# Dynamic Routing:
# Exploiting HiperSockets and Real Network Devices

## Session 8447

Jay Brenneman
rjbrenn@us.ibm.com

# Trademarks

**The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.**

DB2*
DB2 Connect
DB2 Universal Database
e-business logo*
e-business on demand
HiperSockets
IBM*
IBM eServer
IBM logo*
IMS
Resource Link
S/390*
Tivoli*
Tivoli Storage Manager
TotalStorage*
WebSphere*
z/OS*
z/VM*
zSeries*

* Registered trademarks of IBM Corporation

**The following are trademarks or registered trademarks of other companies.**

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries.

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

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

UNIX is a registered trademark of The Open Group in the United States and other countries.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

* 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.

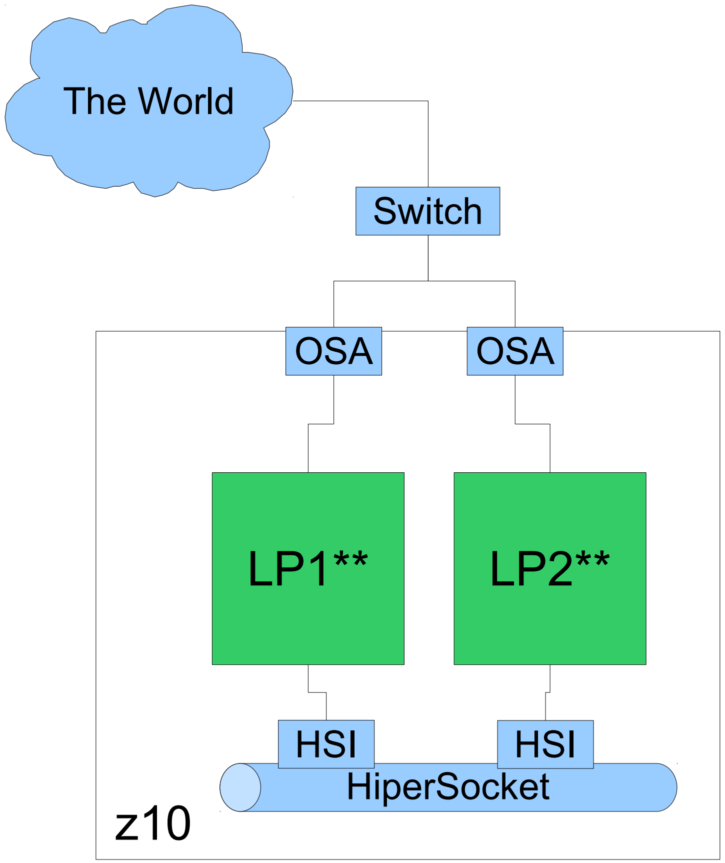
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products 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.

# Agenda

- What is the problem?
- Common solutions and their faults
- OSPF and its faults

# The Problem:

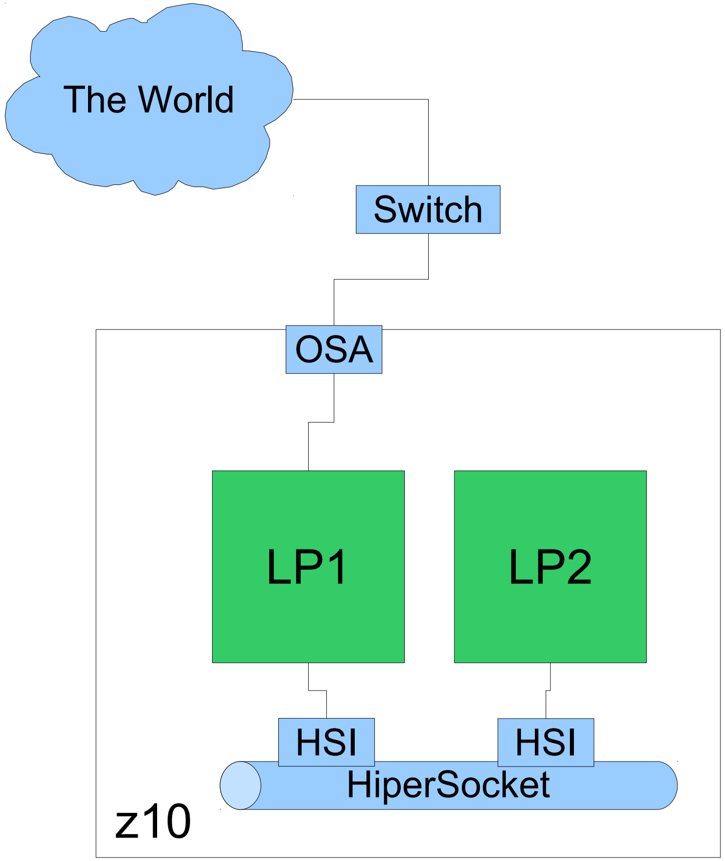


**The diagrams are drawn as LPARs, but these patterns also apply equally to z/VM Guests.

- HiperSockets are very fast, but only work within a CEC*.
- OSA is required to talk to systems outside the CEC.
- How do you exploit HiperSockets while also talking to the rest of the world?

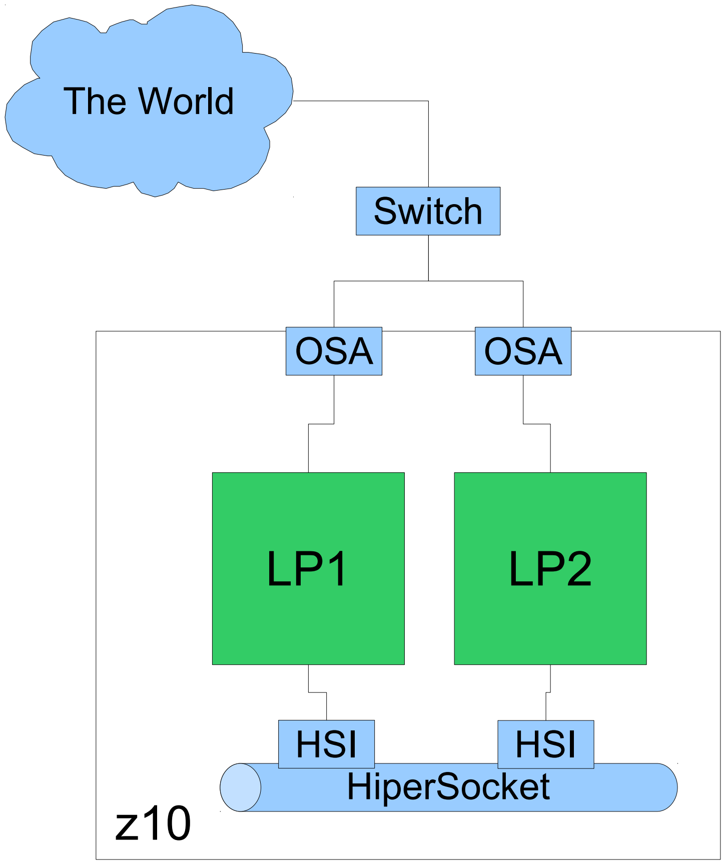
*CEC = Central Electronics Complex
Also Known As: The processor, The CPU, The machine, The big black refrigerator

# The Original Solution:



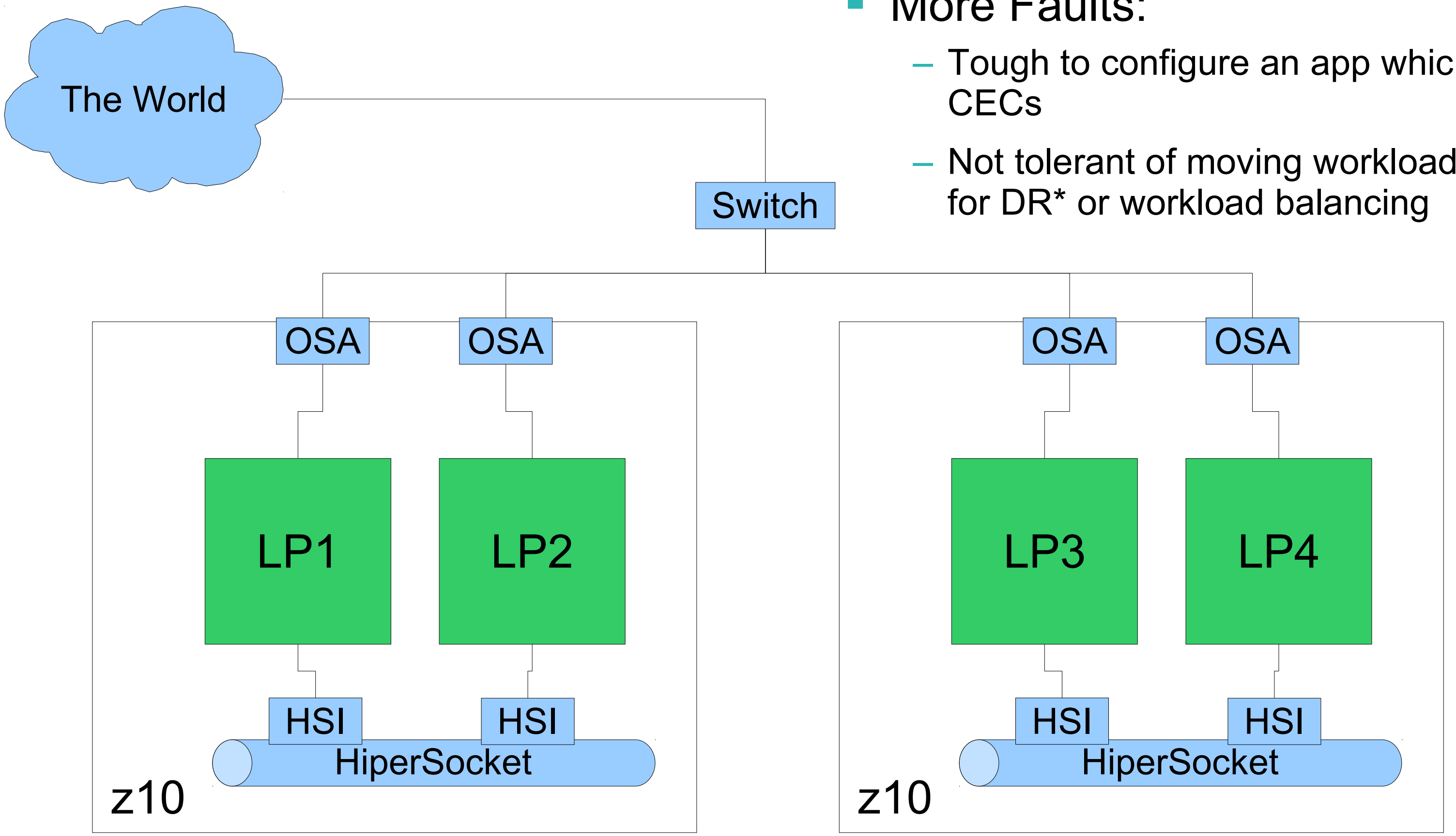
- LP1 becomes a router and forwards packets to and from the HiperSocket Network
- Faults:
  - Pretty expensive for a router: even cheap IFL mips are not really cheap enough to do this
  - LP1 is a Single Point of Failure ( SPoF )
- LP1 is a great place to put a software firewall – so this is still a valid solution if you can solve the SPoF

# The Common Solution:



- Use Naming to choose the interface
  - LP1o and LP2o = OSA side interfaces
  - LP1h and LP2h = HSI side interfaces
  - Both sets of names configured in DNS or hosts
- Manually configure applications to use one name or the other to choose a path
- Faults:
  - Have to pick and choose the correct path for each application in the system
  - Does not handle failures or config errors gracefully

# The Common Solution part 2:

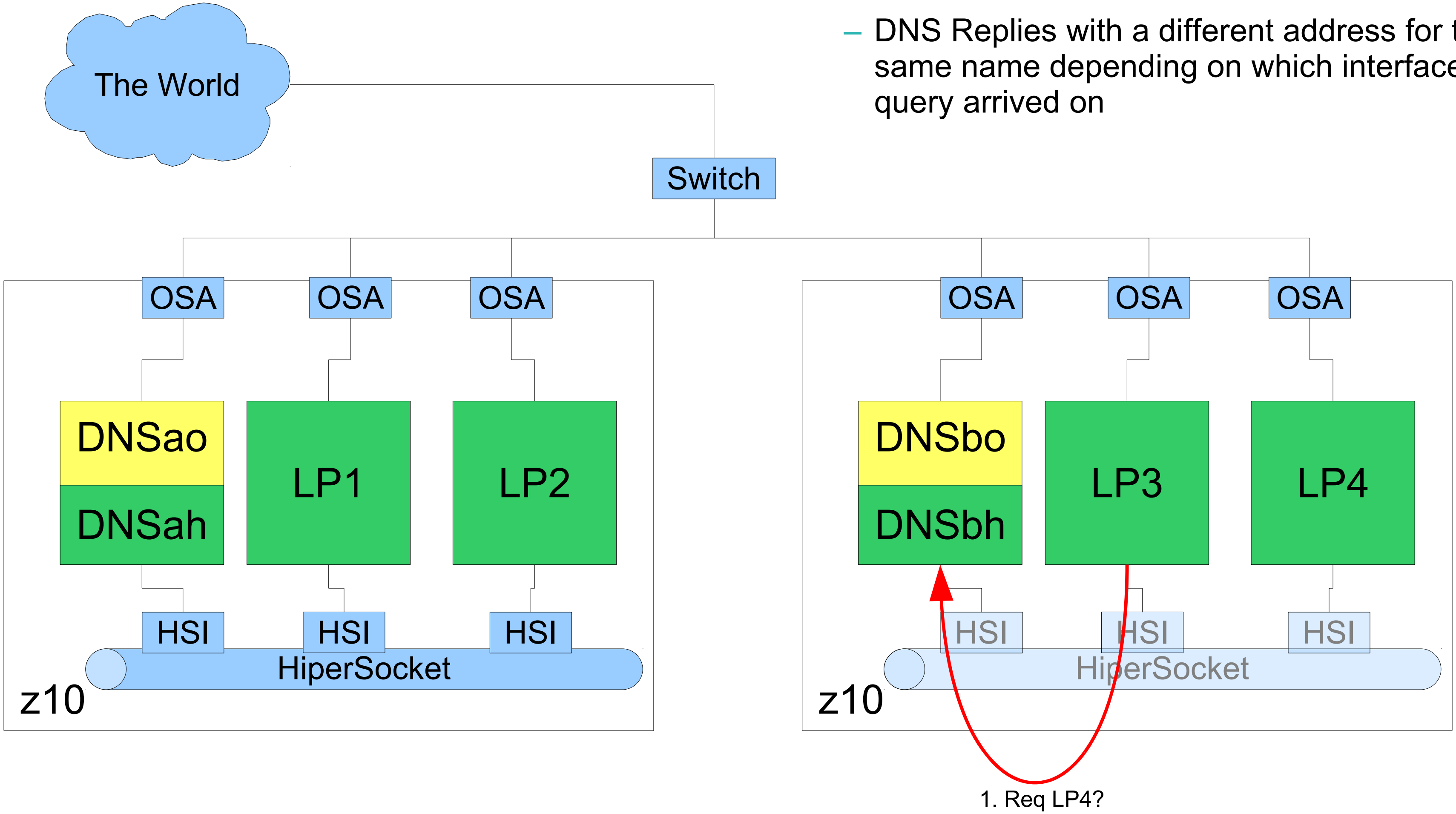


- More Faults:
  - Tough to configure an app which spans multiple CECs
  - Not tolerant of moving workload across CECs for DR* or workload balancing

*Disaster Recovery

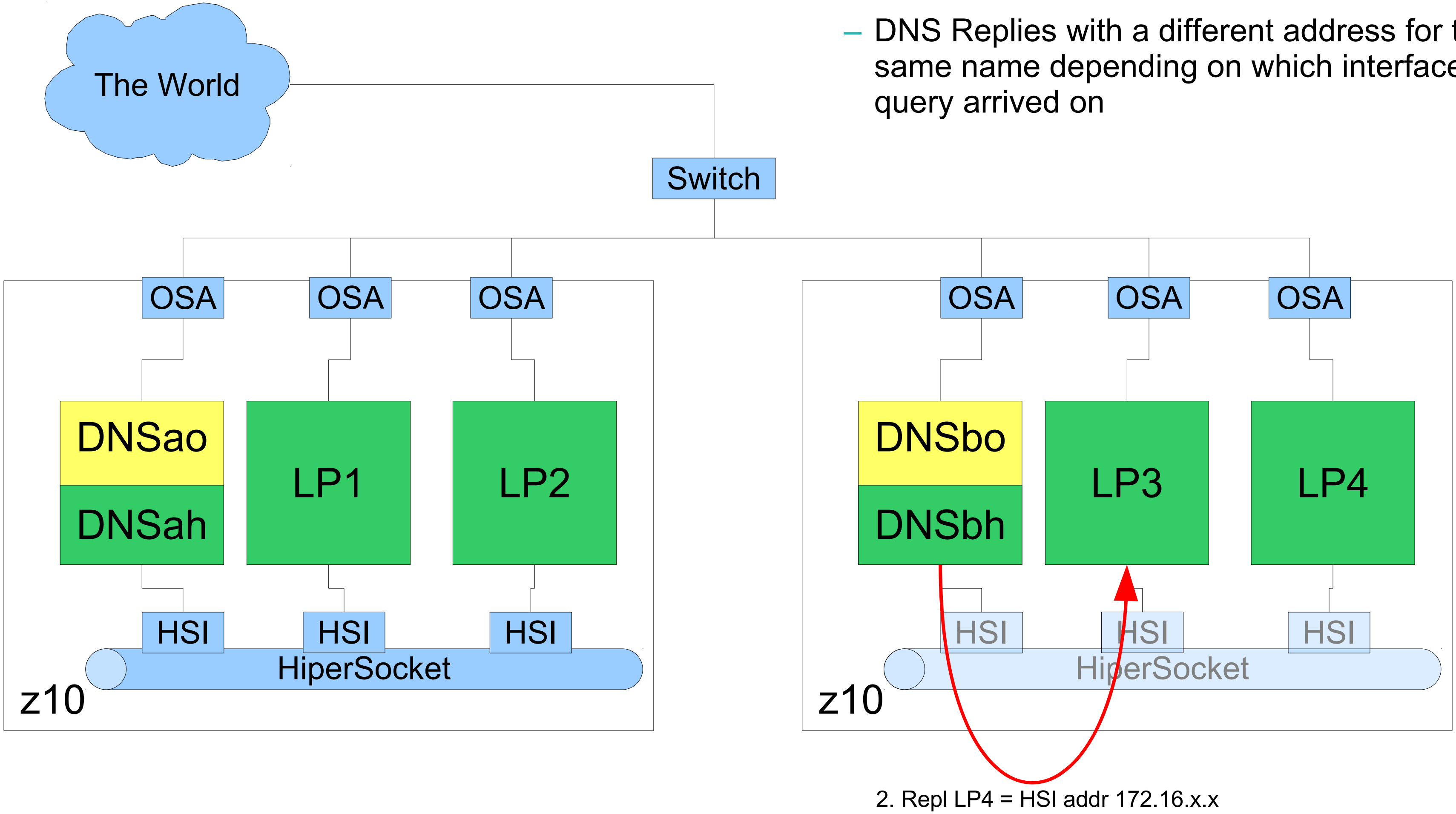
# Split Horizon Solution:

- Split Horizon DNS
  - Single DNS has multiple zones for the same name space
  - DNS Replies with a different address for the same name depending on which interface the query arrived on



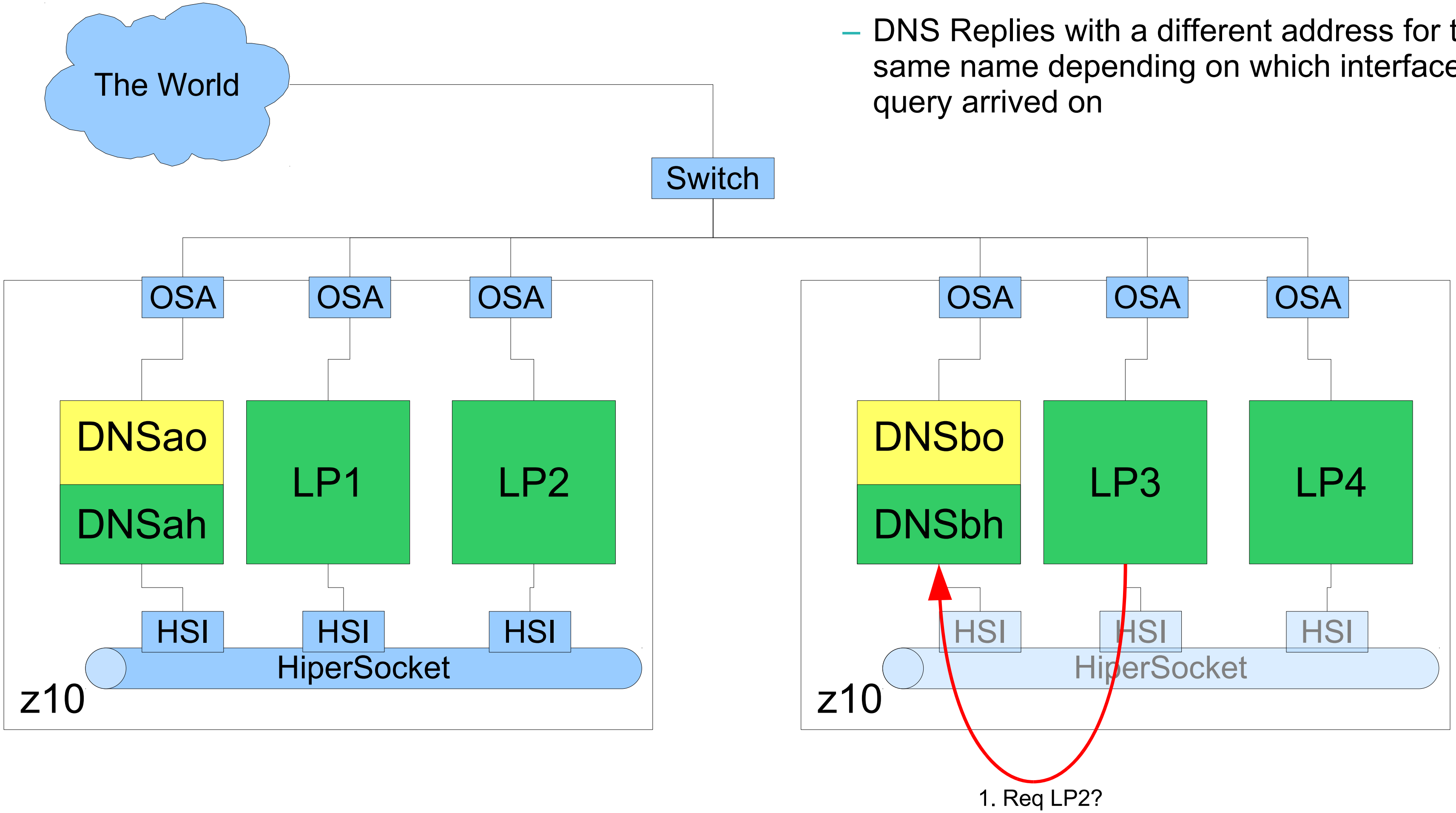
# Split Horizon Solution:

- Split Horizon DNS
  - Single DNS has multiple zones for the same name space
  - DNS Replies with a different address for the same name depending on which interface the query arrived on



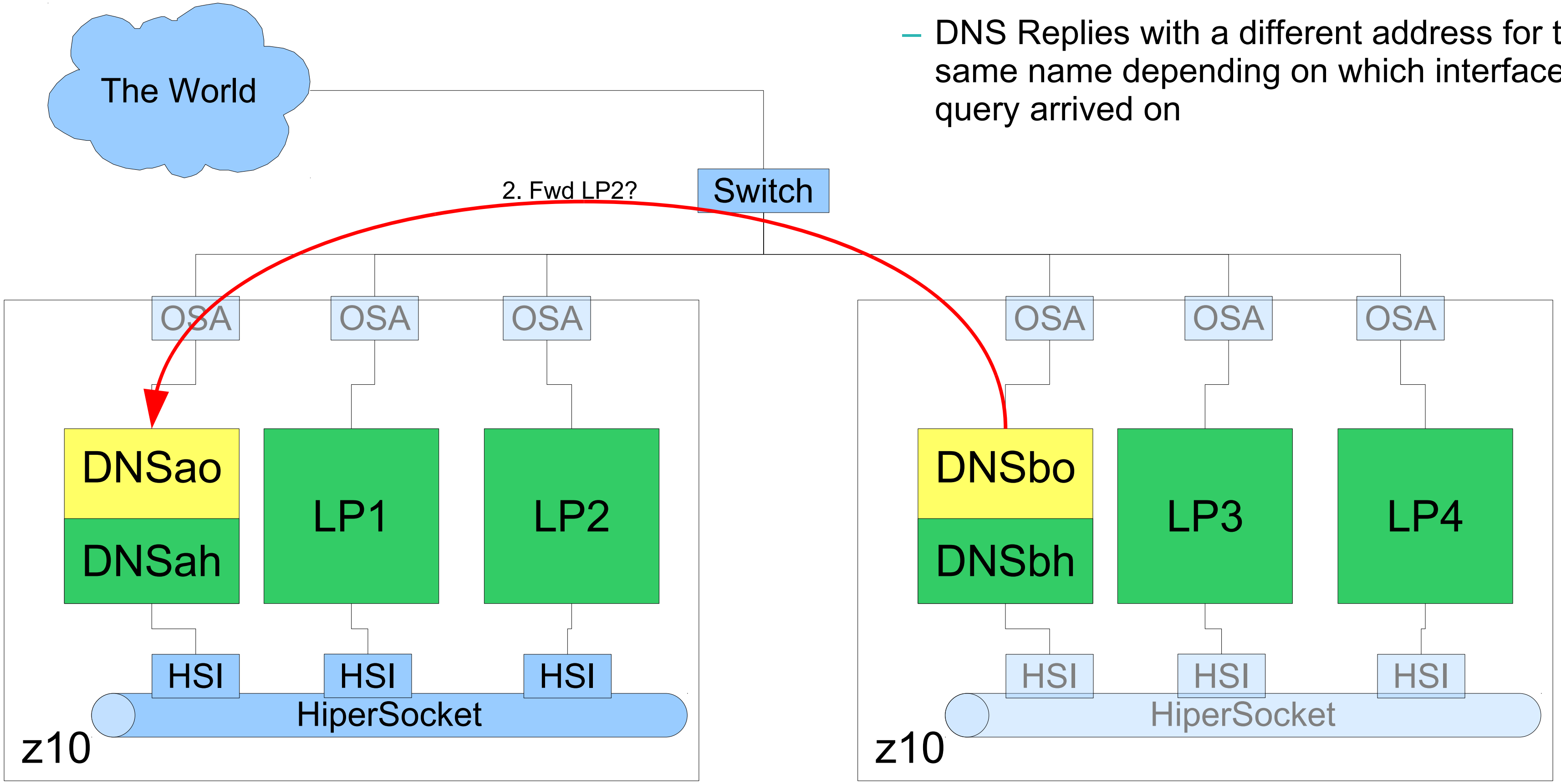
# Split Horizon Solution:

- Split Horizon DNS
  - Single DNS has multiple zones for the same name space
  - DNS Replies with a different address for the same name depending on which interface the query arrived on



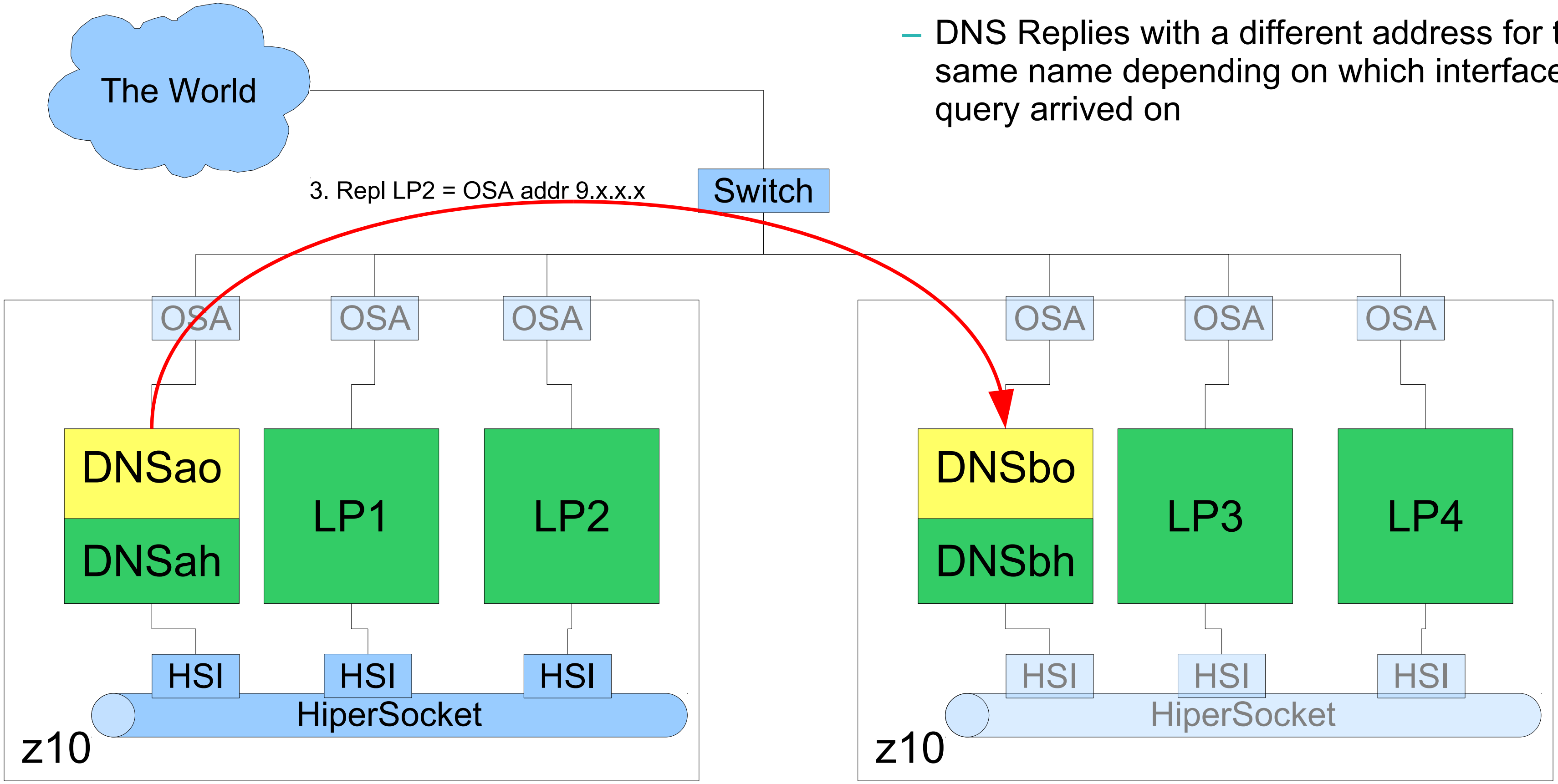
# Split Horizon Solution:

- Split Horizon DNS
  - Single DNS has multiple zones for the same name space
  - DNS Replies with a different address for the same name depending on which interface the query arrived on



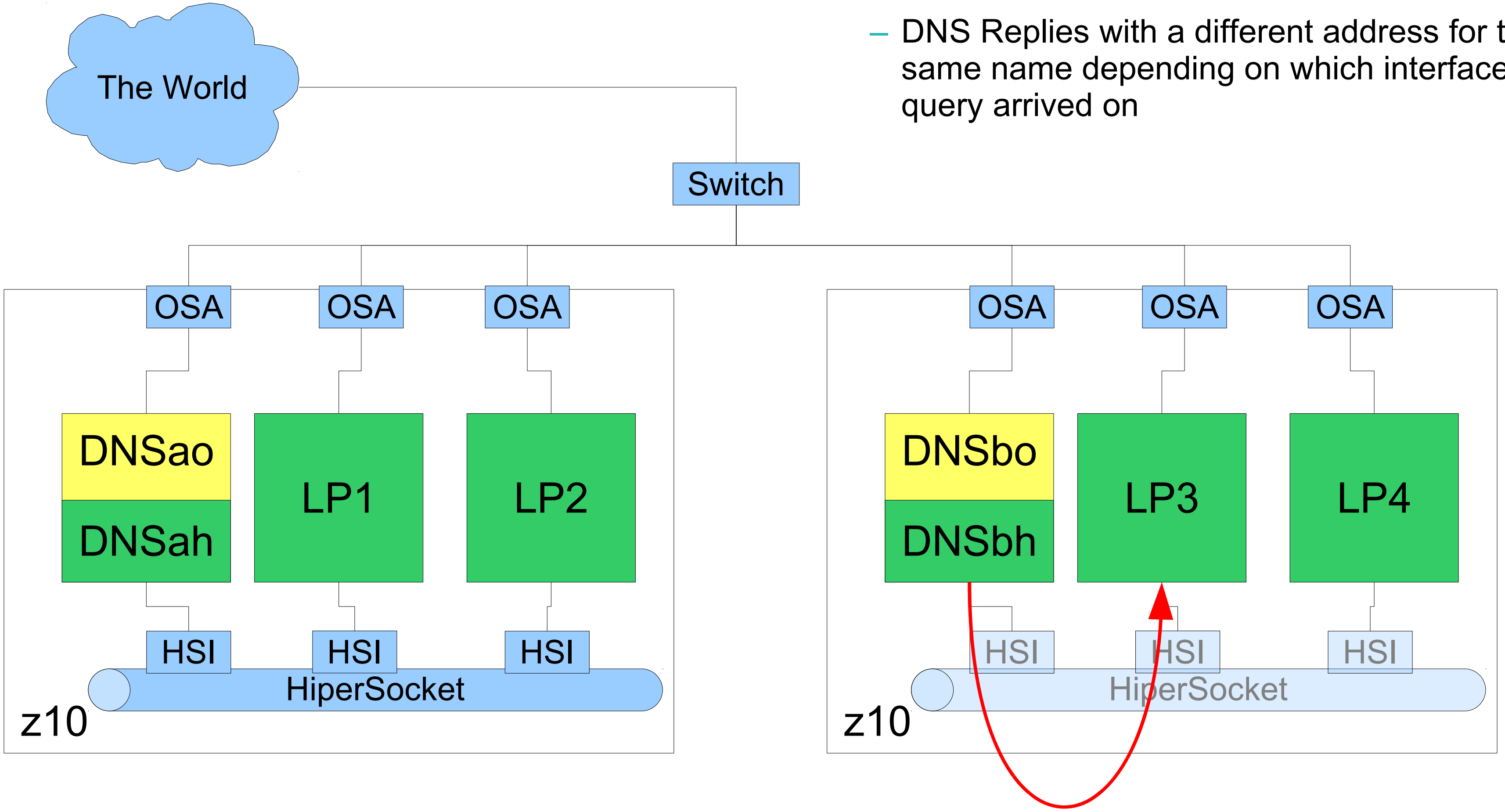
# Split Horizon Solution:

- Split Horizon DNS
  - Single DNS has multiple zones for the same name space
  - DNS Replies with a different address for the same name depending on which interface the query arrived on



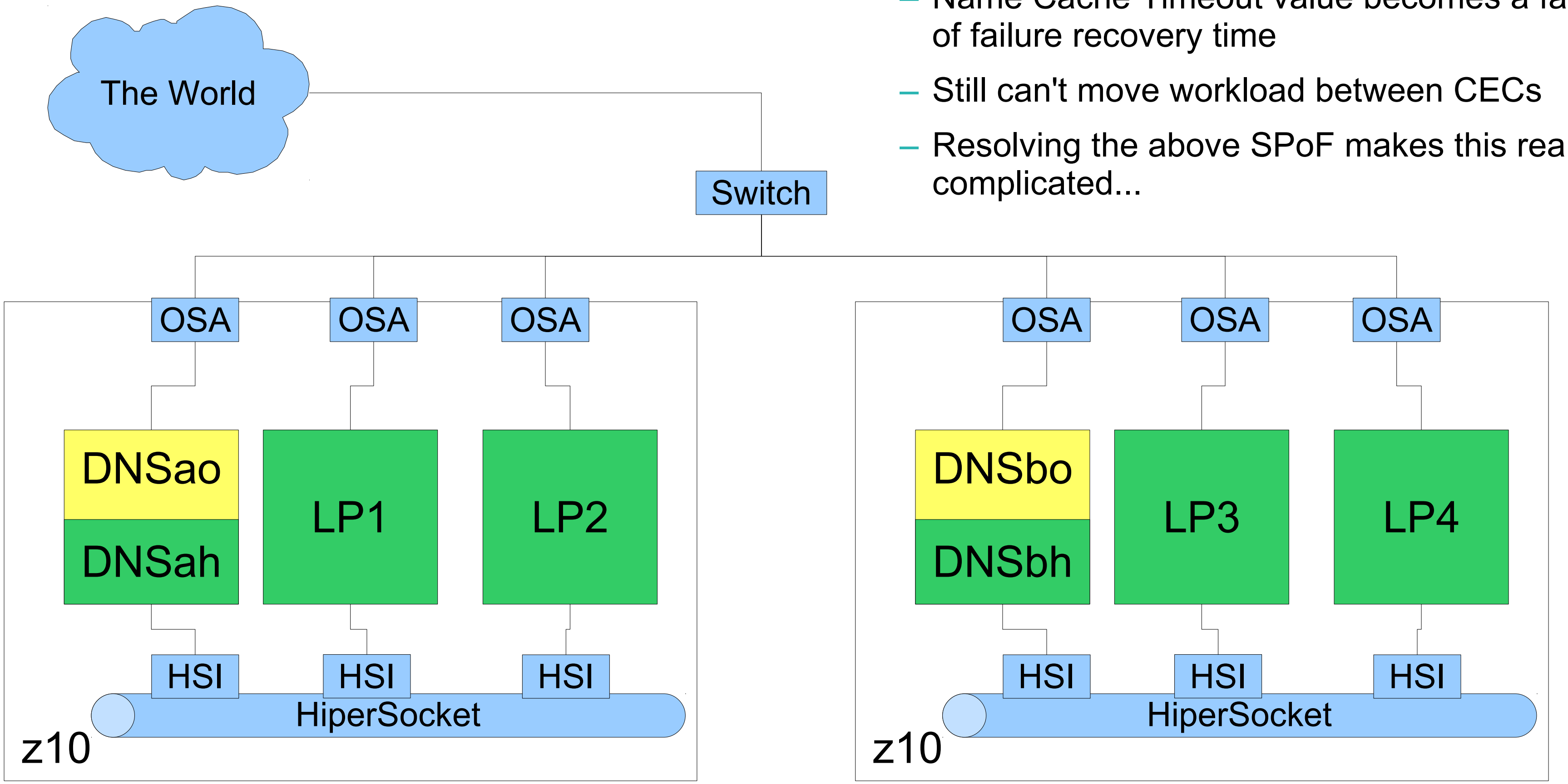
# Split Horizon Solution:

- Split Horizon DNS
  - Single DNS has multiple zones for the same name space
  - DNS Replies with a different address for the same name depending on which interface the query arrived on



4. Repl & Cache LP2 = OSA addr 9.x.x.x

# Split Horizon Solution:



- Faults:
  - The DNS in the CEC is a SPoF
  - Name Cache Timeout value becomes a factor of failure recovery time
  - Still can't move workload between CECs
  - Resolving the above SPoF makes this really complicated...

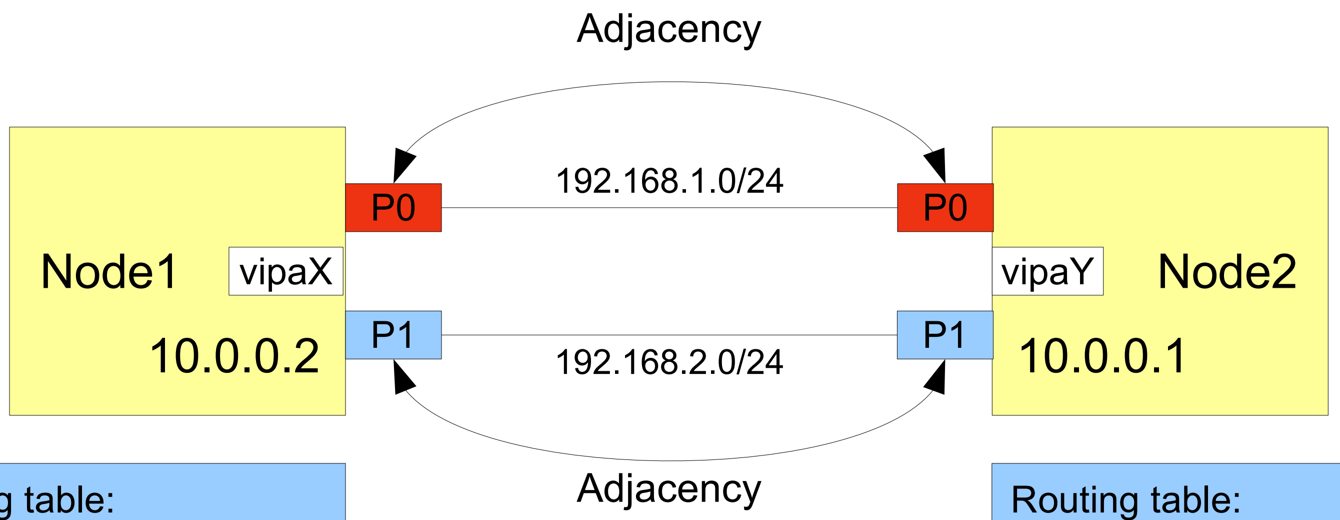
# The Problem ( again )

- When you get right down to it: deciding whether to use the OSA link or the HSI link to talk to a neighbor is a routing decision.
- The LPARs are not themselves routers, since they are not forwarding packets between interfaces.
- There is already a very well designed solution to this problem:

# OSPF: Open Shortest Path First

- Dynamic Routing protocol
- Nodes exchange link state notifications with adjacent nodes to maintain routing tables
  - Node is either an external router, or an LPAR / VM Guest with multiple network interfaces.
- Links have assigned weights to denote link capacity and speed
- Nodes use link state and weights to choose the correct route for packets
- In our case: systems use OSPF to choose the best link to get to the intended destination

# OSPF : getting to know your neighbors

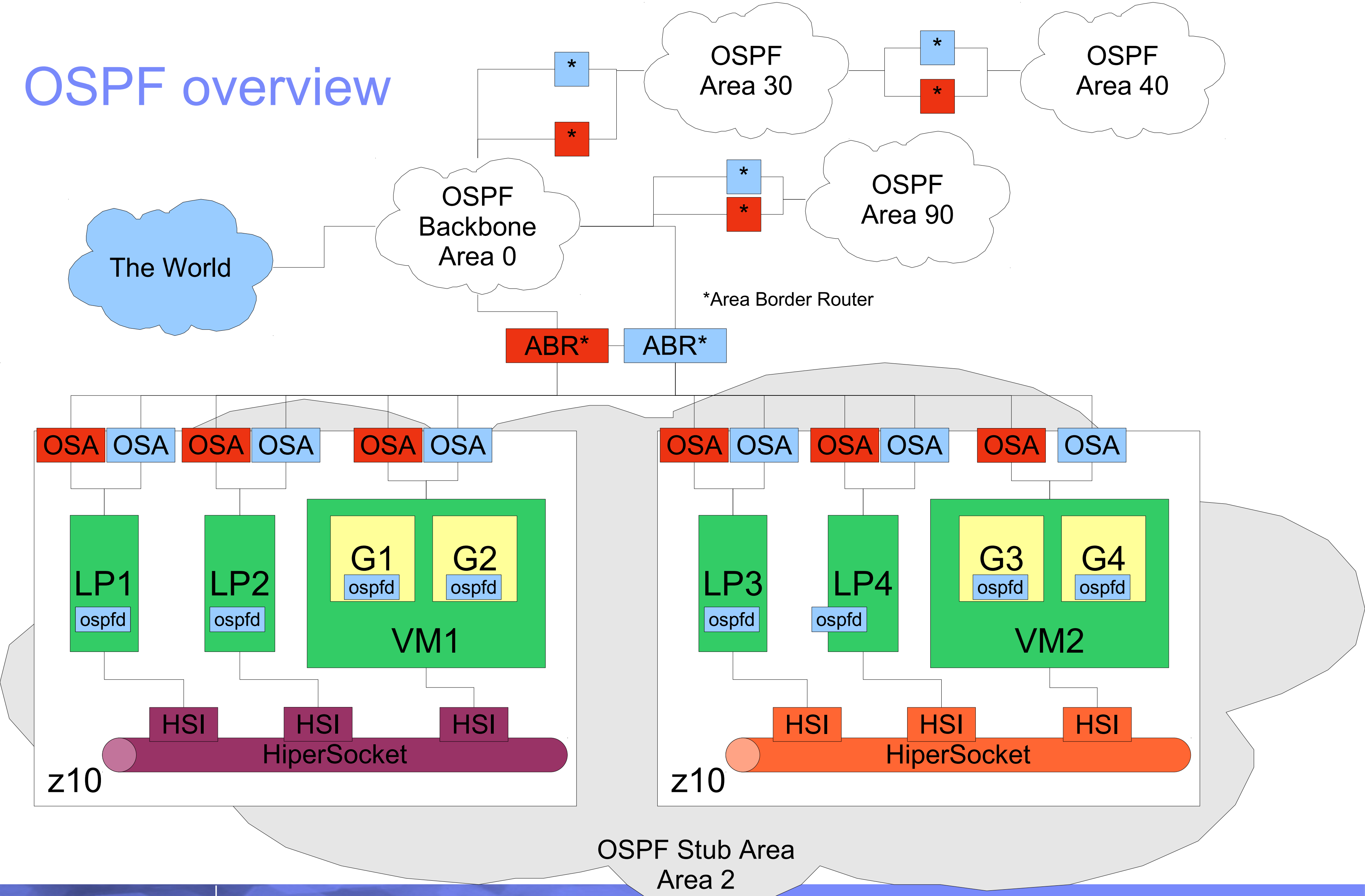


| Routing table: | |
|---|---|
| 192.168.1.0/24: | P0 |
| 192.168.2.0/24: | P1 |
| 10.0.0.1/32: | P0 |

| Routing table: | |
|---|---|
| 192.168.1.0/24: | P0 |
| 192.168.2.0/24: | P1 |
| 10.0.0.2/32: | P0 |

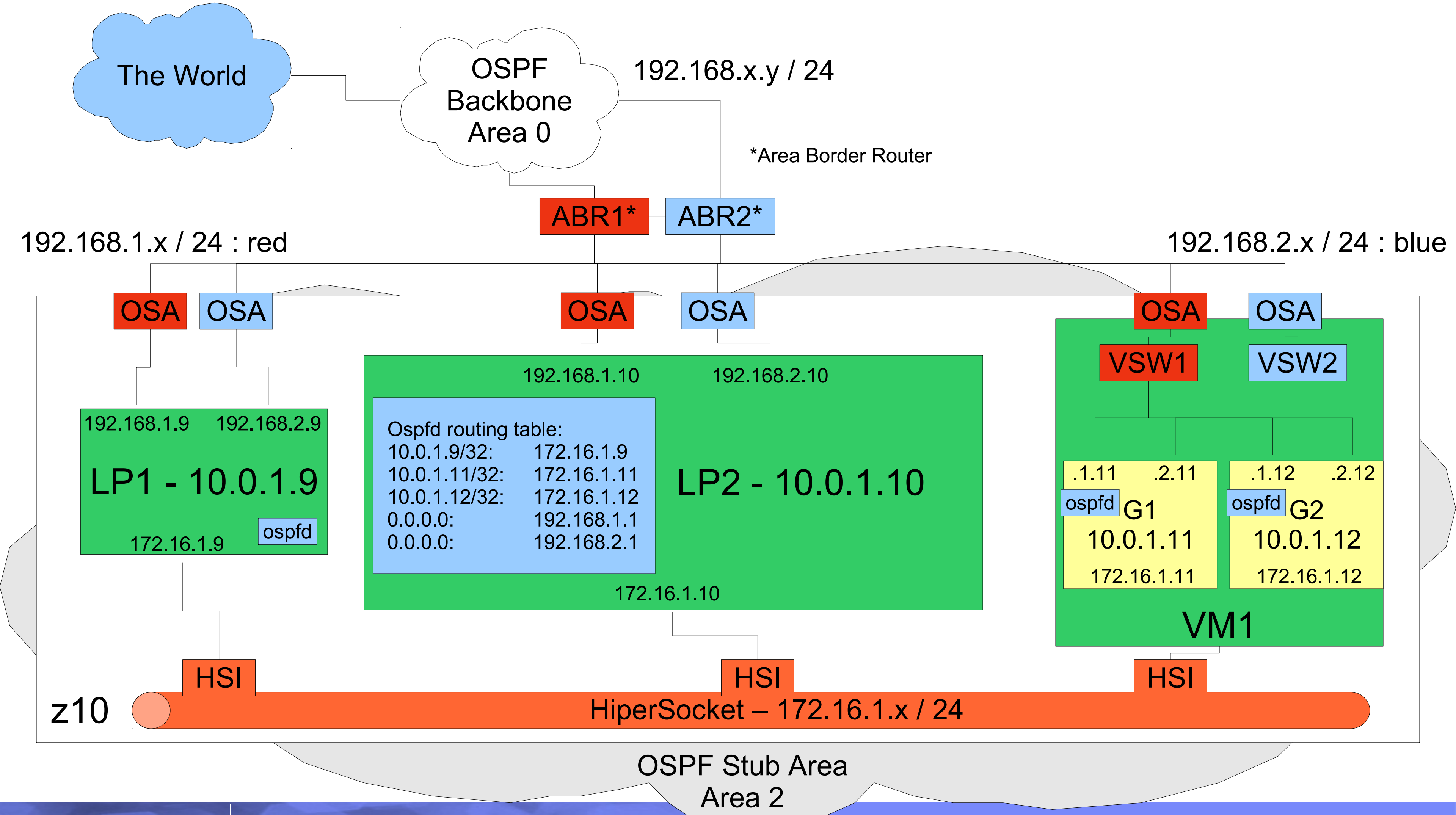
- Form Adjacency with your neighbors
- Advertise all networks you know about to your neighbors
  - Your VIPA!!
- Build routes based on the networks your neighbors advertise to you
- Maintain routes over time

# OSPF overview



10.0.1.x/24 is the VIPA network segment
Equal Cost MultiPathing
also possible with this configuration

# OSPF: zooming in



Ospfd routing table:

| | |
|---|---|
| 10.0.1.9/32: | 172.16.1.9 |
| 10.0.1.11/32: | 172.16.1.11 |
| 10.0.1.12/32: | 172.16.1.12 |
| 0.0.0.0: | 192.168.1.1 |
| 0.0.0.0: | 192.168.2.1 |

# OSPF on Linux

- Our testing used the Quagga package
  - A fork of the zebra package

- Consists of the following components
  - Zebra daemon *
  - OSPF v2 daemon *
  - OSPF v3 ( IPv6 ) daemon
  - Rip daemon
  - RipNG ( IPv6 ) daemon
  - BGP daemon

- All Quagga components have an internal telnet server for interactive configuration and problem diagnosis

*these components are used in this set of examples

# OSPF Config details for Linux

- zebra.conf:

```
! Static VIPA
interface dummy0
 ip address 10.0.1.10/32
 ipv6 nd suppress-ra
!
interface eth1
 ip address 192.168.71.10/24
 ipv6 nd suppress-ra
!
! Hipersocket - 40K packet, 32K MTU
interface hsi1
 ip address 172.16.1.10/16
 ipv6 nd suppress-ra
!
interface lo
!
interface sit0
 ipv6 nd suppress-ra
!
ip forwarding
!
line vty
 exec-timeout 0 0
!
```

- ospfd.conf:

```
! Server - Static VIPA
interface dummy0
 ip ospf cost 1
 ip ospf priority 0
!
interface eth1
 ip ospf cost 10
 ip ospf priority 0
!
interface hsi1
 ip ospf cost 1
 ip ospf priority 10
!
interface lo
!
interface sit0
!
router ospf
 ospf router-id 172.16.1.10
 network 172.16.0.0/16 area 2.2.2.2
 network 172.31.0.34/32 area 2.2.2.2
 network 172.31.200.1/24 area 2.2.2.2
 network 192.168.71.0/24 area 2.2.2.2
 area 2.2.2.2 stub
!
line vty
 exec-timeout 0 0
!
```

# OSPF Test Results

- As expected: no matter which interface was disabled, traffic was able to route around the dud link
- Routes re-converged quickly no matter whether the OSA or HSI side was disabled
- When OSA side links are disabled OSPF enables an eligible OS image on the CEC with a functional OSA link to become a router as in the Original Solution
- Works as Advertised!

# Performance Implications

- Surprisingly – not much
- Tested 66 VM guests in the same OSPF area running on a single VM system
  - Combined CPU Utilization of the zebra & ospfd daemons was less than 1% during normal operations
  - CPU spikes up to 1.5% were noted during re-convergence after a path failure
  - Layer2 networking seems to keep VM guests in queue more so than layer3, which may contribute to the negligible overhead
  - Defining the area containing the Z systems as a Stub Area is critical to minimizing the overhead of running OSPF
    - Using a Completely Stubby Area lowers the overhead even more if your networking configuration supports it.

# OSPF Faults

- Overall Complexity
  - It's not just a single default route anymore

- More customization to be done at each node during provisioning
  - But it can be handled with some creative “sed -i /old/new/” type scripting

- … I can't think of anything else to put on this slide ...

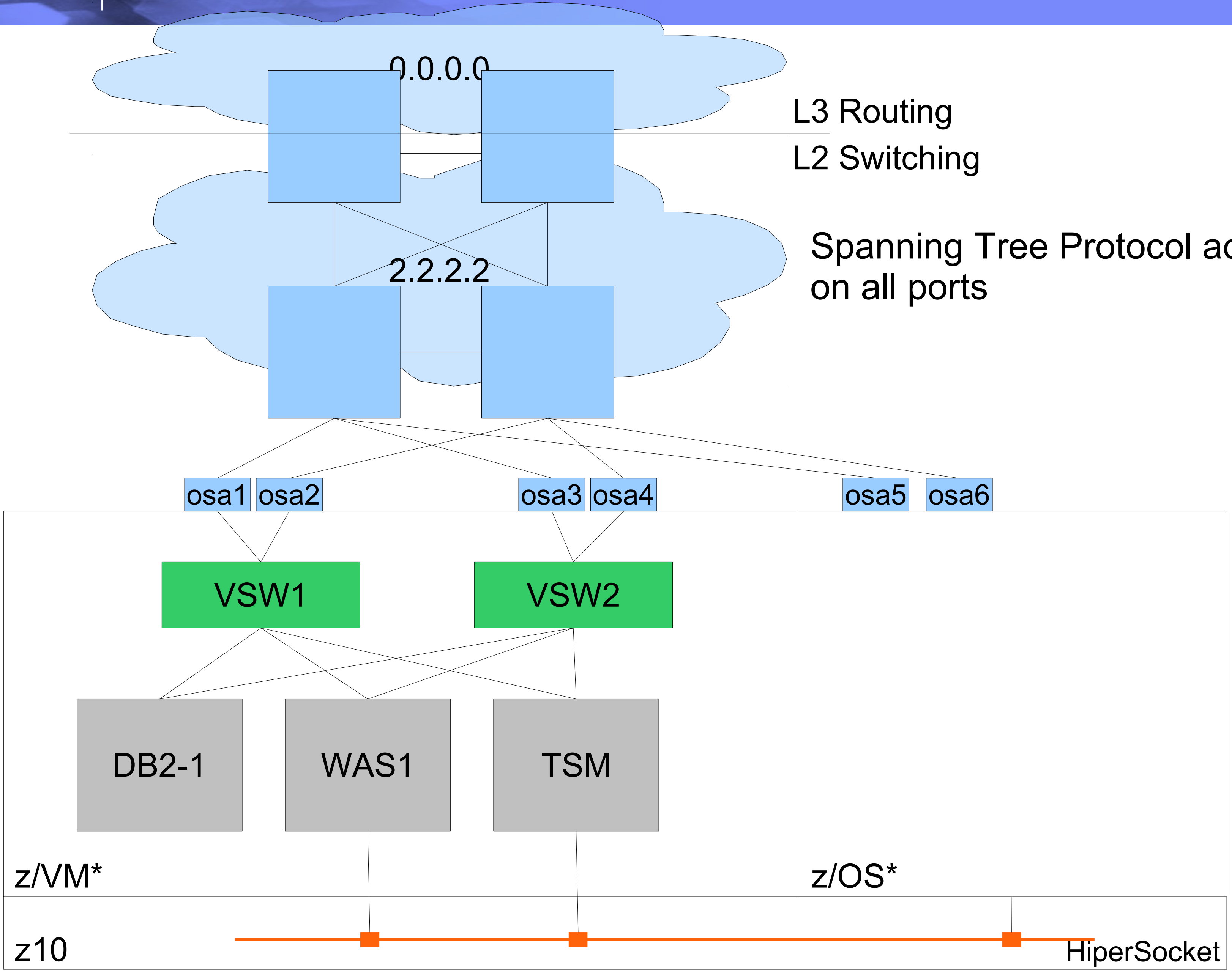
# For more information

- System Z Platform Test library:
  - http://www-03.ibm.com/systems/services/platformtest/servers/systemz_library.html
- The OSPF paper this presentation is based on:
  - http://www-03.ibm.com/systems/resources/linux_ha_ospf.pdf

This slide intentionally left blank

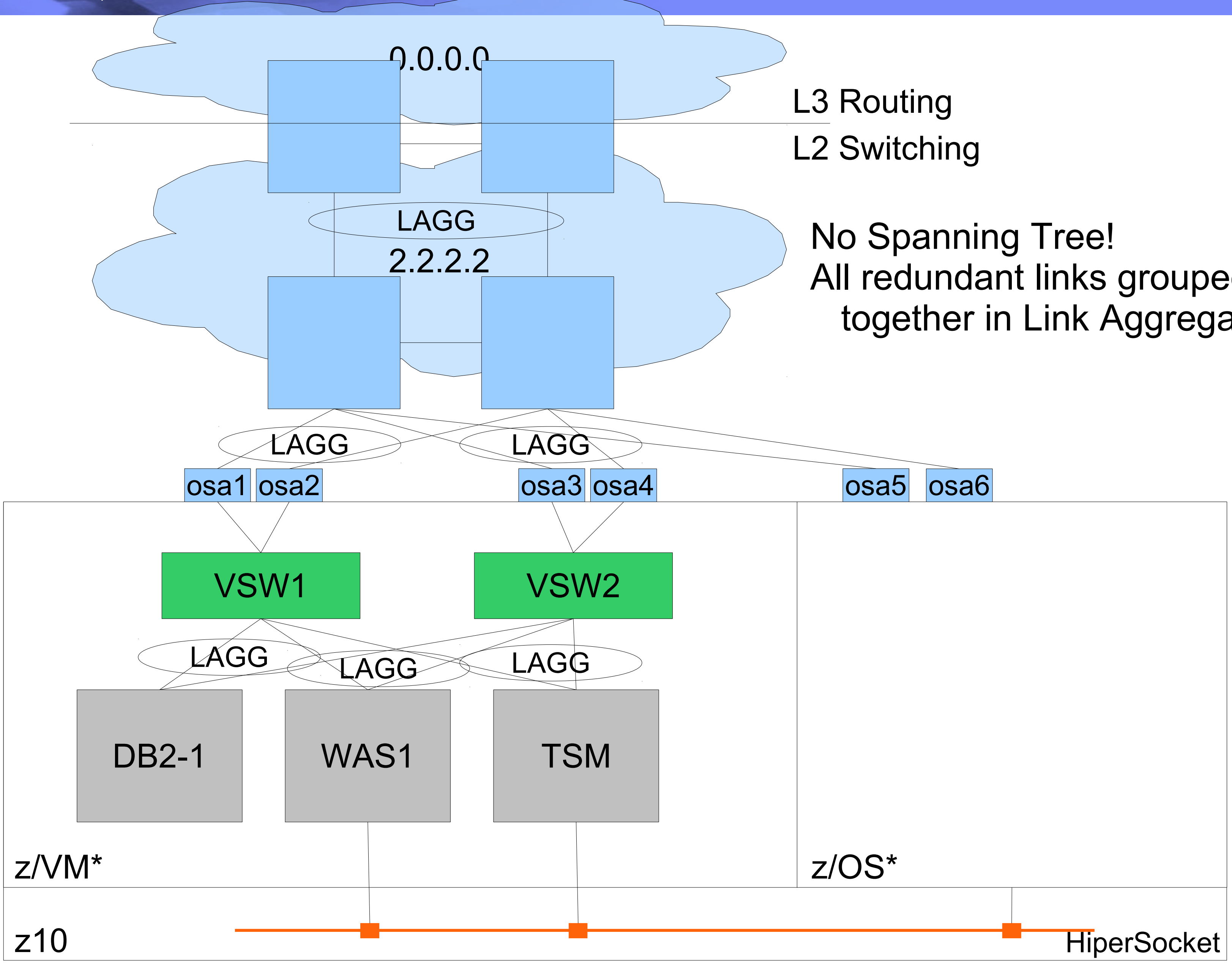
# Old vs New ?

- Does Link Aggregation make OSPF solutions obsolete?



Spanning Tree Protocol active on all ports

*Not to scale



L3 Routing

L2 Switching

No Spanning Tree!
All redundant links grouped
together in Link Aggregation

*Not to scale