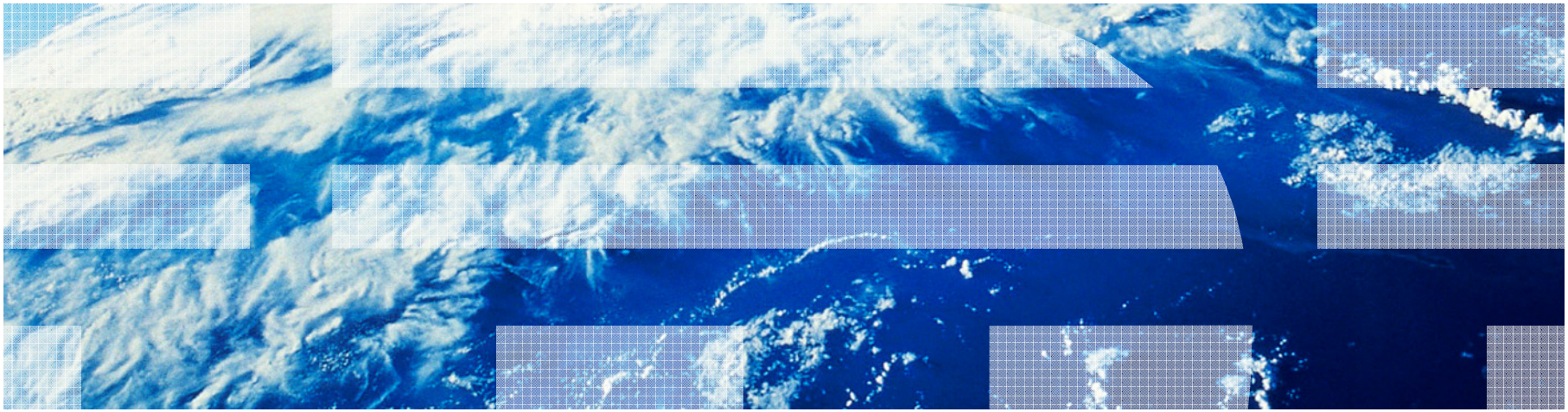
Hochverfügbarkeit und D/R
Wann, warum und wie

Wilhelm Mild
Sen. IT Architect, Böblingen Laboratory
mildw@de.ibm.com

Rene Trumpp
IT Specialist, Böblingen Laboratory
trumpp@de.ibm.com



Trademarks

- This presentation contains trade-marked IBM products and technologies. Refer to the following Web site:

  http://www.ibm.com/legal/copytrade.shtml

Definitions

- High Availability (HA) – Provide service during defined periods, at acceptable or agreed upon levels, and masks *unplanned* outages from end-users. It employs Fault Tolerance; Automated Failure Detection, Recovery, Bypass Reconfiguration, Testing, Problem and Change Management

- Continuous Operations (CO) -- Continuously operate and mask *planned* outages from end-users. It employs Non-disruptive hardware and software changes, non-disruptive configuration, software coexistence.

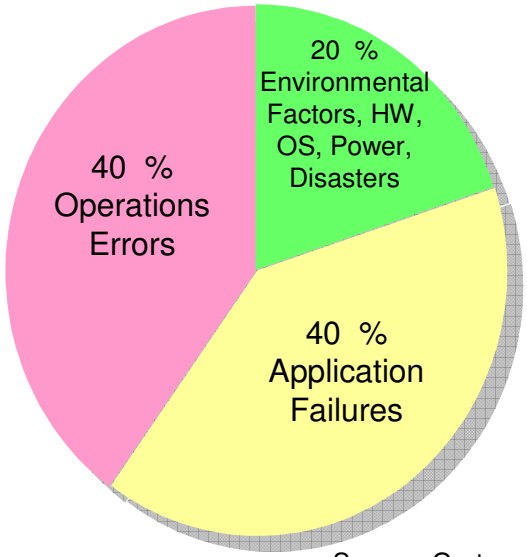
- Continuous Availability (CA) -- Deliver non-disruptive service to the end user 7 days a week, 24 hours a day (there are no planned or unplanned outages).

y.

Business Continuity Issues

What are the reasons for system outages?

- **Planned** outages
  - Maintenance
  - Tests
- **Unplanned** outages
  - Operator errors
    - Lack of application skills
    - Lack of OS skills in heterogeneous environment
  - Application failures
    - SW exceptions
    - Environment / Configuration problems
  - Environmental failures
    - OS failures
    - HW failures
    - Disasters



Source: Gartner Group

Fundamentals of High Availability

- **Redundancy**, Redundancy, Redundancy
  - Duplicate to eliminate single points of failure.
- **Early detection**
  - To keep offline time as short as possible
  - Reduce risk of wrong interpretation and unnecessary failover
  - Keep offline time as short as possible (mean-time-to-repair MTTR)
- **Protect Data Consistency** – Provide ability for data and file systems to return to a point of consistency after a crash.
  - Journaling databases
  - Journaling file systems
  - Mirroring
  - Routine database backups
- **Automate Detection and Failover -** Let the system do the work in order to minimize outage windows.
  - Multipath
  - VIPA –Virtual IP Addresses
  - Monitoring and heart-beating
  - Clustered middleware
  - Clustered operating systems

Differences between HA and DR

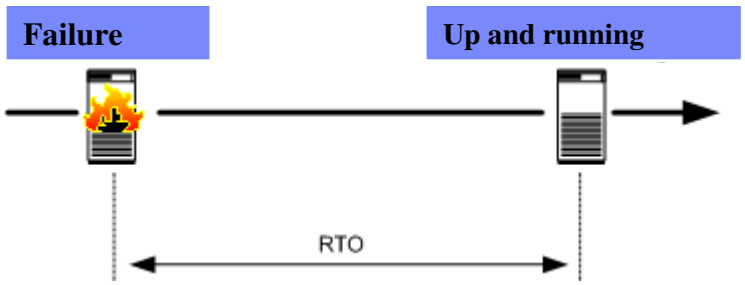
- **High Availability - HA:**
  - Failover is typically realized via duplication and clustering
  - Failover times measured in seconds and minutes
  - Reliable inter-node communication

- **Disaster Recovery - DR:**
  - Failover is typically realized with 2 or more sites in case of disasters
  - Failover times often measured in minutes and hours
  - Unreliable inter-node communication assumed

The borders for HA and DR are soft and the concepts mix depending on the implementation.

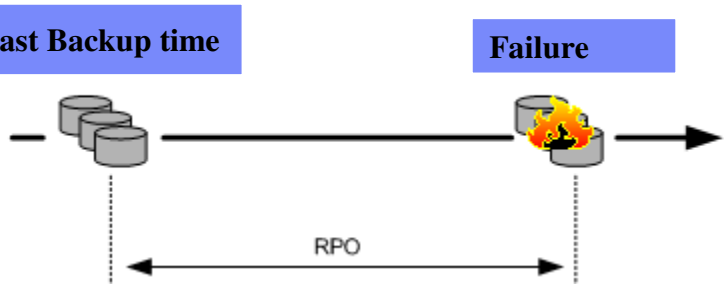
Identify RTO, RPO und NRO

Business Resiliency Plan



Recovery Time Objective (RTO)

What time difference can be between Failure and a total productional run level ?



Recovery Point Objective (RPO)

What is the toleration for data loss?

RPO = "0" means, NULL data loss acceptable
RPO = "5" means, data loss in last 5 min acceptable

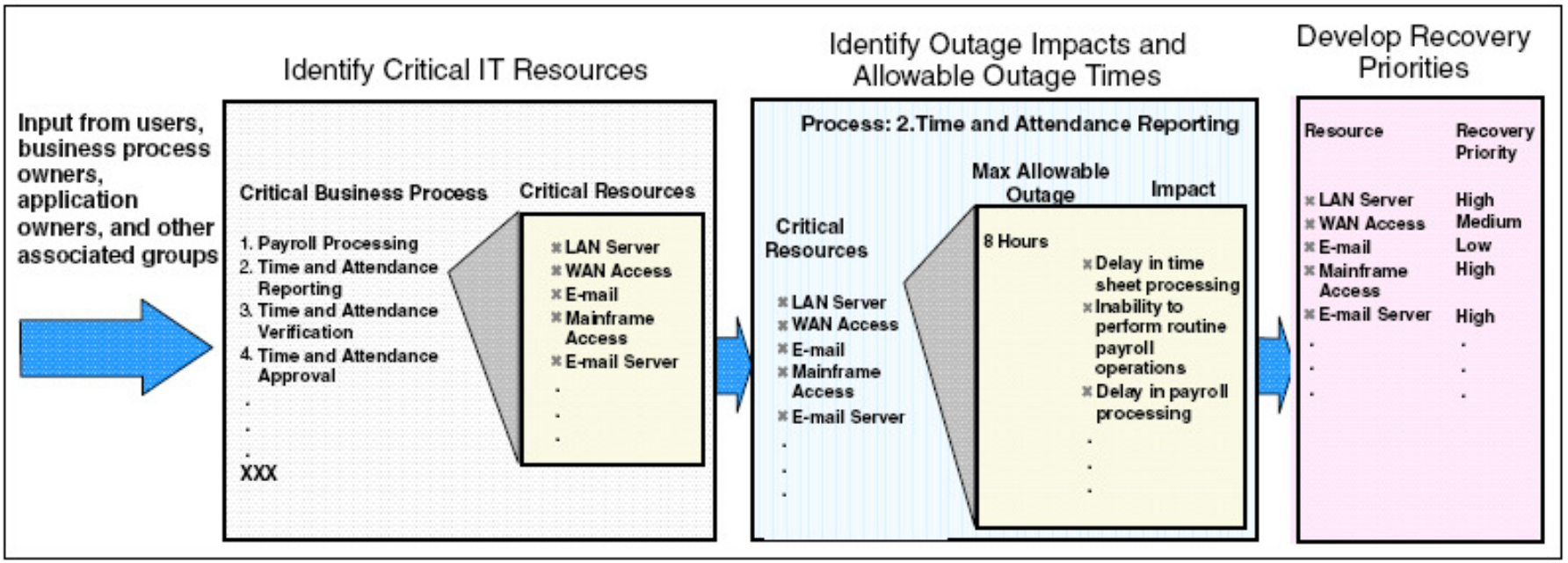
TREND: RPO = 0

Network Recovery Objective (NRO)

Time requirements for network availability.

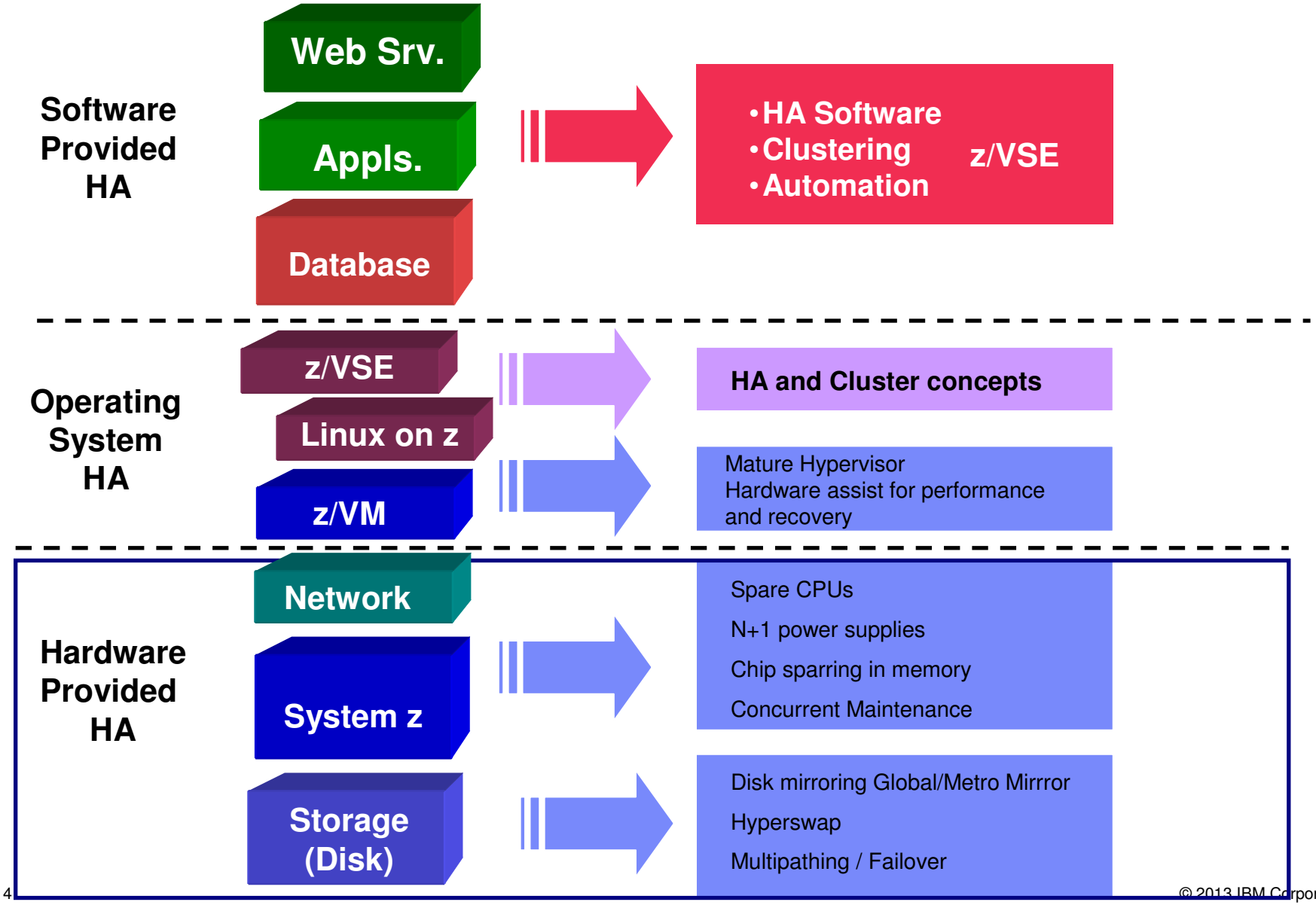
The Business Impact Analysis (BIA)

- IT Resource relation and priorities for DR
- Consider all environments
- Prioritize based on business importance

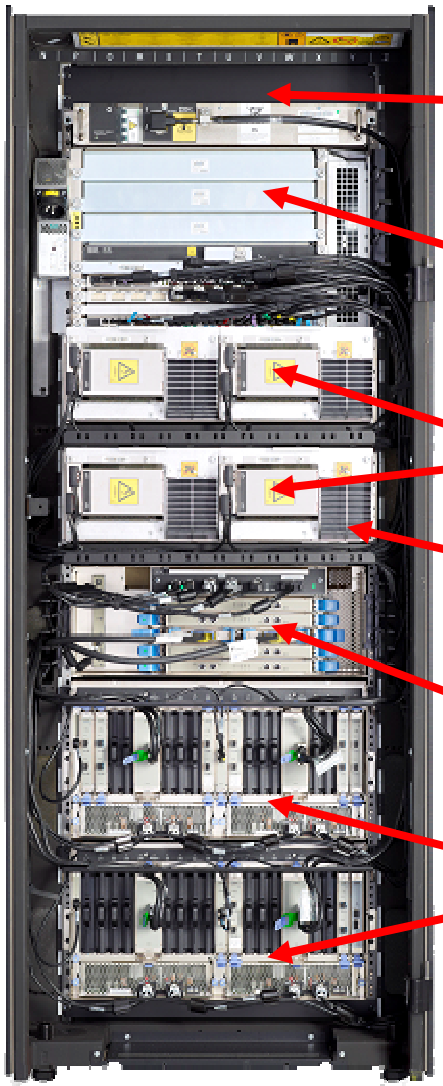


Example of the Business Impact Analysis process

Components of HA with z/VSE and Linux on System z



System z and zEnterprise – HA under the covers



Internal Batteries (redundant)

Power Supplies (redundant)

2 x Processor Drawers, Memory & HCAs

RAIM Memory (Redundant Array of Independent Memory)

I/O Drawer (redundant)

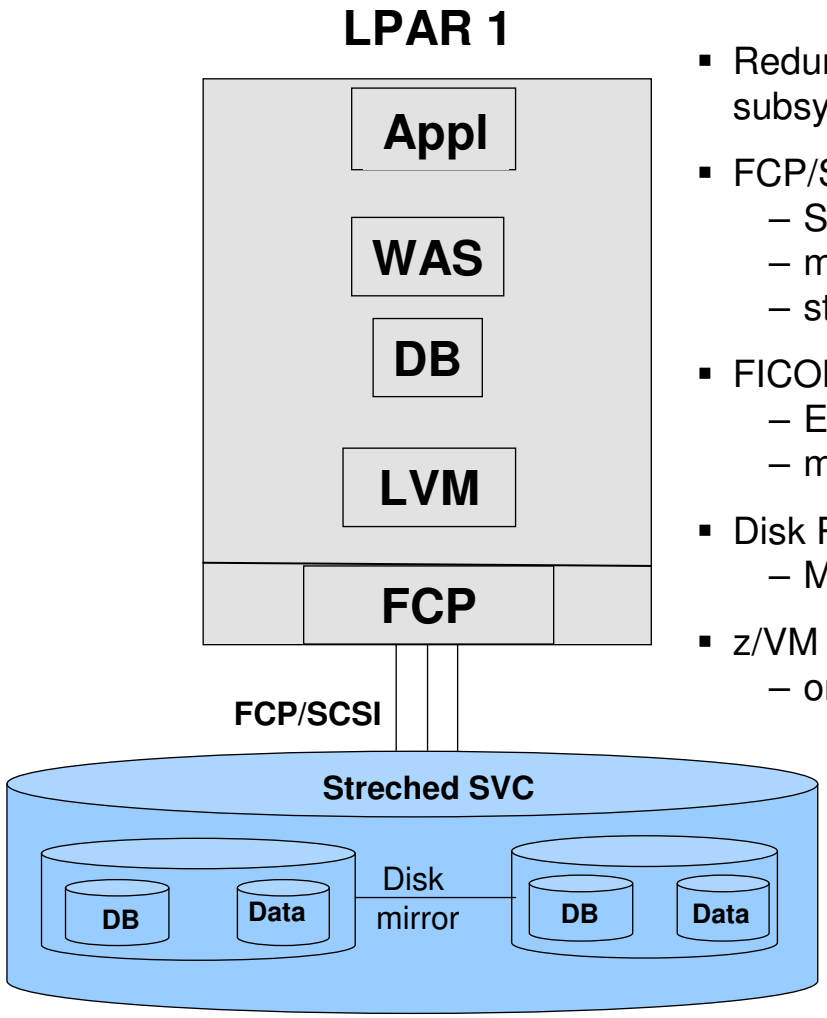
PCIe I/O drawers (redundant)

2 x Support Elements

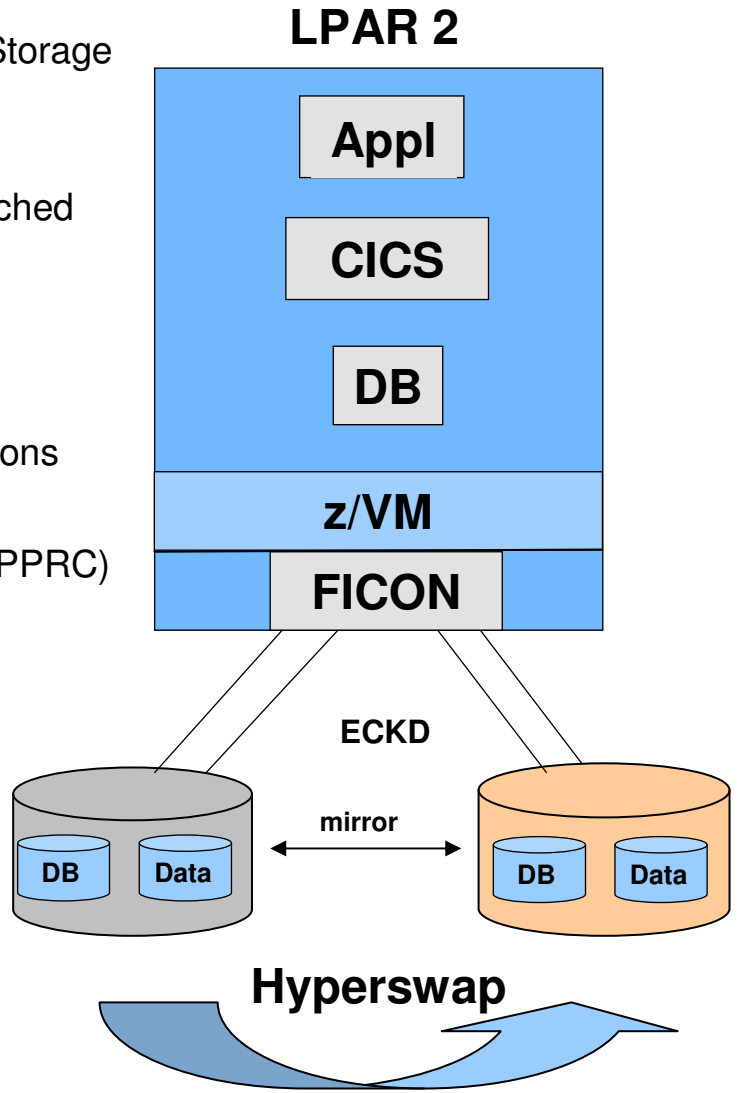
Rear View

Front View

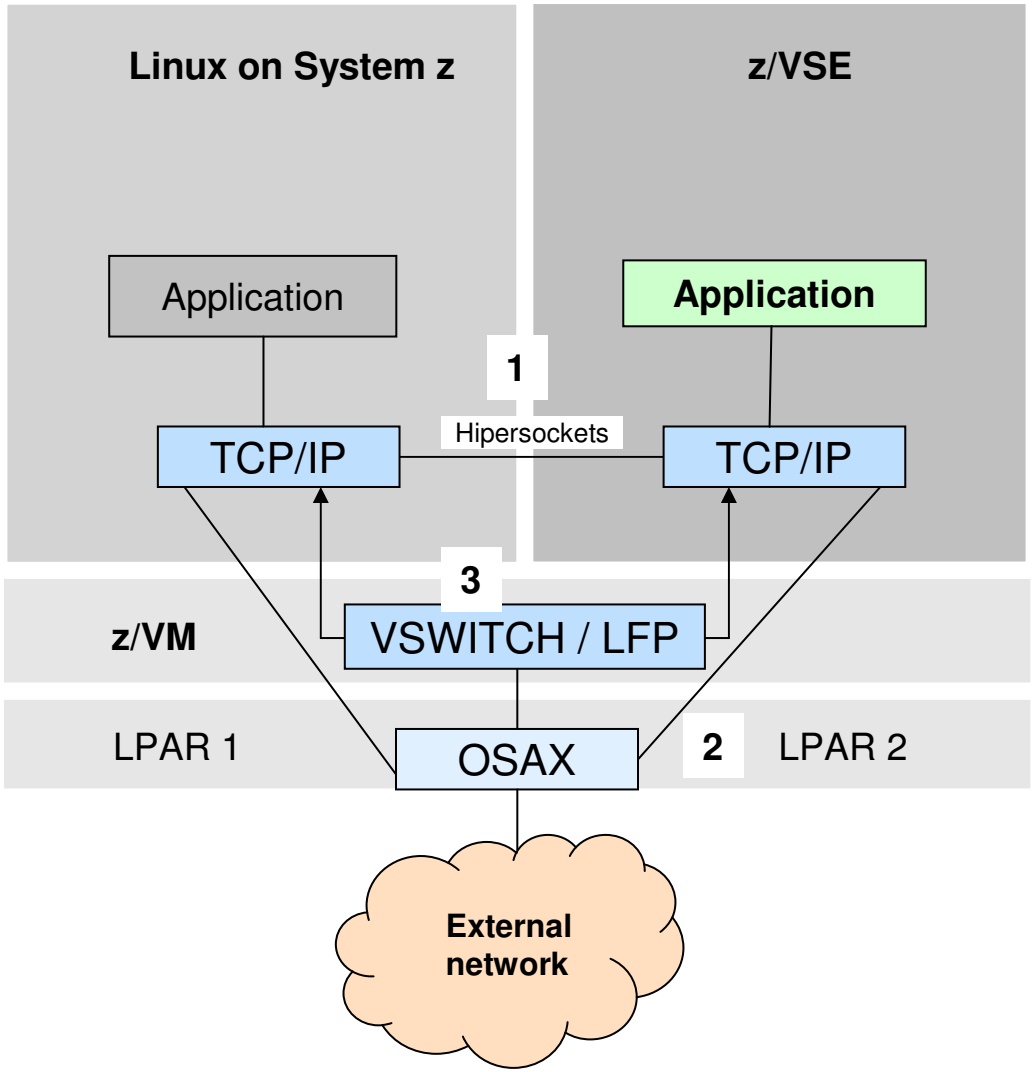
Storage HA Options



- Redundant access to the Storage subsystems
- FCP/SCSI attached disks
  - SCSI disks - LUN attached
  - multi pathing /LVM
  - streched SVC for HA
- FICON attached Disks
  - ECKD disks
  - multi channel connections
- Disk Replication
  - Metro / Global Mirror (PPRC)
- z/VM Hyperswap
  - online disk swap

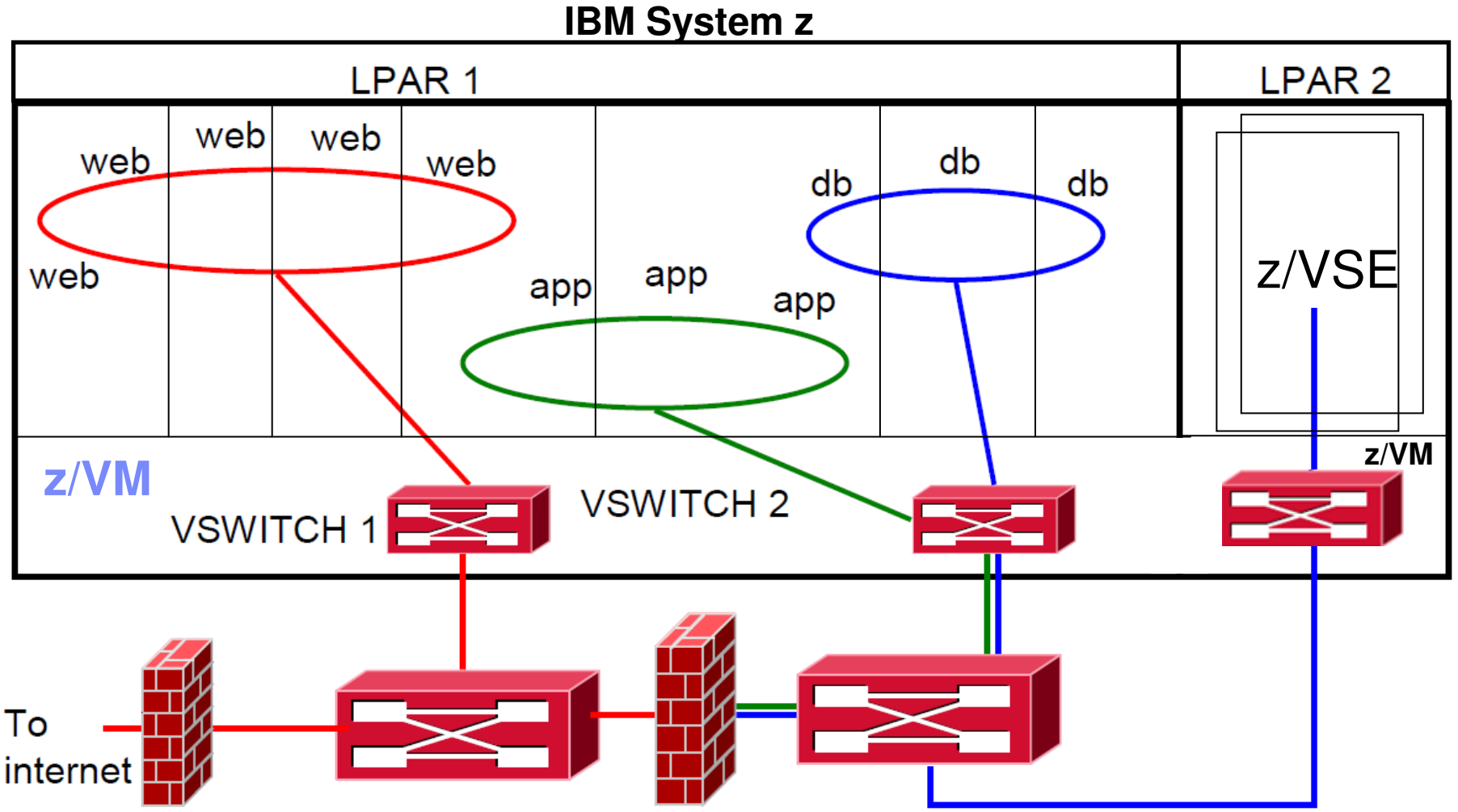


Linux and network alternatives in System z



System z network HA options

Multi-zone Network VSWITCH (red zone physical isolation)



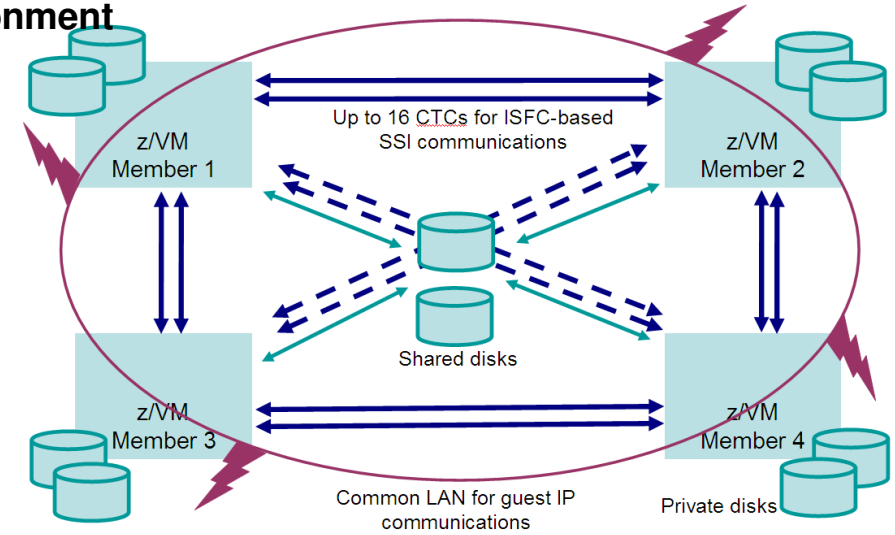
With 2 VSWITCHes, 3 VLANs, and a multi-domain firewall

z/VM V6.2 - Increase Availability for Linux guests

Single System Image, Clustered Hypervisor, Live Guest Relocation

Single System Image (SSI)

- **connect up to four z/VM systems as members of a cluster**
- **Provides a set of shared resources for member systems and their hosted virtual machines**
  - Directory, minidisks, spool files, virtual switch MAC addresses
- **Cluster members can be run on the same or different z10, z196, or z114 servers**
- Simplifies systems management **of a multi-z/VM environment**
  - Single user directory
  - Cluster management from any member
    - Apply maintenance to all members in the cluster from one location
    - Issue commands from one member to operate on another
  - Built-in cross-member capabilities
  - Resource coordination and protection of network and disks



Live Guest Relocation (LGR)

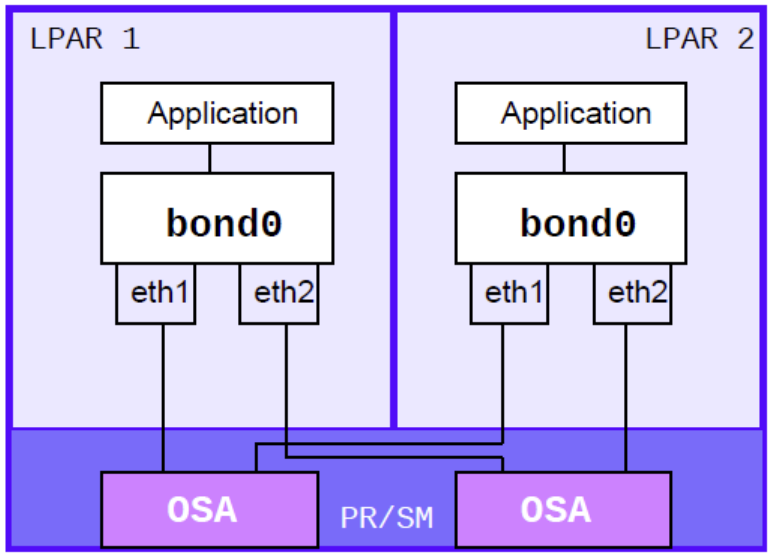
- **– Dynamically move Linux guests from one z/VM member to another**

  Reduce planned outages; enhance workload management
  - Non-disruptively move work to available system resources **and** non-disruptively move system resources to work
  - When combined with Capacity Upgrade on Demand, Capacity Backup on Demand, and Dynamic Memory Upgrade, you will get the best of both worlds

Network HA - Interface Redundancy and Automated Failover

OSA Network HA:

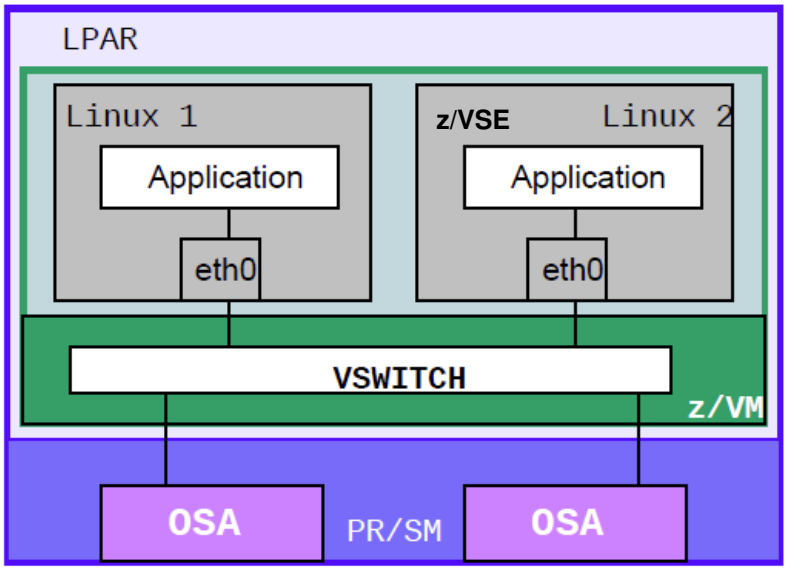
OSA Channel Bonding in Linux



- Linux *bonding* driver enslaves multiple OSA connections to create a single logical network interface card (NIC)
- Detects loss of NIC connectivity and automatically fails over to surviving NIC
- Active/backup & aggregation modes
- **Separately configured for each guest**

z/VM Network HA:

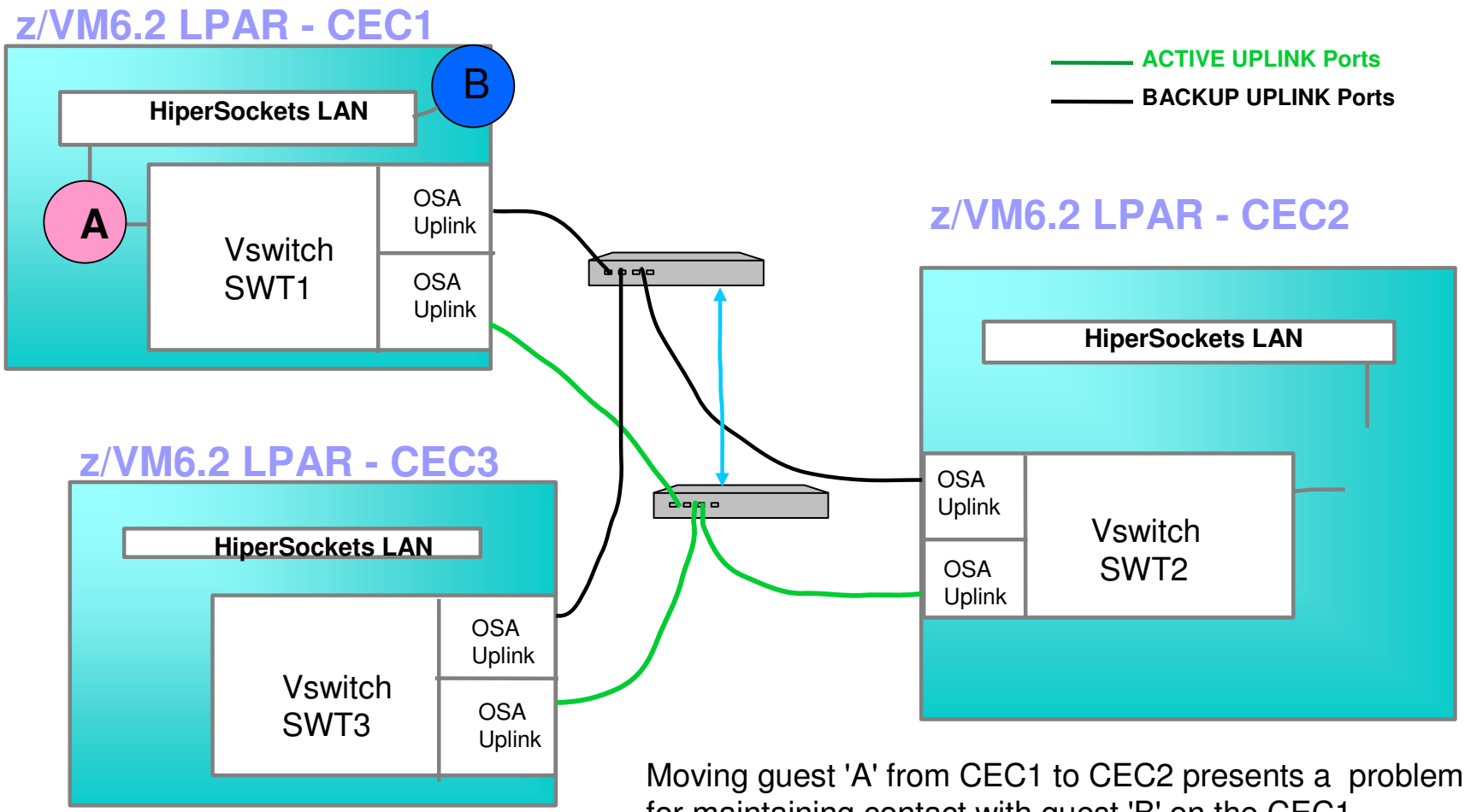
Port aggregation



- z/VM *VSWITCH* enslaves multiple OSA connections. Creates virtual NICs for each Linux guest
- Detects loss of physical NIC connectivity and automatically fails over to surviving NIC
- Active/backup & aggregation modes
- **Centralized configuration benefits all guests**

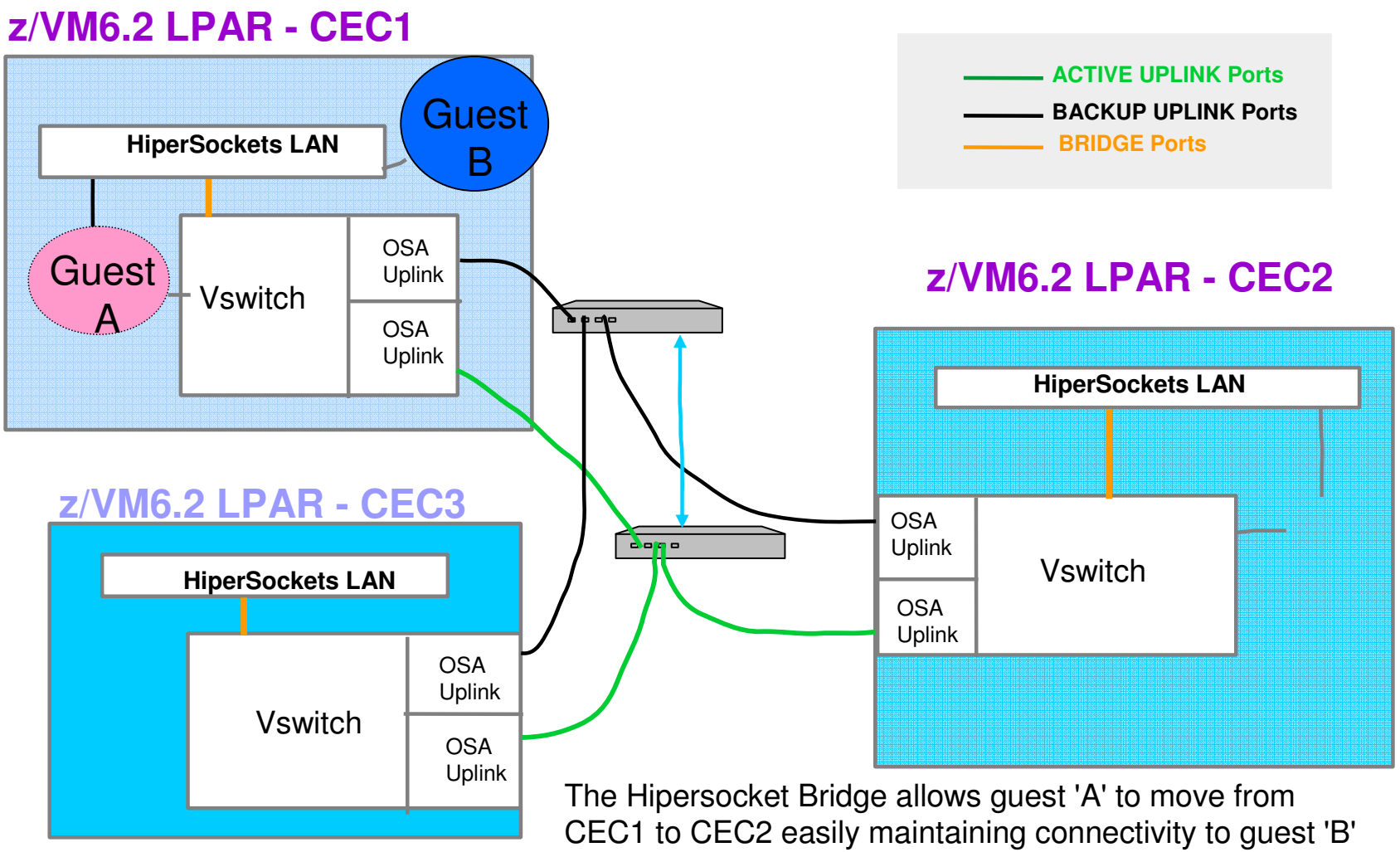
VSWITCH Topology for HA or DR

A typical Vswitch topology for multiple CECs. Active and Backup Uplink Ports to redundant Ethernet switches.



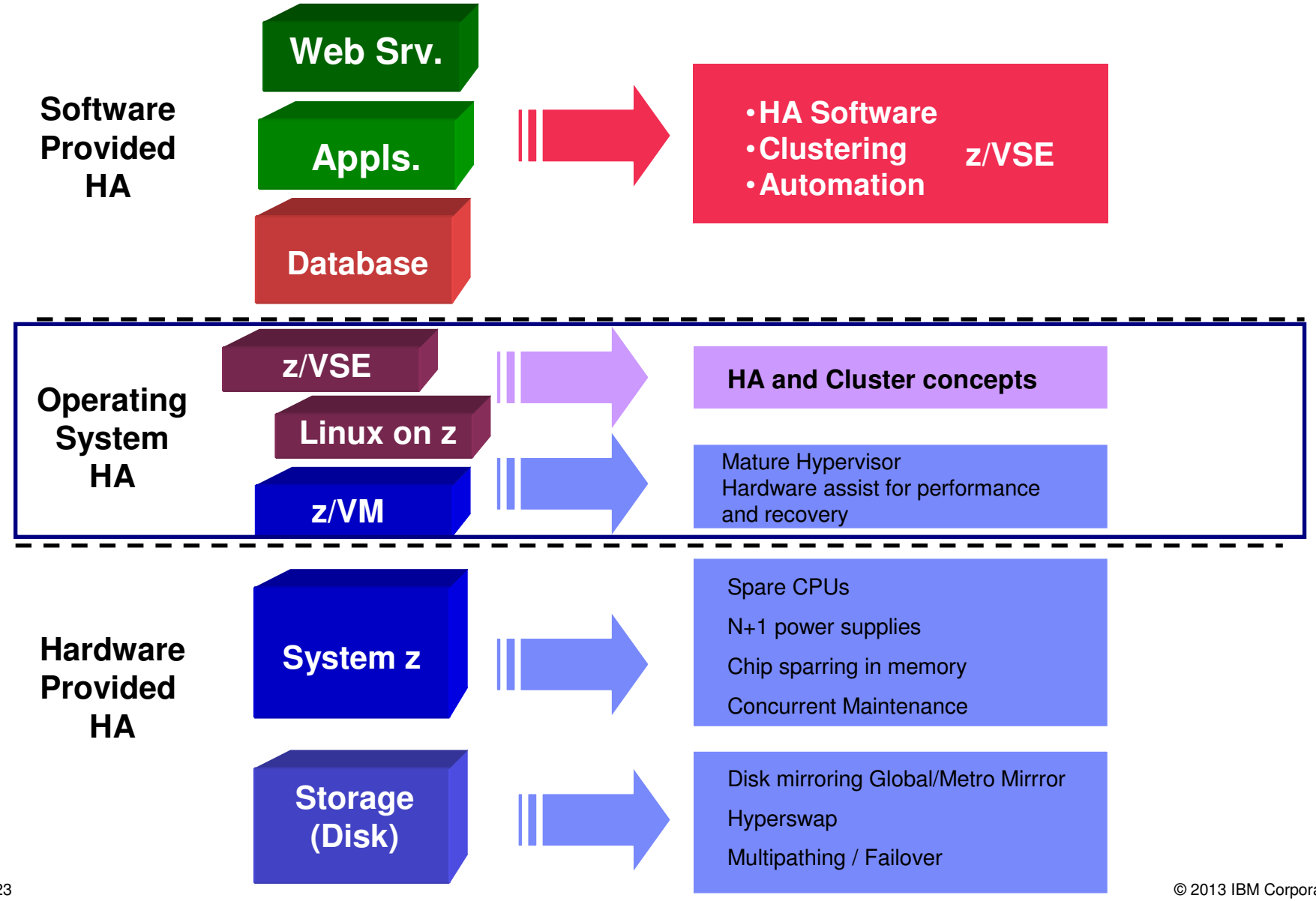
Moving guest 'A' from CEC1 to CEC2 presents a problem for maintaining contact with guest 'B' on the CEC1 hipersockets LAN segment.

VSwitch – With Hipersocket Bridge



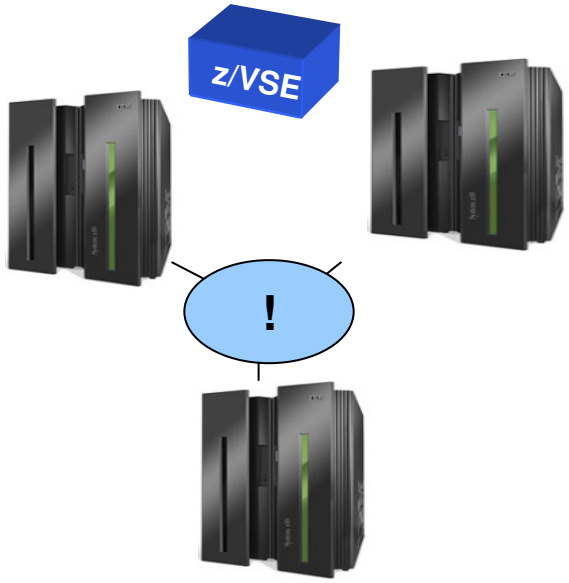
The Hipersocket Bridge allows guest 'A' to move from CEC1 to CEC2 easily maintaining connectivity to guest 'B'

Components of HA with z/VSE and Linux on System z

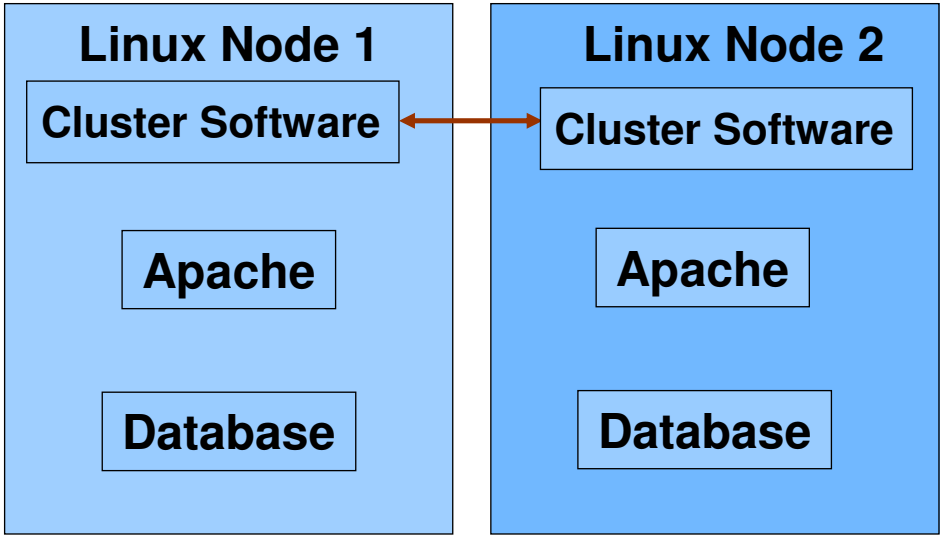


High Availability for Operating Systems: z/VSE and Linux

- z/VSE HA concepts



- Linux on z HA using cluster software



Program directed re-IPL

Program directed re-IPL is designed to allow an operating system on System z to re-IPL without operator intervention. This function is supported for both SCSI and ECKD™ devices.

Computer Cluster Definition

- A computer cluster consists of a set of loosely connected computers that work together so that in many respects they can be viewed as a single system. (Wikipedia definition: Computer Cluster)

High Availability Cluster

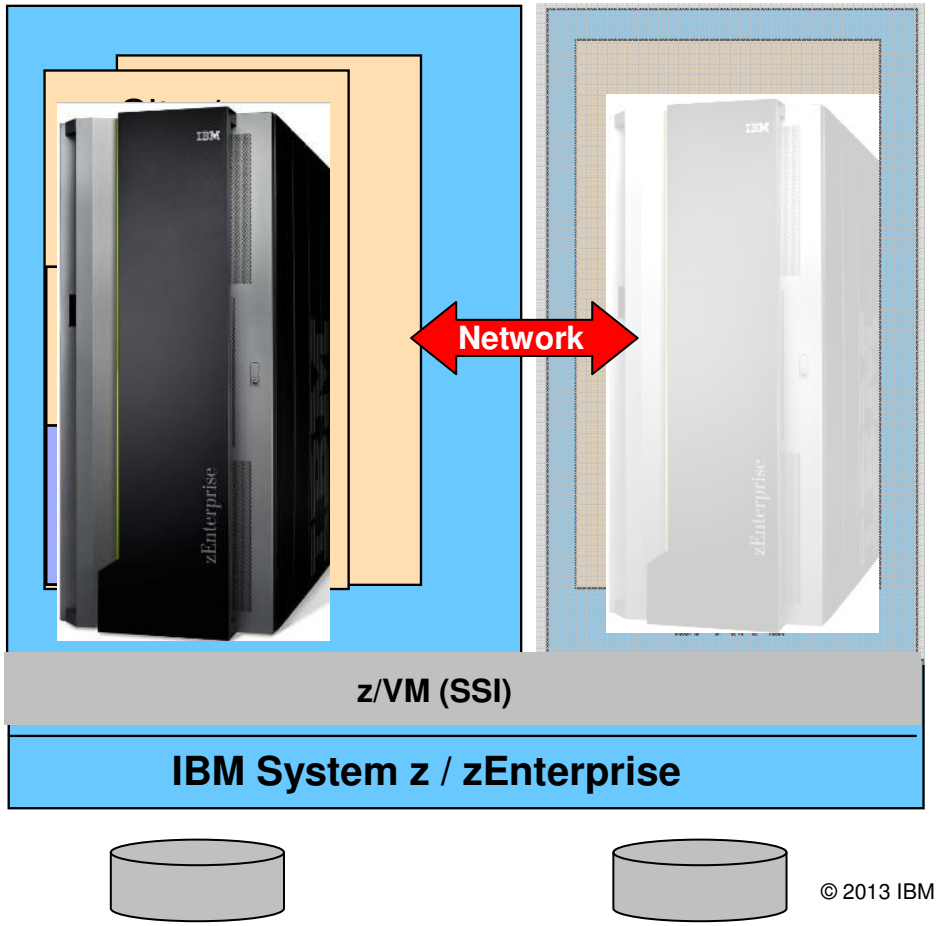
- A computer cluster where each cluster operates as workload node. When one node fails another node takes over the entire workload: IP address, data access, services, etc.
- The key of High Availability is avoiding single points of failure
- High Availability adds costs because of added complexity due to redundant resources in the environment

High Availability scenario as Active/Passive with System z

- **Active / Passive Deployment**.
  - Workload normally contained at Site 1, standby server capability at Site 2
  - Primary and secondary disk configurations active at both sites.
  - During fail over, Capacity Upgrade on Demand (CUoD) adds resources to operational site, and standby servers are started. Helps save hardware and software costs, but requires higher recovery time.

- **Hot / Cold scenario**
  - Workload is not split.
  - Each site is configured to handle all operations
  - Cold environment needs longer to get active – often used in DR

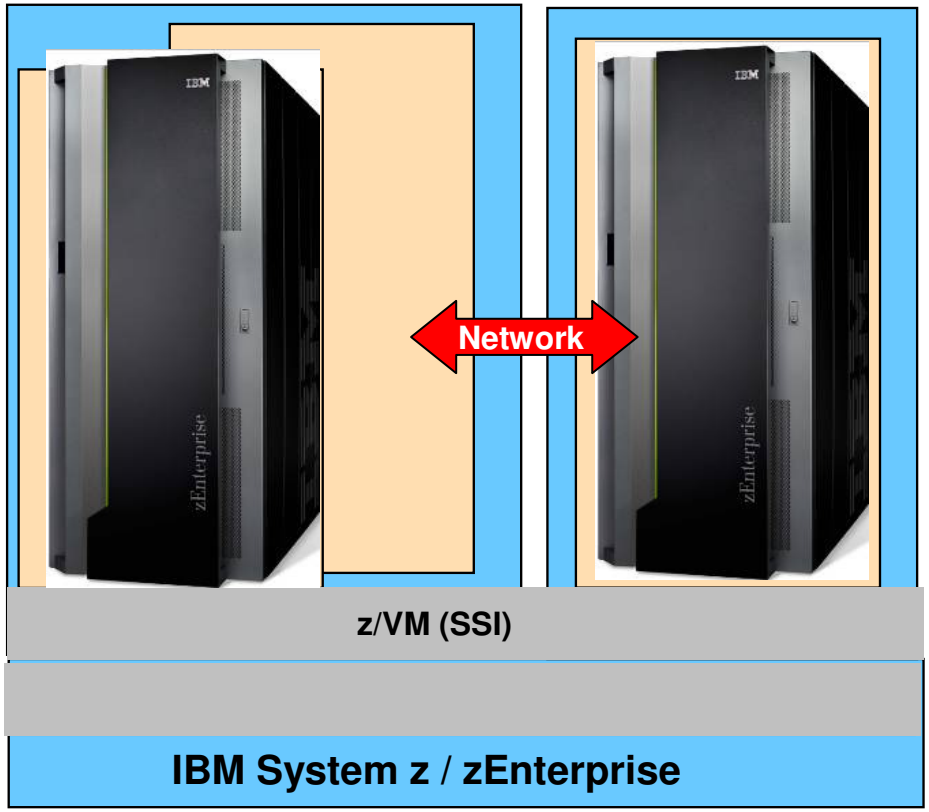
- **Hot / Warm scenario**
  - Workload is not split
  - Each site is configured to handle all operations
  - Warm environment is idling.



High Availability with an active/active environment on System z

- **Active / Active Deployment -Expendable work.**
  - Workload is normally split between 2 or more sites
  - Each site is (over) configured to be able to instantly cover the workload if needed.
  - During normal operation, excess capacity at each site is consumed by lower priority, work like development or test activitries
  - In a failover situation, low priority work is stopped to free up resources for the production site's incoming work.

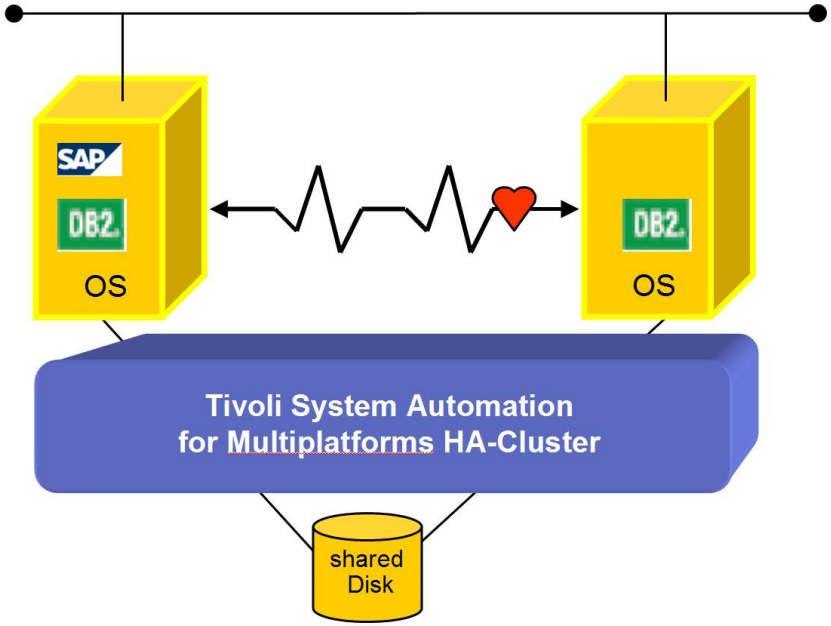
- **Capacity Upgrade on Demand (Active / Active)**
  - Workload is normally split between sites.
  - Each site is configured with capacity to handle normal operations
  - Special setup with Capacity Upgrade on Demand (CUoD) and CBU.
  - In a failover situation, additional CPUs are enabled at the operational site.



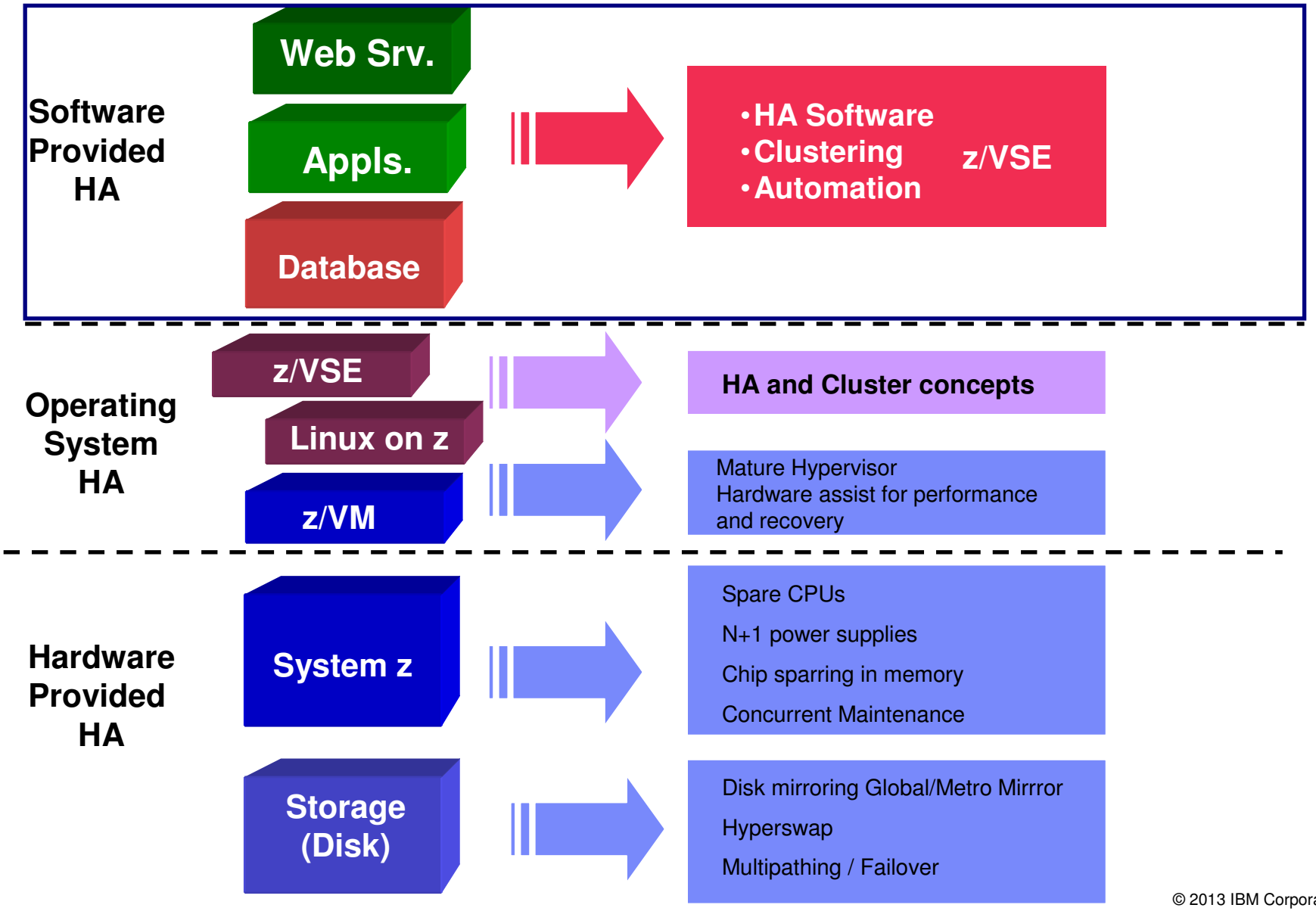
High Availability Clustering

- **Tivoli System Automation for Multiplatforms**
  - **for multiplatforms**
  - **for distributed heterogeneous environments**

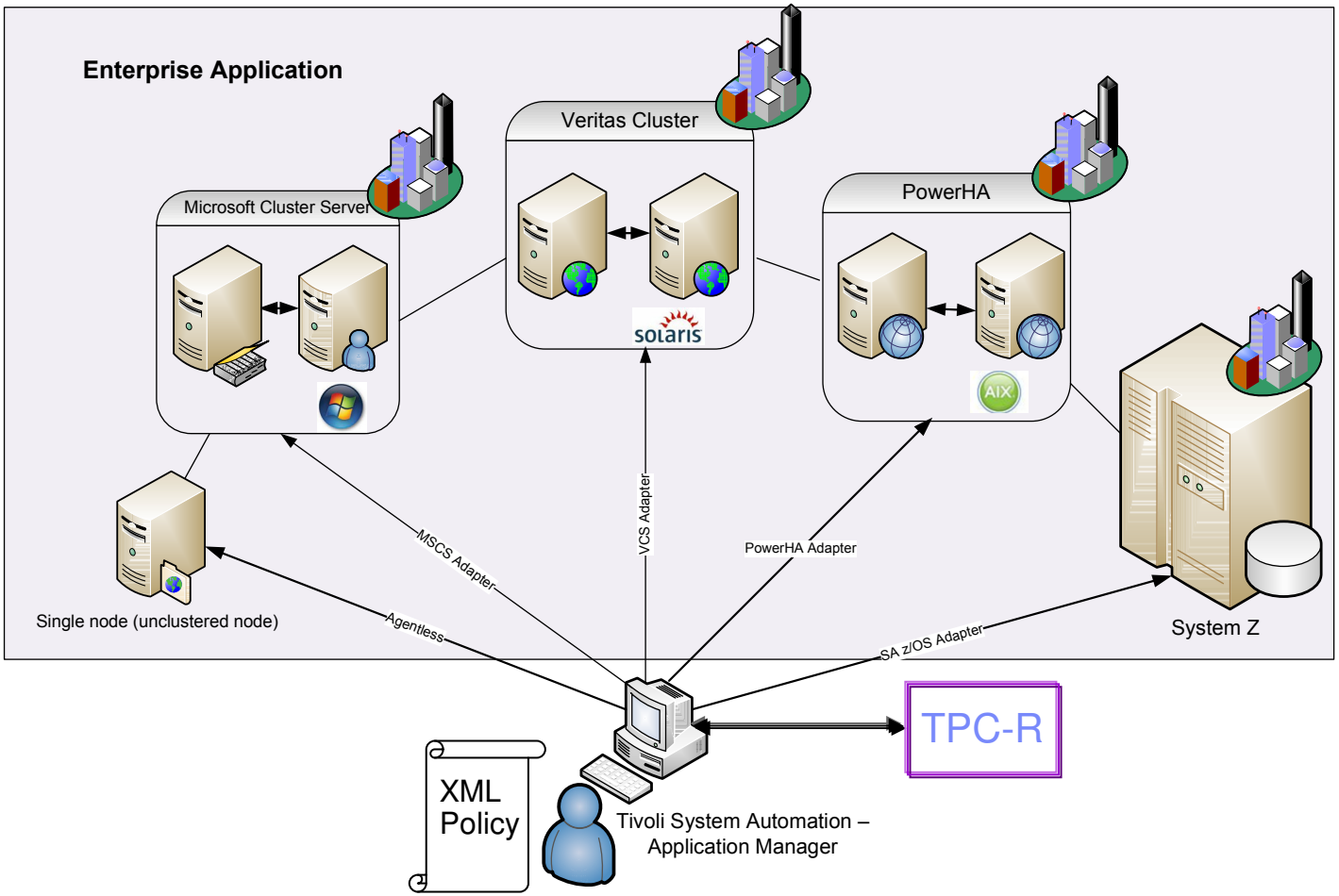
- **Linux-HA Open Source package**
  - **for Linux environments**



Components of HA with z/VSE and Linux on System z

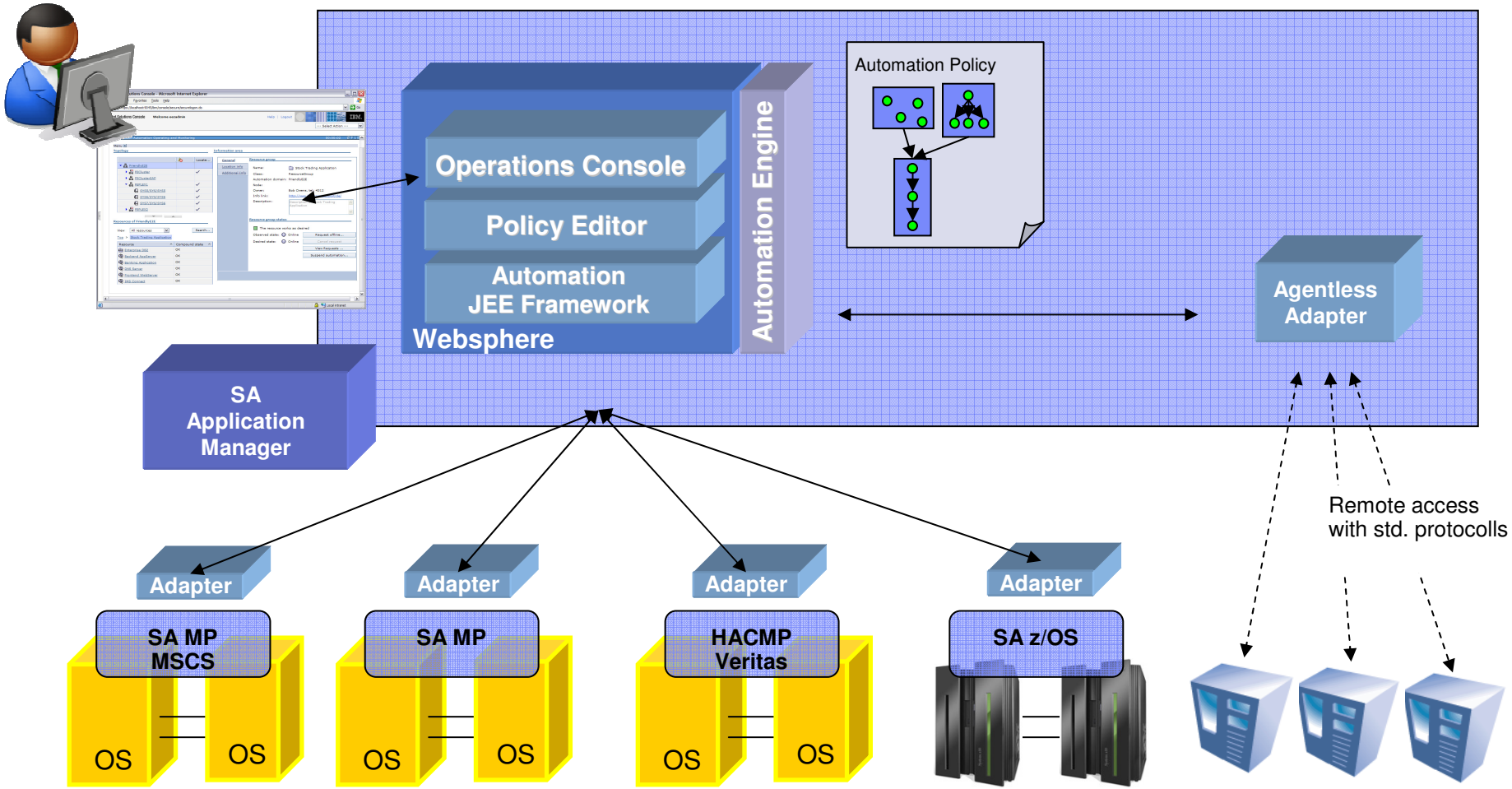


Tivoli System Automation (TSA) – Application Manager Overview



"SA MP and SA AM help us manage business applications over 200 AIX Lpar's and saves us over 3 person-years of scripting effort. We also manage application components on SA z/OS from SA AM for end to end application management" – large financial customer

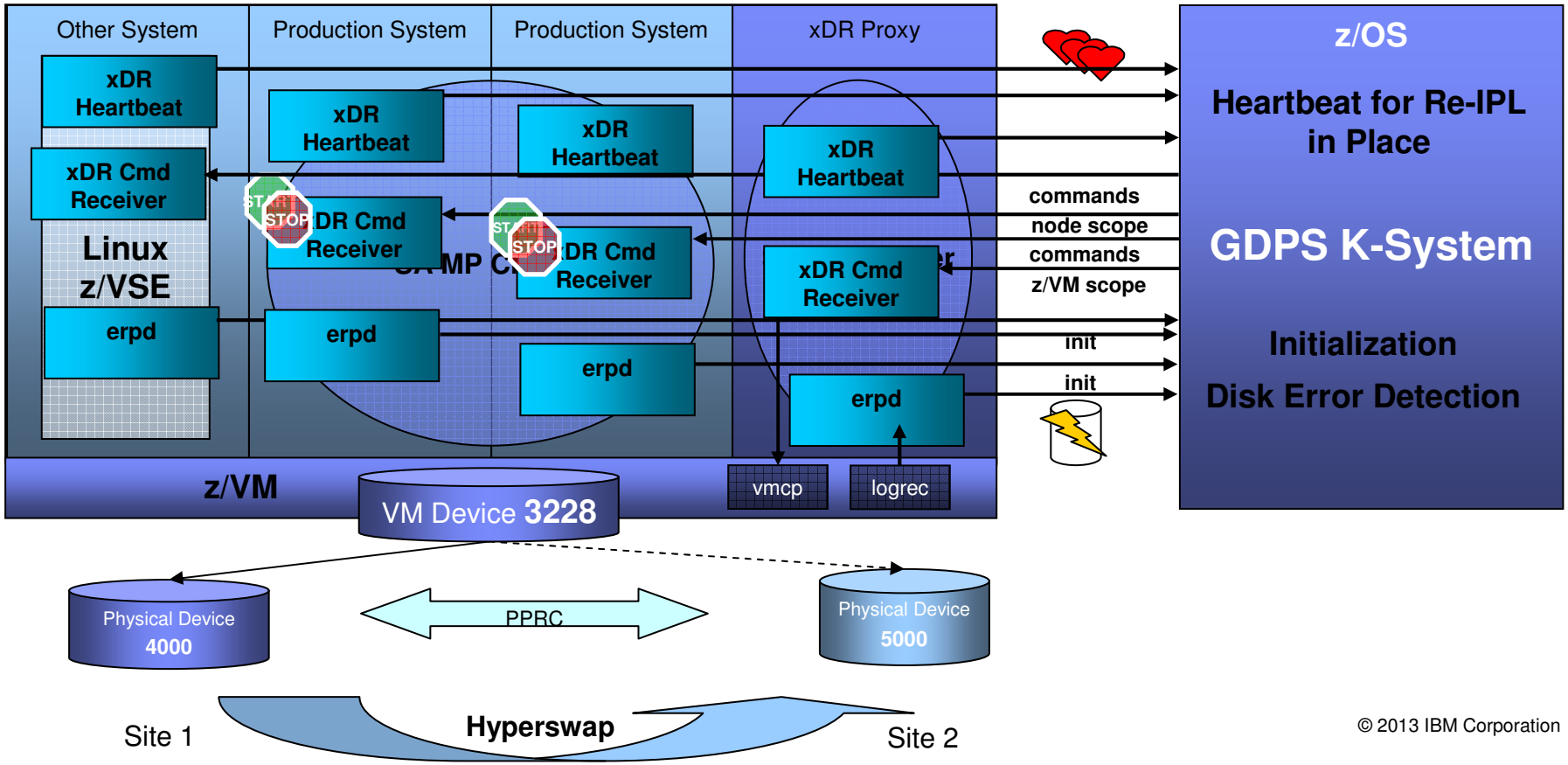
SA Application Manager Adapter Infrastructure



GDPS and xDR with z/VM guests – High availability

- **Proxy**
  - One linux system is configured as Proxy for GDPS which has special configuration
    - (Memory locked, Access rights to VM, One-Node-Cluster)
  - Is used for tasks that have z/VM scope
    - HyperSwap, shutdown z/VM, IPL z/VM guest

- **Production Nodes**
  - Run Linux Workload
  - Are used for local actions (Shut down node, Maintenance Mode)
- **Other Systems**
  - Enabled for HyperSwap via xDR Proxy (Linux, z/VSE)
  - No re-IPL in place, no start/stop via GDPS (init, reipl, maint)



z/VSE GDPS Client

Easy setup:

To setup the z/VSE GDPS Client, you just have to

- Use Job SKGDPSCF (provided in VSE/ICCF library 59) to create the configuration file
- Use job SKSTGDPS (provided in VSE/ICCF Library 59) to start the GDPS Client

Easy usage:

- During startup, it sends the „init“ command to the GDPS K-System
- When it is running, it sends heartbeats
- GDPS K-System can force a switch to a second GDPS K-System (automatically)
- If you want to shutdown z/VSE (planned outage), you can send a „Start maintenance“ command to the GDPS K-System

Configuration

Before the GDPS Client can be started, you must configure the GDPS Client using sample Job SKGDPSCF (provided in VSE/ICCF library 59). Per default, the job creates a member with name IESGDPCF.Z in sublibrary PRD2.CONFIG. If required, change the default values before submitting this job.

```
$$ JOB JNM=GDPSCFG,DISP=D,CLASS=0
// JOB GDPSCFG CATALOG GDPS CLIENT CONFIGURATION MEMBER
// EXEC LIBR,PARM='MSHP'
ACCESS S=PRD2.CONFIG
CATALOG IESGDPCF.Z REPLACE=Y
* NAME OF THIS GDPS CLIENT (MUST BE SAME AS YOUR HOSTNAME)
GDPSCNAME='MYVSE'
* OPTIONAL CLUSTERNAME
* CLUSTERNAME='MYCLUSTER'
* RECEIVER (LOCAL) PORT
RECEIVERPORT='16111'
* INTERVAL (IN SECONDS) BETWEEN HEARTBEATS
HEARTBEATINTERVAL='10'
* TIMEOUT
TIMEOUT='60'
* CURRENT MAINTENANCE STATUS
MAINTENANCE='OFF'
* CURRENTLY ACTIVE SITE
SITE='1'
* SITE 1 CONFIGURATION
* IP/HOSTNAME OF THE GDPS K-SYSTEM
SITE1_KSYSTEM='123.123.1.1'
* PORT OF THE GDPS K-SYSTEM
SITE1_KSYSTEMPORT='16112'
* SITE 2 CONFIGURATION
SITE2_KSYSTEM='123.123.1.2'
SITE2_KSYSTEMPORT='16112'
/+
/*
/&
$$ EOJ
```

The GDPSCNAME must be same as your hostname. It is used to identify your z/VSE system in your GDPS K-System

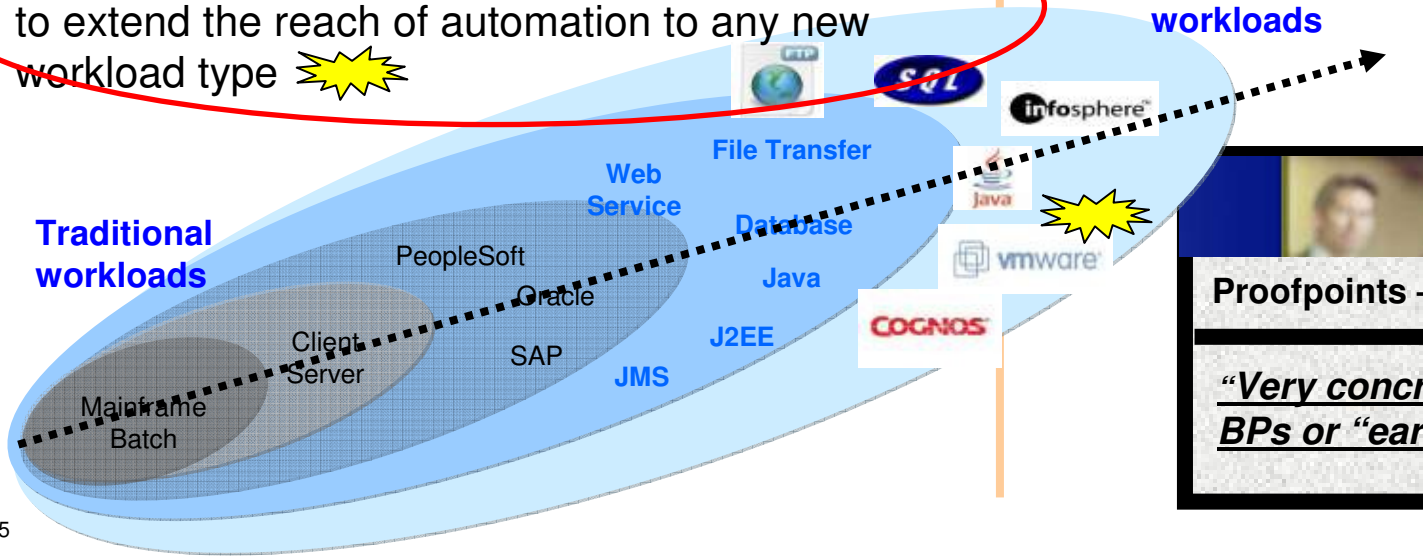
The Clustername is optional. It can be used to build „groups“ of z/VSE systems in you GDPS K-System

You can configure two different GDPS K-Systems.

Application Extensions allow business users to take advantage of processes in a managed approach

New Tivoli Workload Automation application extensible framework

- Customers shifting from traditional backend transaction focused systems to modern systems running web applications and heterogeneous applications
- Workload Automation role is maintaining a single point of control over workloads
- TWS 8.6 easily build and deploy application plugins to extend the reach of automation to any new workload type



Business benefits

- ★ ***Share infrastructure among applications***
- ★ ***Reduces labor costs, enabling to automate new workloads with the same staff of people***
- ★ ***No request for new skill: re-using of workload automation processes and procedures already in place***

Proofpoints – Customer quotes

"Very concrete needs" from BPs or "early adopters"

Summary: High Availability und Disaster Recovery

- **Konzepte haben fließenden Übergang**
- **ALLE Systeme sollten einbezogen werden**
- **Die Schichten von HW bis Software sind zu berücksichtigen**
- **Netzwerk und Drucker sind oft Sonderbehandlung**
- **System z ist hoch verfügbar durch Design**

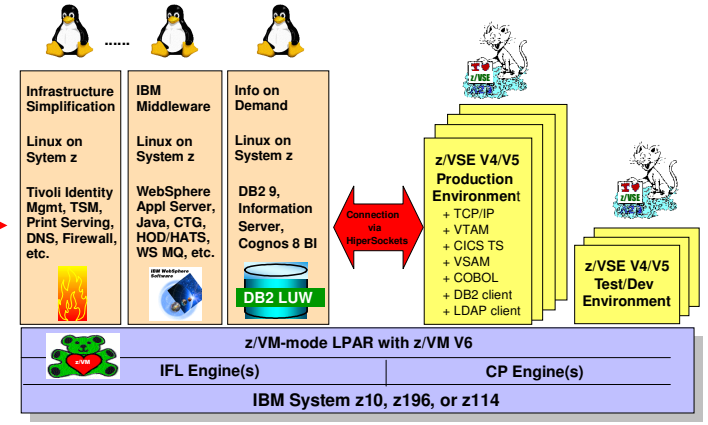
aber denken Sie dran:

Die Kette ist so 'z' stark wie das schwächste 'x' Glied !!!

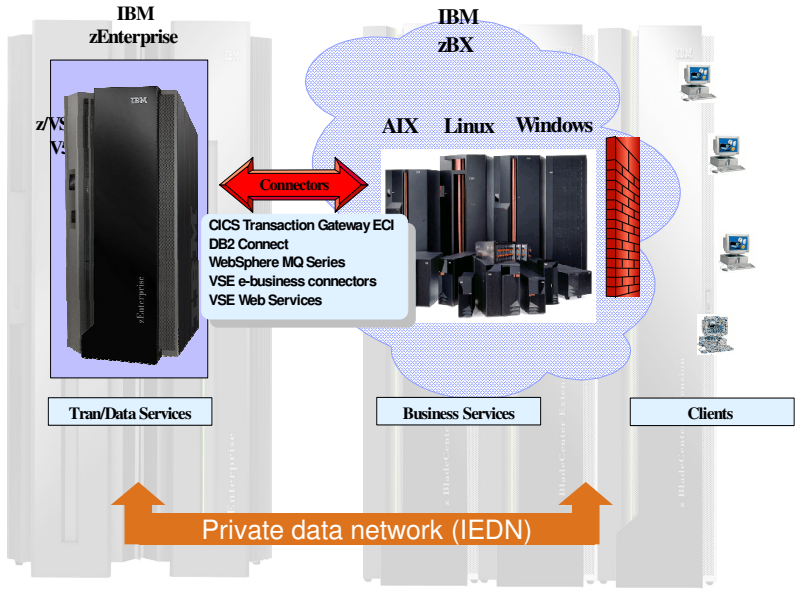


IBM zEnterprise can do IT all - Think inside the Box and/or think global !

Rational
DB2
Lotus
Tivoli
WebSphere
COGNOS



- Protect **existing VSE investments**
- Integrate **using middleware and VSE connectors**
- Extend **with another platform to access new applications & solutions**



Questions?

Wilhelm Mild

IBM IT Architect

IBM Deutschland Research & Development GmbH
Schönaicher Strasse 220
71032 Böblingen, Germany

Office: +49 (0)7031-16-3796
mildw@de.ibm.com

Rene Trumpp
IBM Specialist

IBM Deutschland Research & Development GmbH
Schönaicher Strasse 220
71032 Böblingen, Germany

Office: +49 (0)7031-16-2106
trumpp@de.ibm.com