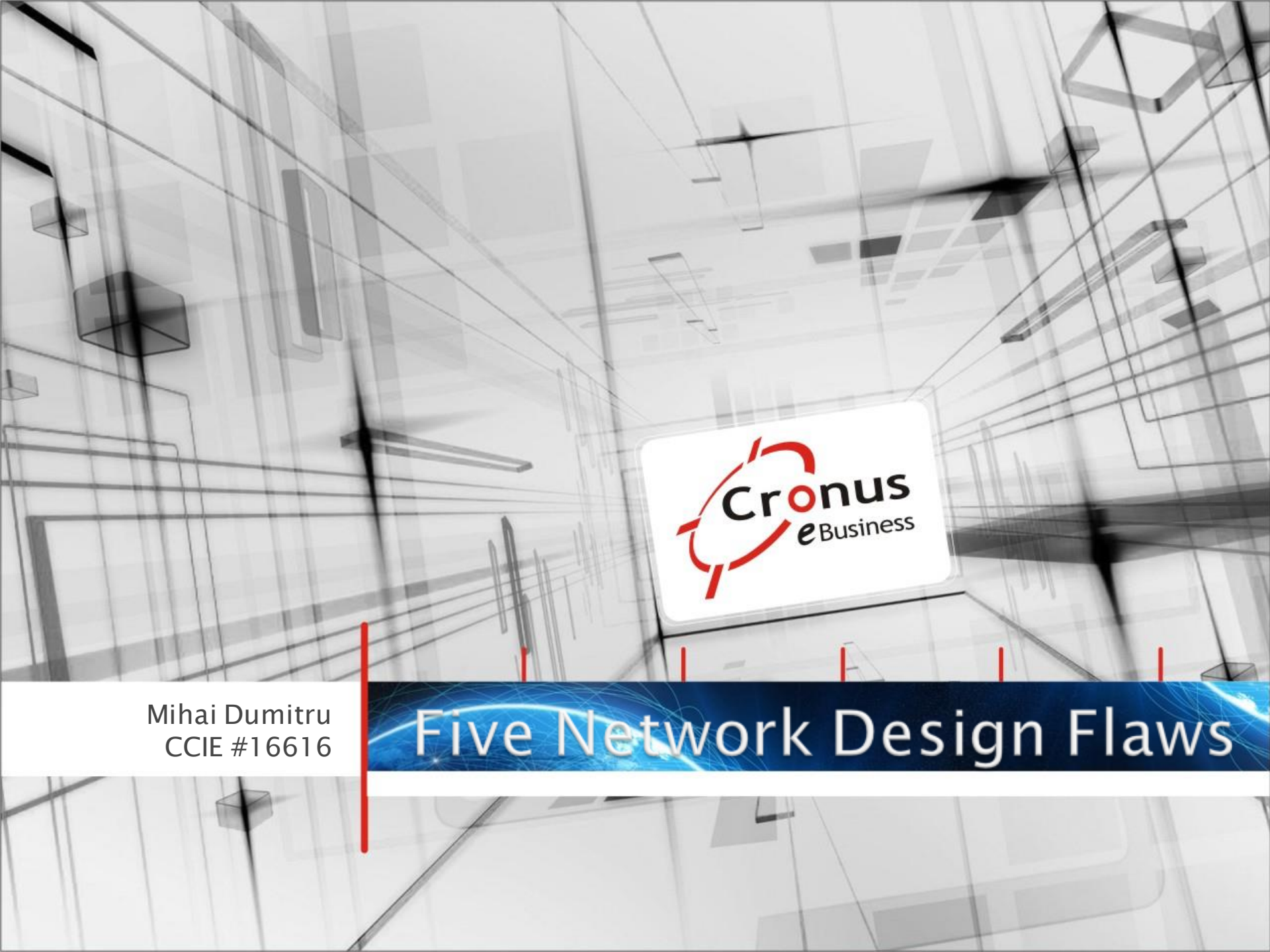


Mihai Dumitru
CCIE #16616

Five Network Design Flaws

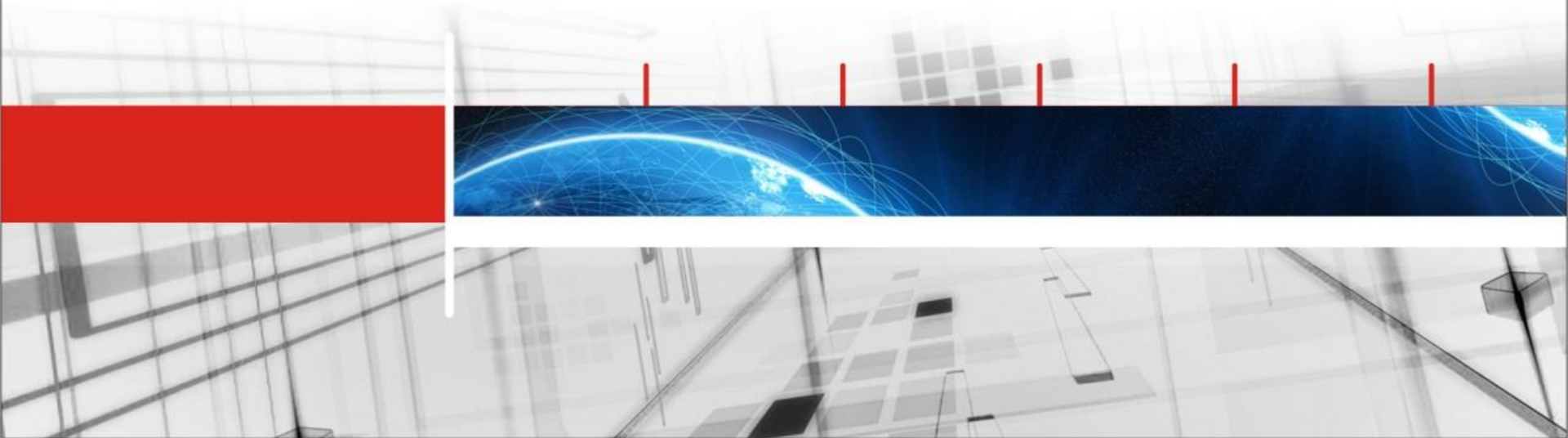
- Design flaws are design decisions that may produce unwanted results in the future.
- Usually, they cannot be fixed by just entering a few commands at the CLI.
- We will take a look at five examples.

- All designs and events appearing in this work are fictitious. Any resemblance to real designs, living or dead, is purely coincidental.

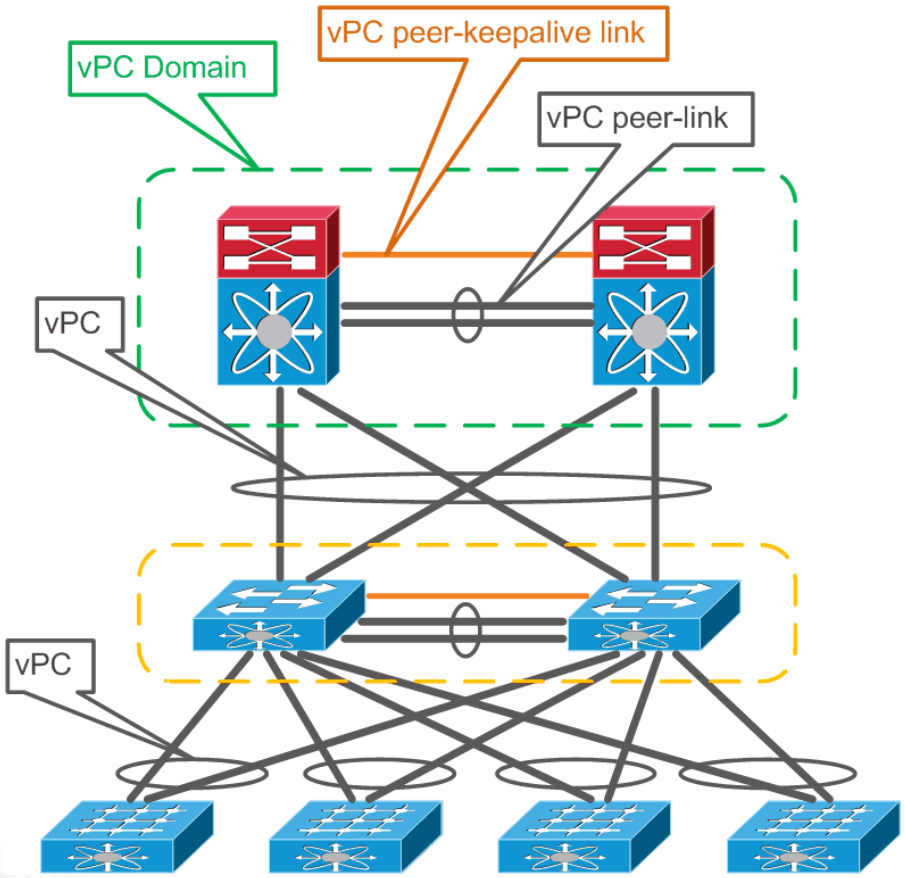
Cronus eBusiness

Take 1

Nexus Virtual Port-Channel



Nexus vPC

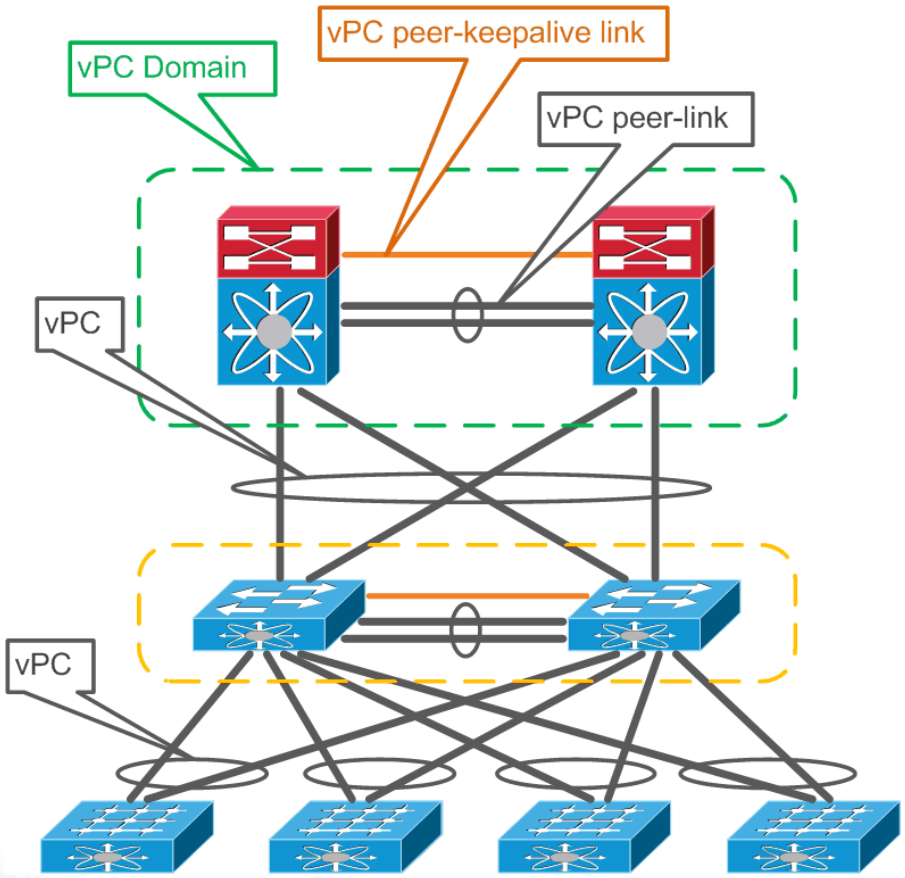


- Eliminates STP blocked ports
- Uses all available uplink bandwidth
- Provides fast convergence

- Terminology:
  - vPC Domain — a pair of vPC switches
  - vPC — the combined port channel between the vPC peers and one downstream device
  - vPC peer-link — link used to synchronize state between peers, carry STP, HSRP, IGMP and flooded traffic – must be 10GbE

Cronus eBusiness

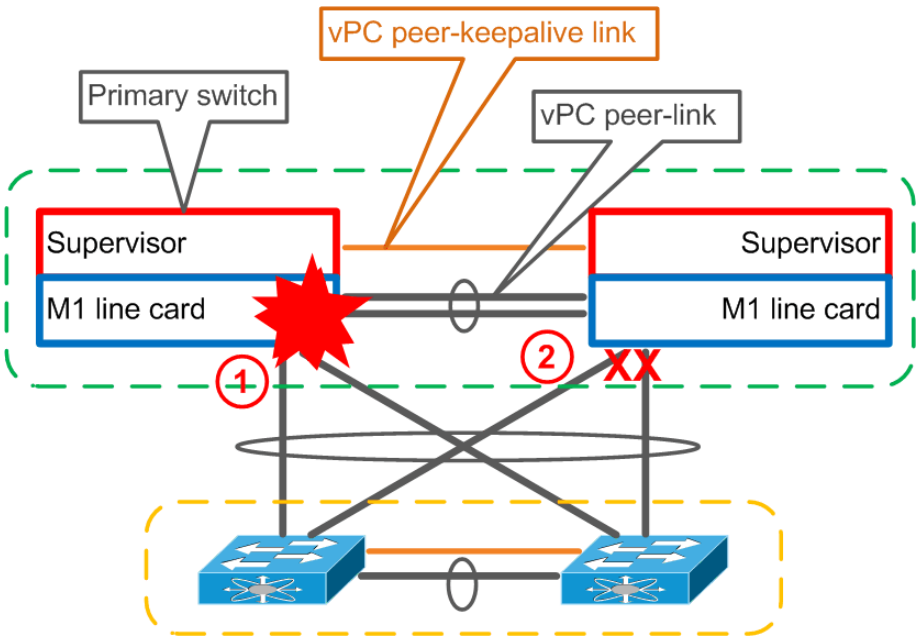
Nexus vPC



- Terminology:
  - vPC peer-keepalive link — link used for active-active detection and role election (nothing happens if this link goes down temporarily)
- One vPC peer switch is elected as the vPC primary switch, based on the configured role priority.
- How can this reference design go wrong?

Cronus eBusiness

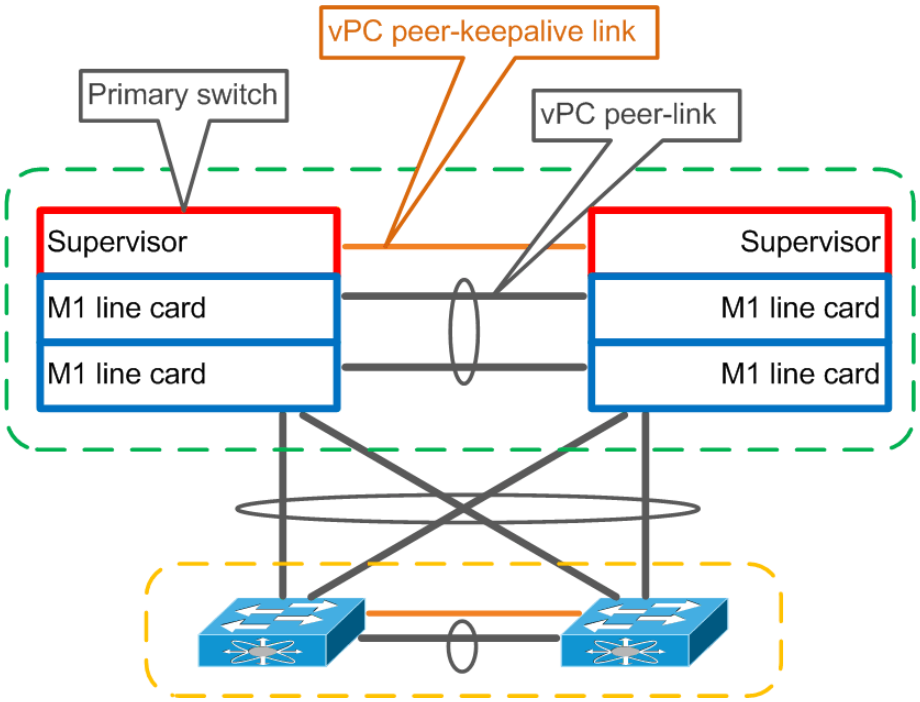
Nexus vPC



- The M1 line card on the vPC primary switch failed; the vPC peer link went down.
- As a result, the vPC secondary switch shuts down vPC member ports, *if it can still receive heartbeat messages from the vPC primary switch*.
- This is by design, in order to avoid a possible STP loop.

Cronus eBusiness

Nexus vPC

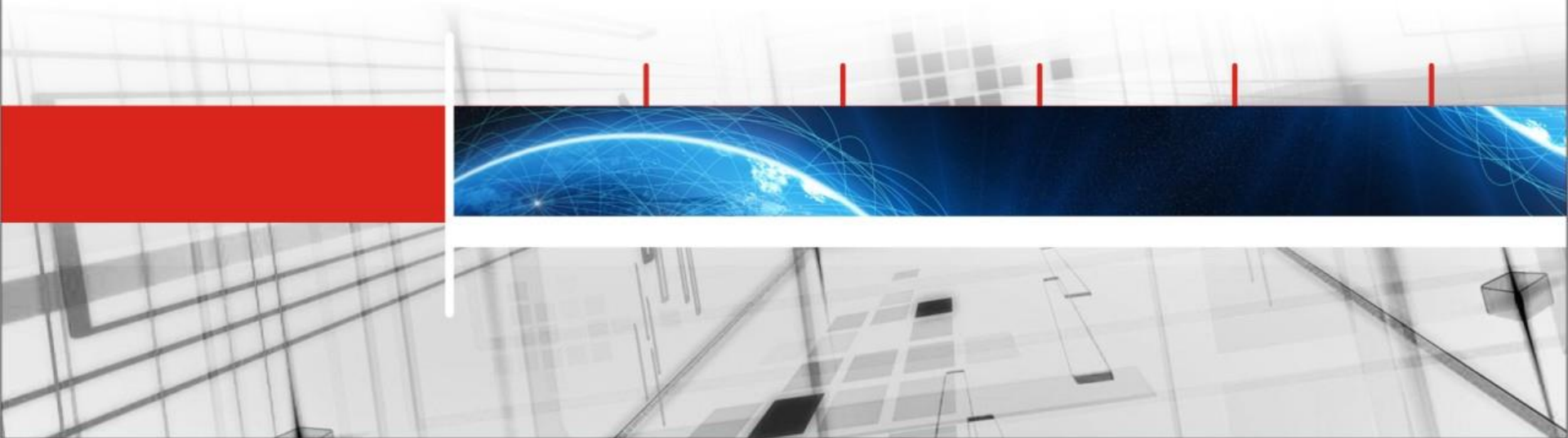


- When building a vPC peer link, use at least two ports *and* at least two line cards of the same type.
- What if the vPC peer-keepalive link was on the M1 line card?

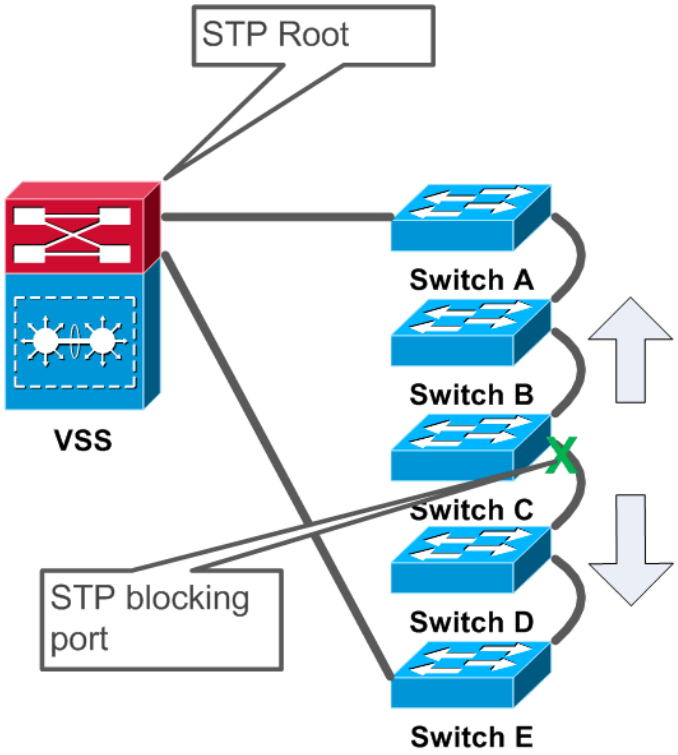
Cronus eBusiness

Take 2

Daisy Chain to StackWise



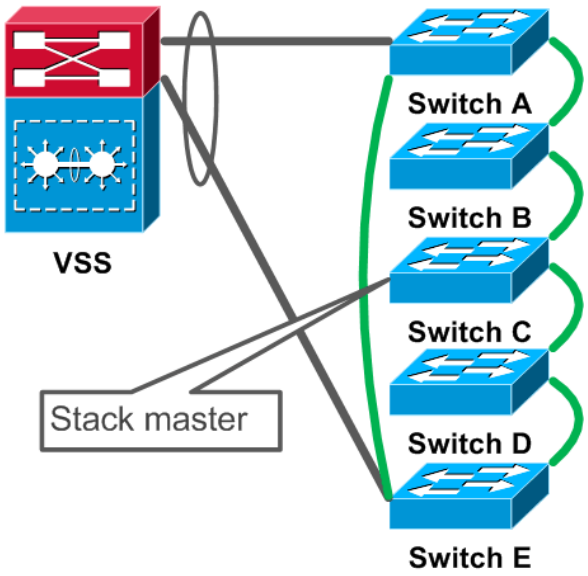
Daisy Chain to StackWise



- Uplink failure would temporarily isolate half of the switches.
- STP convergence time was in tens of seconds.

Cronus eBusiness

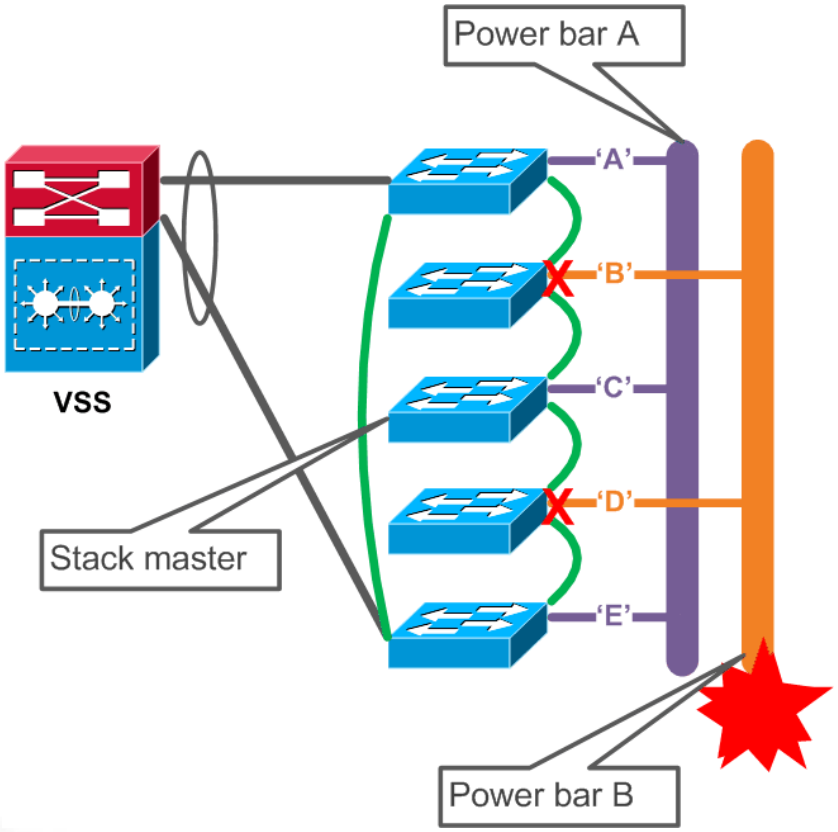
Daisy Chain to StackWise



- The old access switches have been replaced with a stack.
- No more STP loop
- Convergence time on uplink failure fell under 300msec, given that the uplinks were not on the stack master.
- A stack master and a slave can be predefined. But…

Cronus eBusiness

Daisy Chain to StackWise

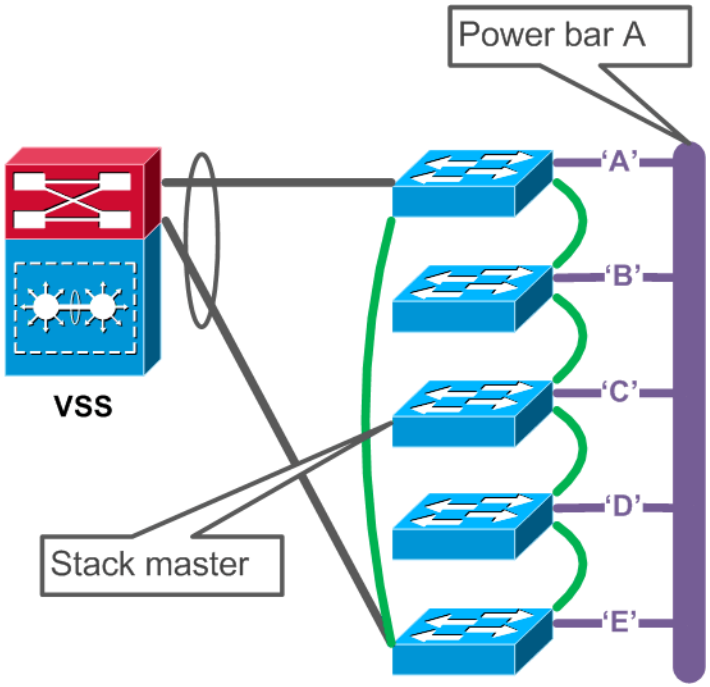


- Switch B and switch D were on power bar B.
- Power bar B failed.
- A and E reloaded (to elect master).
- Two active stack "islands" appeared: A plus E, and C.

- After power came back, the two "islands" attempt to merge: A, E and C reload, then B and D *continue to boot* as isolated masters.

Cronus eBusiness

Daisy Chain to StackWise



<Fri Mar 6 14:20:07 2015> Message from sysmgr: Reason Code:[4] Reset Reason:Reset/Reload requested by [stack-manager]. [stack merge]

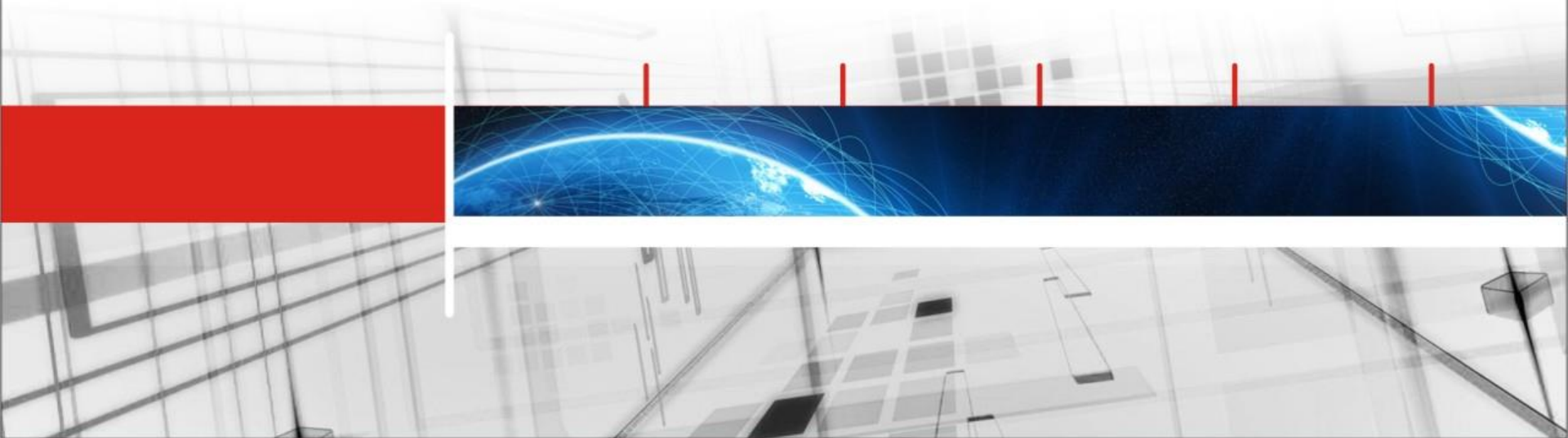
Switch rebooted continously for atleast 5 times. **Disabling autoreboot**

- Make sure you have consistent/ redundant power sources.
- Or limit the stack size to three switches.

Cronus eBusiness

Take 3

Network Load Balancing



NLB

- RTFM:
- "In its default unicast mode of operation, Network Load Balancing reassigns the station MAC address of the network adapter for which it is enabled, and *all cluster hosts are assigned the same MAC address*.
- Incoming packets are thereby received by all cluster hosts and passed up to the Network Load Balancing driver for filtering.“

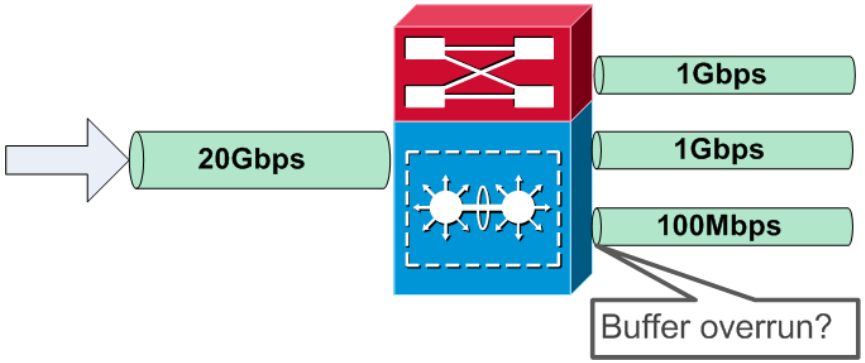
Cronus eBusiness

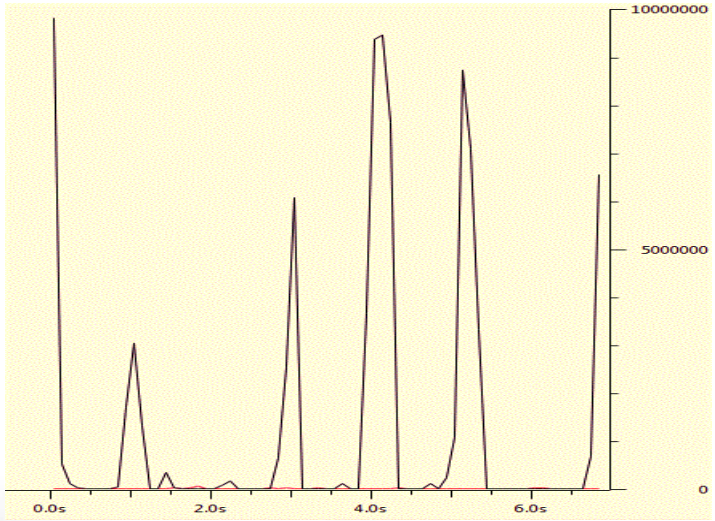
NLB

- "The use of a common MAC address would create a conflict since layer-two switches expect to see unique source MAC addresses on all switch ports.
- To avoid this problem, *Network Load Balancing uniquely modifies the source MAC address for outgoing packets*; a cluster MAC address of 02-BF-1-2-3-4 is set to 02-h-1-2-3-4, where h is the host's priority within the cluster.
- This technique prevents the switch from learning the cluster's actual MAC address, and as a result, *incoming packets for the cluster are delivered to all switch ports*."

Cronus eBusiness

NLB

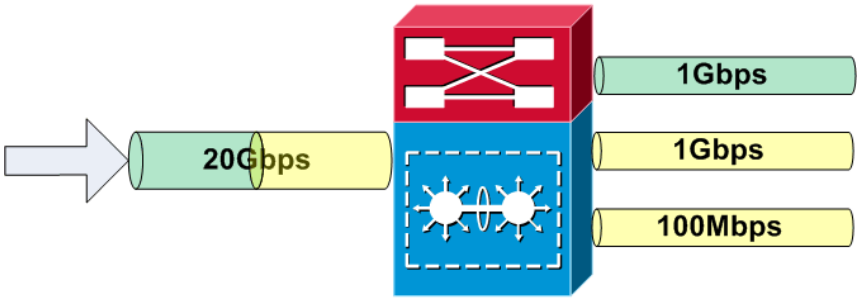




- What if the net admin had absolutely no idea what the server guy was doing?
- He had installed the NLB cluster in the same server VLAN… The same VLAN that the iLO interfaces belonged to…
- Intermittent loss, intermittent delays, RDP freezes sometimes, iLO not responding…

Cronus eBusiness

NLB

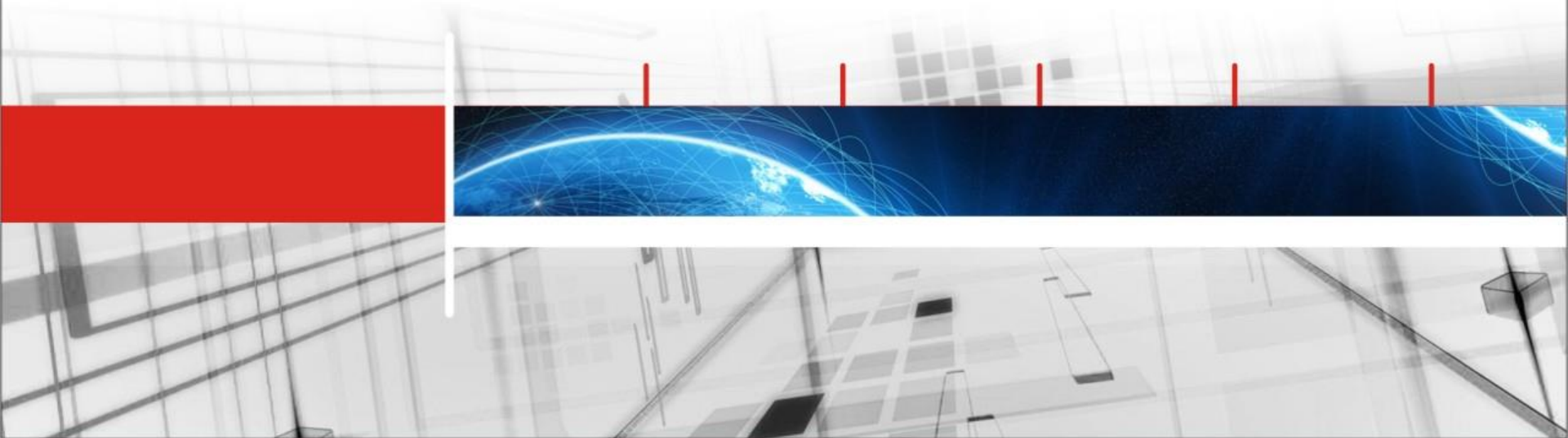


- This is a micro-segmentation issue.
- The NLB should have been installed in its own new VLAN.
- The same for the iLO interfaces.

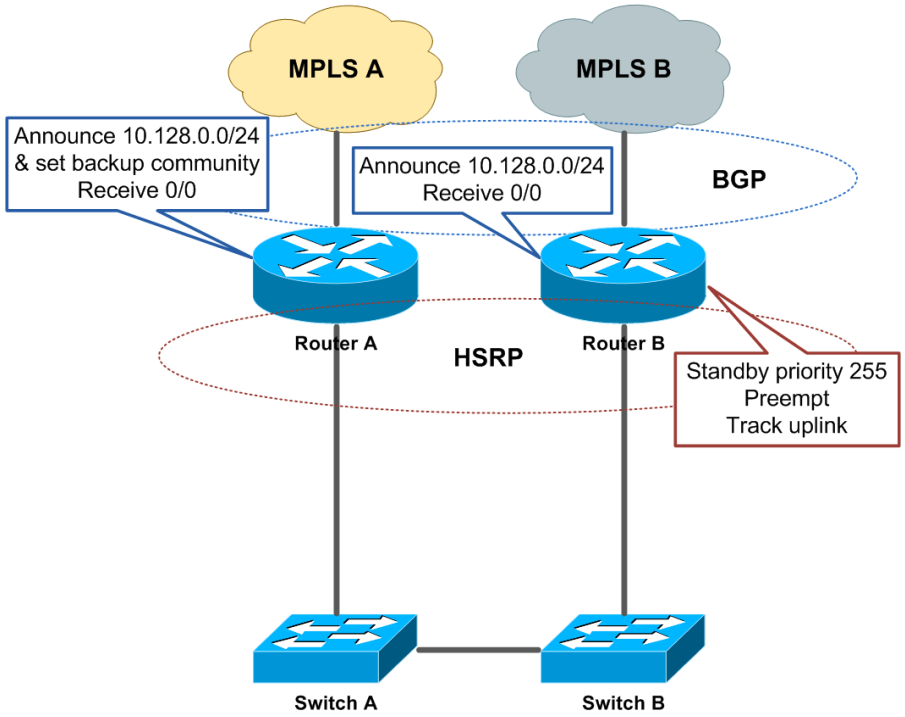
Cronus eBusiness

Take 4

Inline Appliance



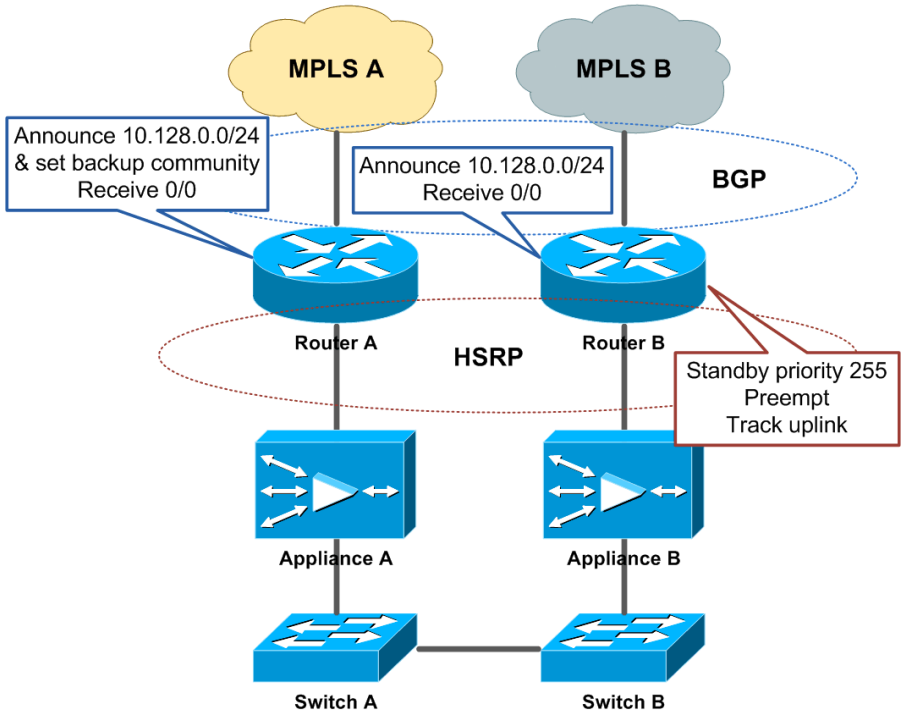
Inline Appliance



- If uplink B failed, then router A becomes primary gateway. Beware of the track delay and BGP hold-time.
- If router B failed, it's easy.
- If switch B failed, then router B withdraws the BGP announcement.
- So far so good.

Cronus eBusiness

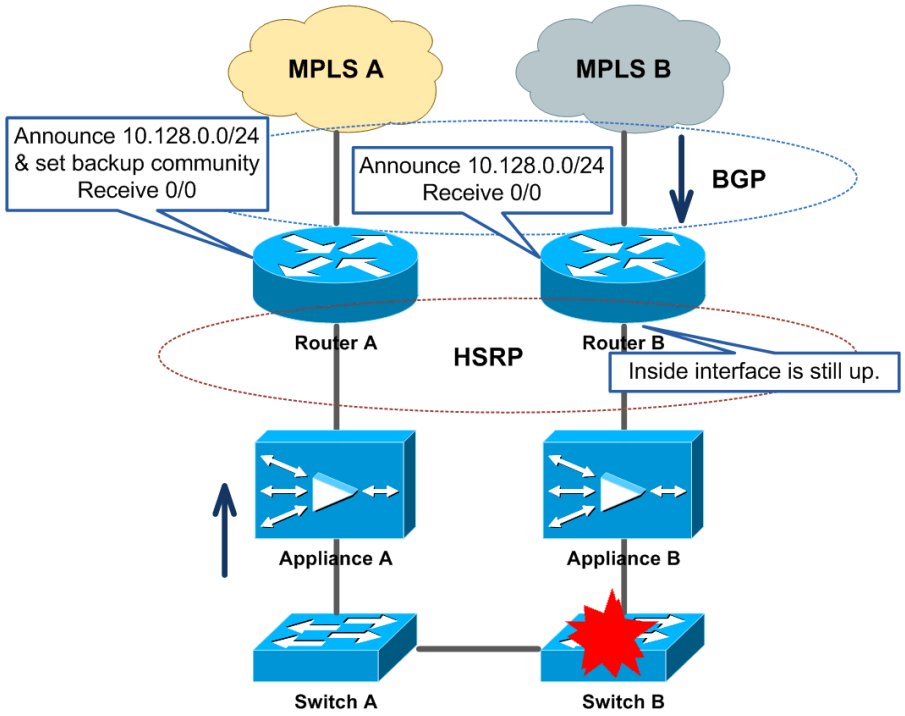
Inline Appliance



- Two appliances were introduced inline.
- They work transparently and are capable of failing to wire (given that correct cables and duplex settings were used).
- However, they do not feature Link State Propagation.

Cronus eBusiness

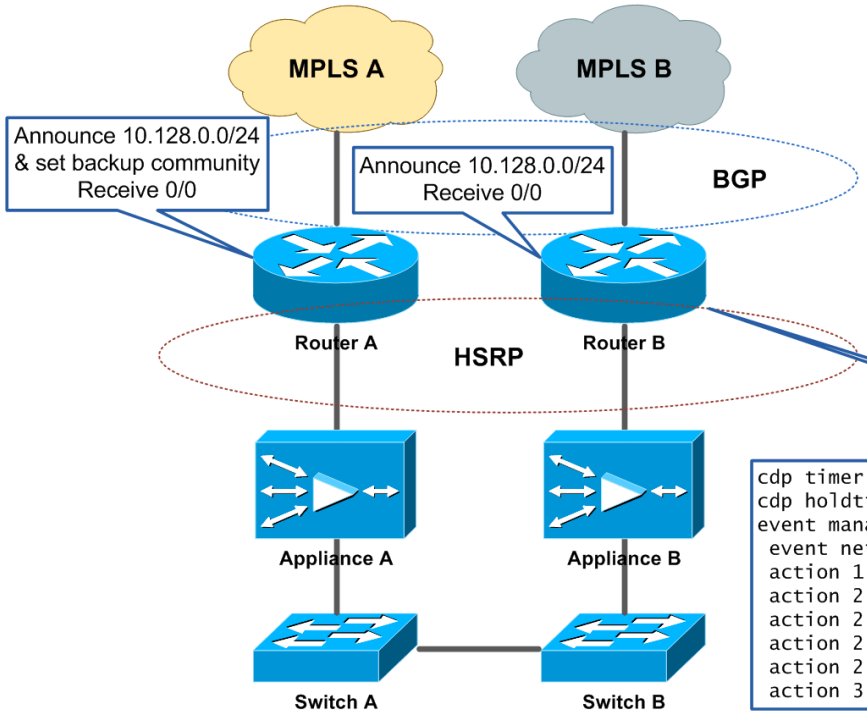
Inline Appliance



- Switch B fails, sometimes in the future, after appliances went in production.
- Router B has no knowledge of the failure and happily sends traffic to nowhere.
- How to detect and react to failure? Switch B is Layer 2 only, so there is no routing protocol hello (or BFD).

Cronus eBusiness

Inline Appliance



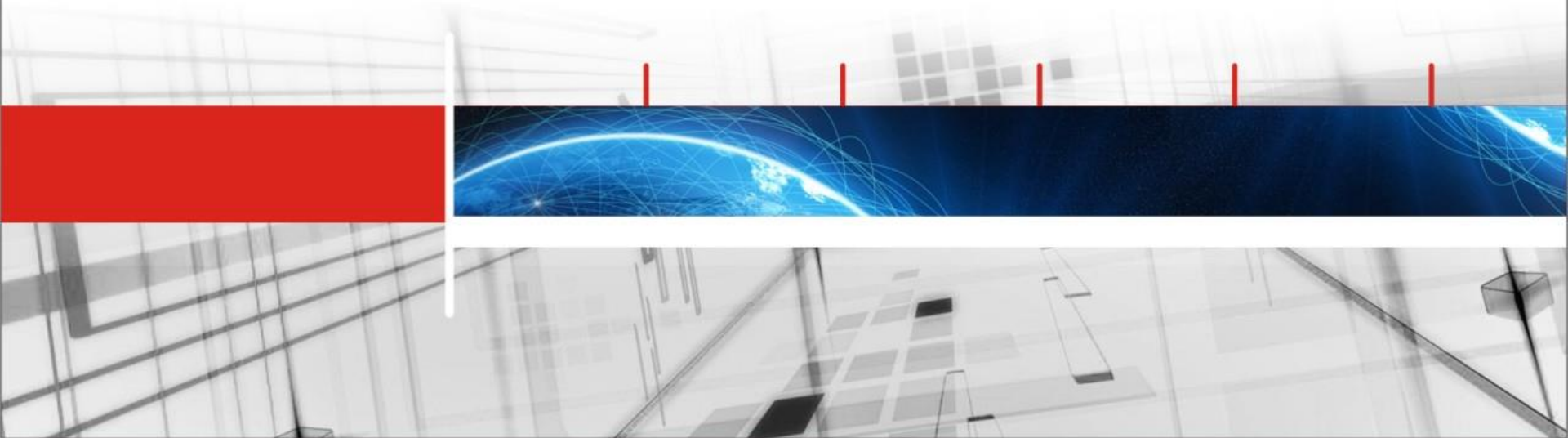
- Luckily, EEM can help.
- Otherwise, we will have to remove the appliances from the physical path and to redirect traffic to them.

```
cdp timer 5
cdp holdtime 15
event manager applet BGP-network-off authorization bypass
 event neighbor-discovery interface GigabitEthernet0/1 cdp delete
 action 1   cli command "enable"
 action 2   cli command "configure terminal"
 action 2.1 cli command "router bgp 65535"
 action 2.2 cli command "no network 10.128.0.0 mask 255.255.255.0"
 action 2.3 cli command "end"
 action 3   syslog msg "Local switch unreachable, BGP prefix removed."
```

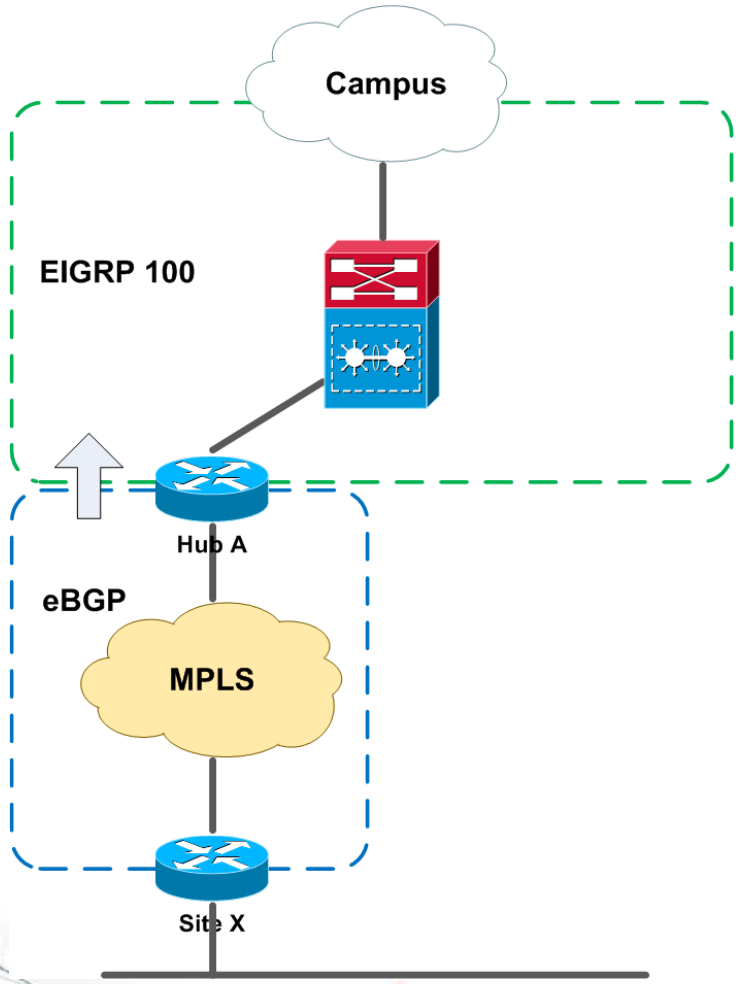
Cronus eBusiness

Take 5

MPLS + Internet WAN



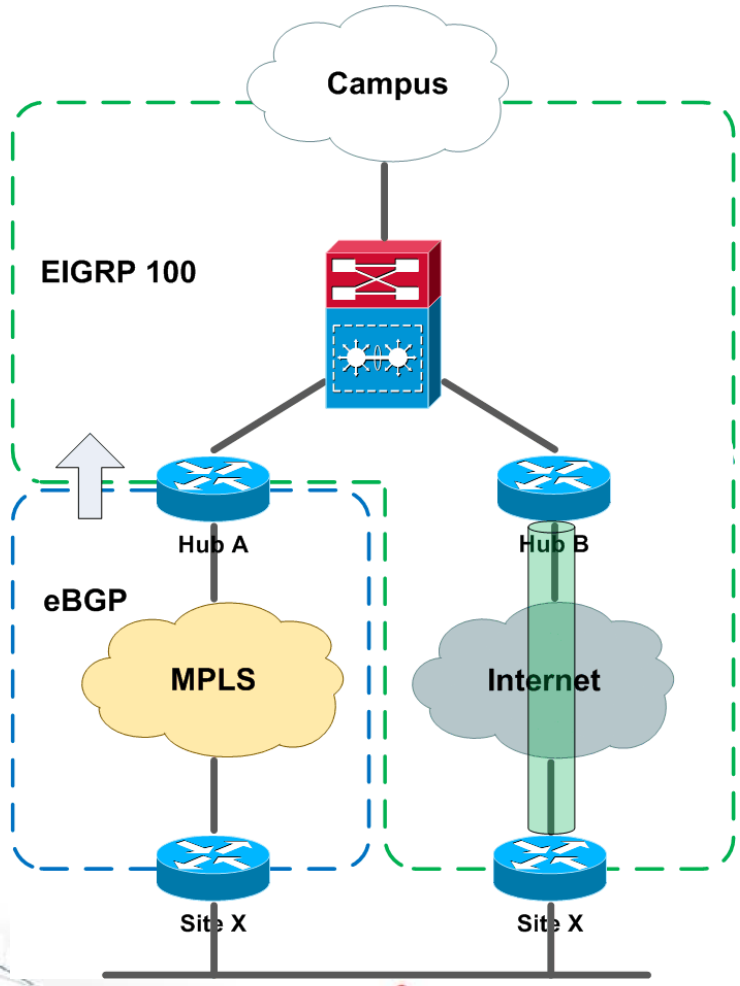
MPLS + Internet WAN



- The initial WAN design is based on an MPLS VPN cloud.
- BGP routes are redistributed into EIGRP 100 as external routes with default Admin Distance 170.
- Redundancy is now required at spoke sites, but the operational budget is low.

Cronus eBusiness

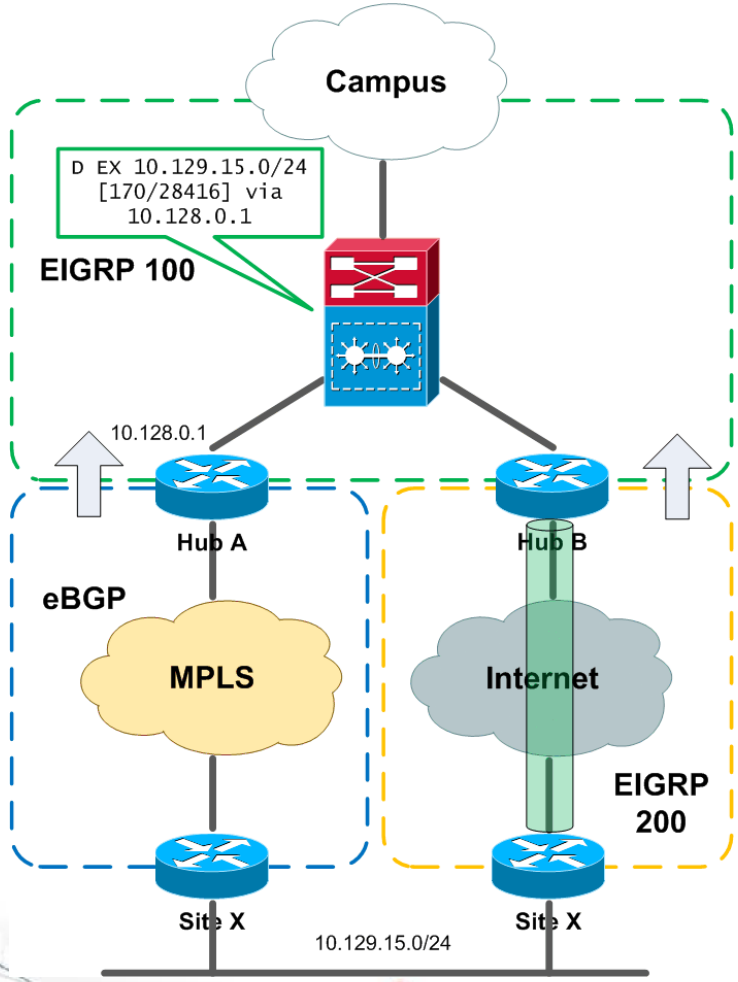
MPLS + Internet WAN



- A DMVPM cloud over the public Internet was introduced.
- But… running the same EIGRP AS for both campus and DMVPN network results in Internet path preferred over MPLS path.

Cronus eBusiness

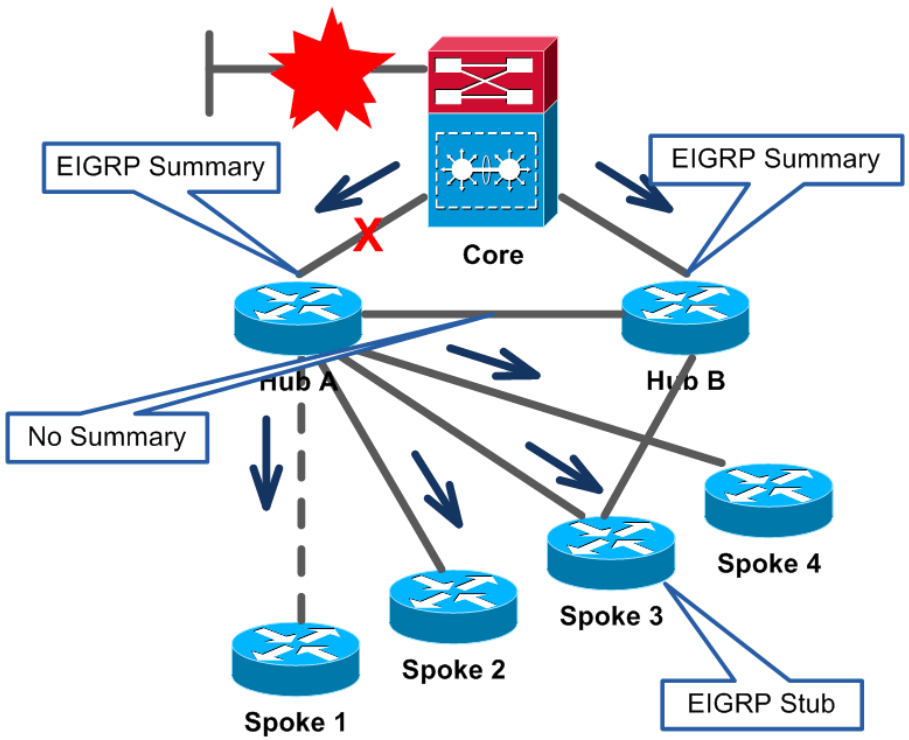
MPLS + Internet WAN



- Two EIGRP AS processes can be used to provide control of the routing.
- Routes from EIGRP 200 redistributed into EIGRP 100 appear as external route (distance = 170).
- We use EIGRP delay to modify path preference.
- Alternative(s): iWAN w/ PfR

Cronus eBusiness

MPLS + Internet WAN

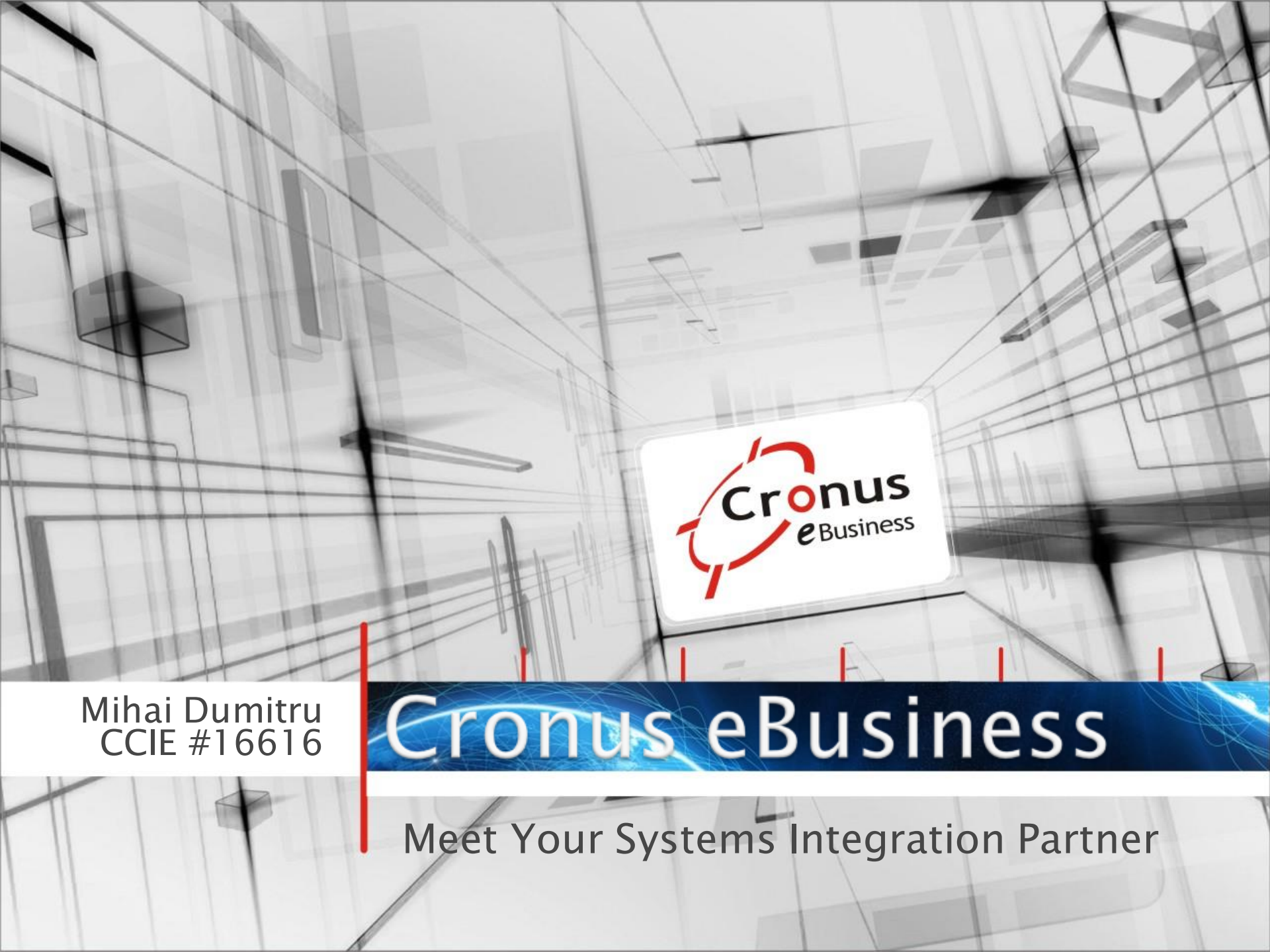


- The Internet links at the spokes were unreliable.
- EIGRP Stuck In Active (SIA): the Core–Hub A adjacency gets reset just because Core sees no reply from Spoke.
- Use EIGRP summary to hide spokes (have an inter-hub link to avoid black holing).
- Use EIGRP stub on spokes.

Cronus eBusiness

Conclusions

- Bad design may reveal itself *in the future*.
- Bad design is expensive: usually, there is no immediate fix, and sometimes it takes time to troubleshoot.
- Good design is a matter of good reference (Cisco Validated Designs, Cisco Live) and rich personal experience.
- You can Google for error messages or configuration examples, but how do you find a “free” suitable design and… *how do you know it is good?*



Mihai Dumitru
CCIE #16616

Cronus eBusiness

Meet Your Systems Integration Partner