IBM Germany R&D

IBM Deutschland Research
& Development GmbH
Schoenaicher Strasse 220
71032 Boeblingen, Germany

Phone +49 7031 16-4093
gnirss@de.ibm.com