IBM i TCP/IP redundancy and virtual Ethernet

Colin R. DeVilbiss April 23, 2014

Network virtualization and redundancy are the key techniques for keeping systems available while simplifying hardware management and reducing capital costs. This article describes how to combine Ethernet and IP virtualization and redundancy techniques using IBM i.

Network virtualization and redundancy technologies

IBM® Power Systems™ support several technologies that can provide network redundancy for IBM i partitions. This article discusses Ethernet link aggregation, IBM i Virtual IP Addressing (VIPA), and Ethernet layer-2 bridging, and explains how they interact with each other.

Ethernet link aggregation combines multiple physical Ethernet ports into a single Ethernet interface, providing redundancy by using all of the available ports at the same time. However, virtual Ethernet adapters do not support link aggregation, so an IBM i partition without physical Ethernet resources cannot directly take advantage of it.

An IBM i virtual IP address (VIPA) is based on multiple other IP interfaces. At any given time, the VIPA is using one of those IP interfaces for all of its outgoing traffic, and incoming traffic for the VIPA comes through those interfaces. If the underlying interface fails, the VIPA will change to use a different interface, and therefore the VIPA is resilient against individual interfaces failing.

With Ethernet layer-2 bridging, a partition can *share* its physical connection to the network by bridging virtual Ethernet traffic to the physical network, and bridging traffic from the physical network to the virtual Ethernet. This technology allows a *client* partition to access the network transparently without assigning any physical Ethernet resources. IBM i supports Ethernet layer-2 bridging, both as a *client* and as a *server* or *bridge provider*. Virtual I/O Server (VIOS) partitions also support layer-2 bridging, under the name *Shared Ethernet Adapter* (SEA). Layer-2 bridging does not provide redundancy by itself, but is a key enabler for redundant configurations with virtual Ethernet, as we'll see below.

These technologies combine in multiple interesting ways.

Link aggregation and layer-2 bridging

Even though virtual Ethernet does not directly support link aggregation, a bridge provider can use link aggregation on the physical interface that is used for the bridge. This transparently provides

the same benefits to all the client partitions simultaneously: traffic spreading across the physical Ethernet ports (for better utilization) and resilience against link failures.

From the client, this requires no additional configuration beyond the virtual Ethernet adapter that is required for layer-2 bridging. In the bridge provider partition, this just requires creating an aggregate interface or line description, then using it as the physical side of the bridge.

Shared Ethernet Adapter failover

The VIOS provides Shared Ethernet Adapter failover between two VIOS partitions. In this configuration, one of the SEAs is active, and the other serves as the backup. If the active SEA stops or fails for any reason, the backup SEA takes over and bridges the same traffic. This provides the same resilience against link failures on the active SEA, but also provides redundancy against whole-partition failures in the VIOS that owns the active SEA. However, the resources for the backup SEA are left idle.

This also uses the simplest client configuration; nothing is required beyond the virtual Ethernet adapter. Configuration in the VIOS partitions and Hardware Management Console (HMC) is more involved than a single bridge, and is described in detail in an IBM technote.

Shared Ethernet Adapter active/active (load sharing)

As mentioned above, with SEA failover, the Ethernet resources assigned to the backup SEA are left idle. In recent versions, VIOS has added support for a configuration that allows both physical Ethernet resources to be active at the same time. However, this configuration requires that the client partitions split their traffic across two or more virtual local area networks (VLANs) on the physical network (as defined in the IEEE 802.1q standard).

Within the client partitions, this environment still does not require any more configuration. However, the physical network design, VIOS configuration, and HMC configuration are significantly more complicated than any other environment described in this article. The detailed requirements and configuration steps can be found in the IBM Power Systems Information Center.

IBM i physical Ethernet with bridged virtual backup

An IBM i partition with access to physical Ethernet resources can use a hybrid configuration that uses the physical Ethernet resources while they are available, but fall back on a virtual Ethernet adapter if the physical interface becomes unavailable. For partitions that benefit from direct access to the physical network, this provides a fallback that can be shared among multiple partitions, allowing things such as network hardware replacement without an outage.

This environment requires more configuration in the client partition, including the following elements:

- A line description and IP interface for the physical Ethernet (which can be a single port or an aggregate),
- A line description and IP interface for the bridged virtual Ethernet adapter, and

- A VIPA with preferred interface (`PREFIFC`) with the physical interface first and virtual Ethernet second.

For example, given line descriptions `PHYSETH` and `VIRTETH` and appropriate IP addresses available on the network, these commands will configure the appropriate IP interfaces:

```
ADDTCPIFC INTNETADR('192.168.1.4') LIND(*VIRTUALIP) SUBNETMASK('255.255.255.255')
ADDTCPIFC INTNETADR('192.168.1.2') LIND(PHYSETH) SUBNETMASK('255.255.255.0')
    LCLIFC('192.168.1.4')
ADDTCPIFC INTNETADR('192.168.1.3') LIND(VIRTETH) SUBNETMASK('255.255.255.0')
    LCLIFC('192.168.1.4')
CHGTCPIFC INTNETADR('192.168.1.4') PREFIFC('192.168.1.2' '192.168.1.3')
```

After all three interfaces are started, the VIPA should be accessible from the network through the physical Ethernet port; if that port fails, then the virtual Ethernet port takes over.

IBM i failover between bridged virtual Ethernet adapters

The design for failover from physical Ethernet to virtual Ethernet works because the VIPA support detects the failure of the underlying physical Ethernet IP interface. However, failures in a bridge provider do not cause failures in the client partitions. When a bridge fails, outgoing client traffic will not reach the physical network, and physical network traffic intended for the client will not reach the virtual Ethernet, but any IBM i client line descriptions and IP interfaces will still remain active and will be able to send and receive internal traffic. Because there is no IP interface failure, VIPA support will not switch away from that interface. Therefore, failover between virtual Ethernet ports using VIPA requires some kind of monitor to detect failures in end-to-end connectivity to some external entity.

If the network already has multiple subnets with an external IP router between them, then one option is to use the VIPA support for routing protocols, such as the Routing Information Protocol (RIP) or Open Shortest Path First (OSPF). This requires that the underlying interfaces are on separate IP subnets and that the router is accessible by both paths. The IBM i partition will constantly attempt to advertise the availability of the VIPA on both paths. The routing protocol will discover any broken paths and the router and IBM i VIPA support will route traffic through working paths.

Another option is to configure VIPA using proxy Address Resolution Protocol (proxy ARP) as described in the previous section, but to run a program that can constantly monitor the availability of some external resource (such as an IP default gateway) through both paths. If the default or primary path is ever unable to access the gateway while the backup is able to, the program can switch over by modifying the order of the preferred interface (PREFIFC) list.

With VIPA and OSPF, RIP, or a monitor script, IBM i can use two bridged virtual Ethernet adapters as the backing interfaces to support a VIPA. In turn, that enables designs that can do active/active bridging and detect any end-to-end failure without using separate IEEE 802.1q VLANs in the physical network or VIOS SEA failover configuration. In these designs, the failover logic is located in a more complicated design in the clients, while the network and bridge provider configuration is left simple.

In this design, there must be two bridge provider partitions (either VIOS or IBM i), with the bridges configured on separate VLANs or switches at the HMC. Each bridge provider just has a basic layer-2 bridge configured. In each client partition, there should be one virtual Ethernet adapter for each bridge, and each needs a line description and an IP interface. If using VIPA with RIP or OSPF, the routing configuration is sufficient. If using proxy ARP instead, create the VIPA and run a monitor program. In the following example script, `IFC1` and `IFC2` are the addresses of the underlying virtual Ethernet-based interfaces, `VIPA` is the address of the VIPA, and `TARGET` is the address of an external resource that must be reachable for an interface to be considered usable (for example, a network gateway for the interface). In order for the routing and monitor program to work correctly together, apply V7R1 PTF MF57894.

```
PGM
DCL &TARGET *CHAR 15 '192.168.1.1'
DCL &IFC1   *CHAR 15 '192.168.1.2'
DCL &IFC2   *CHAR 15 '192.168.1.3'
DCL &VIPA   *CHAR 15 '192.168.1.4'

DCL &IFC1STAT *INT 4 0
DCL &IFC2STAT *INT 4 0
DCL &PREF     *INT 4 1

/* MAKE SURE PROXY IN PREFERRED ORDER */
CHGTCPIFC &VIPA PREFIFC(&IFC1 &IFC2)

/* BEGIN LOOP FOREVER TO MONITOR THE INTERFACES */
LOOP:
  /* PING THROUGH INTERFACE 1 AND COUNT FAILURES */
  PING &TARGET LCLINTNETA(&IFC1) NBRPKT(1) MSGMODE(*VERBOSE *ESCAPE)
  MONMSG MSGID(TCP3210) EXEC(GOTO IFC1FAIL)
  CHGVAR &IFC1STAT 0
  GOTO IFC2
IFC1FAIL:
  IF (&IFC1STAT < 9999) +
  THEN(CHGVAR &IFC1STAT (&IFC1STAT + 1))

IFC2:
  /* PING THROUGH INTERFACE 2 AND COUNT FAILURES */
  PING &TARGET LCLINTNETA(&IFC2) NBRPKT(1) MSGMODE(*VERBOSE *ESCAPE)
  MONMSG MSGID(TCP3210) EXEC(GOTO IFC2FAIL)
  CHGVAR &IFC2STAT 0
  GOTO UPDVIPA
IFC2FAIL:
  IF (&IFC2STAT < 9999) +
  THEN(CHGVAR &IFC2STAT (&IFC2STAT + 1))

UPDVIPA:
  /* TREAT 2 CONSECUTIVE FAILURES AS LINK DOWN INDICATION */
  IF (&IFC1STAT < 2) THEN(GOTO PREF1)
  IF (&IFC2STAT < 2) THEN(GOTO PREF2)
  /* NO INTERFACES APPEAR TO BE UP... NO CHANGES */
  GOTO DELAY

PREF1:
  IF (&PREF ^= 1) +
  THEN(CHGTCPIFC &VIPA PREFIFC(&IFC1 &IFC2))
  CHGVAR &PREF 1
  GOTO DELAY
PREF2:
  IF (&PREF ^= 2) +
  THEN(CHGTCPIFC &VIPA PREFIFC(&IFC2 &IFC1))
  CHGVAR &PREF 2
  GOTO DELAY
```

```
DELAY:
  DLYJOB 30
  GOTO LOOP
ENDPGM
```

Conclusion

When designing a multipartition system with IBM i, you have several virtualization techniques available, and they combine in powerful ways. This article describes several points in the design space that can give your system the resiliency it needs and potentially reduce your I/O capital expense by sharing Ethernet capacity among partitions.

Resources

See the following information center topics for more detail:

- Ethernet on IBM i
- IBM i TCP/IP Setup
- Virtual I/O Server