Port aggregation using a Cisco Catalyst 6513 switch and the IBM AIX operating system

John Lumby (lumby@ca.ibm.com)
Joyce Coleman (colemanj@ca.ibm.com)

1. Introduction

Port aggregation refers to the process of bonding two or more Ethernet ports together to create a single virtual network interface. All ports bonded together using port aggregation are given the same Media Access Control (MAC) address, and they therefore are treated by remote servers as a single port.

There are two primary reasons for using port bonding:

- It increases the link throughput beyond the throughput capable by any single port
- It provides redundancy in case any single port fails

This paper will give configuration guidelines for setting up port aggregation using a Cisco switch. Setting up port aggregation requires configuration on both the server side and the switch side. If incompatible settings are chosen on each side, performance will suffer. However, most of the available information on port aggregation is specific to either the server or the switch. The goal of this article is to fill in this gap by providing several examples of recommended configurations for both the server and the switch.

Because of the many variations that exist in the configuration steps for different models of switches and operating systems, it is not possible to provide a single set of configuration steps that covers all scenarios. Instead, this paper will provide configuration guidelines for connecting a Cisco Catalyst 6513 switch to servers running the IBM® AIX® operating system (version 5.2 or later) and performance measurements for a variety of configuration settings.

We will focus on two widely used port aggregation protocols. The first is the Link Aggregation Control Protocol (LACP), which is based on the IEEE 802.3ad standard. The second is the Port Aggregation Protocol (PAgP), a Cisco-proprietary protocol that can be run only on Cisco switches and on switches released by licensed vendors. We will also cover an additional method for setting up port aggregation that uses neither of these protocols.

All of our experiments were done on small clusters (typically two machines), and in all cases the port aggregation was created from two ports on the server.

2. Terminology

Terminology relating to port aggregation can be confusing, particularly since terms are not used consistently in the product documentation from different manufacturers, and in some cases are not used consistently even within the documentation for a single product. This section describes terms and definitions as they are used in this paper.

- *port aggregation*: The general concept of bonding two or more Ethernet adapters together to create a single virtual network interface, regardless of the protocol used.
- *bonding*: Port aggregation on Linux® platforms.[1]
- *channel*: A synonym for port aggregation.
- *Link Aggregation Control Protocol (LACP)*: A non-proprietary protocol for port aggregation defined in IEEE 802.3ad. Also referred to as the IEEE 802.3ad or 802.3ad protocol. On the AIX platform, the mode that corresponds to this protocol is called "8023ad LACP"; on the Linux platform it is called "802.3ad".
- *802.3ad Link Aggregation (Link Aggregation)*: A port aggregation created using the LACP protocol.
- *Port Aggregation Protocol (PAgP)*: A Cisco-proprietary protocol for port aggregation. If the PAgP protocol is specified in the switch, then a non-LACP mode must be specified on the server for the AIX port aggregation or Linux bond policy.
- *EtherChannel*: In AIX documentation, this term is generally used to refer to any non-LACP port aggregation. We use this term in the AIX sense to refer to a port aggregation created using a protocol other than LACP, or using no protocol at all. (In Cisco documentation, by contrast, this term is used to refer to any port aggregation, whether or not it follows the LACP protocol.)

The following two terms relate specifically to Cisco switches. The descriptions provided are not formal definitions, but rather a guide for readers of this document.

- *channel group*: The group of ports in a port aggregation. A channel group is assigned a number called the administrative group number, which identifies it uniquely. This term has the same meaning for both Cisco IOS (IOS) and CatOS (two operating systems that run on Cisco switches). The channel-group view of a port aggregation is the view of the **composition** of the aggregation, that is, what this aggregation consists of.
- *port channel*: The interface that is created from a channel group. In other words, whereas the channel group is the internal composition of a port aggregation, the port channel is the external view of that channel group as a special kind of interface, that is, what attributes (such as maximum transmission unit, enabled/disabled) this aggregation has as a single entity. This term has similar but not quite identical meanings for IOS and CatOS. Under IOS, it is a black-box view of the port aggregation, as described above. Under CatOS, it also includes some attributes of the ports that make up the port channel. Thus a *port channel* is closer in meaning to a *channel group* under CatOS than under IOS.

[1] For information about bonding on Linux, consult the following Web site: http://www.linuxfoundation.org/en/Net:Bonding

3. Hardware prerequisites

On the server side, only certain network interface cards (NICs) are supported. All NICs used must have the same link speed and duplex setting. Furthermore, for LACP, the duplex setting must be full-duplex.

Both LACP and PAgP require support in the switch, if a switch is used. Although it is possible to use port aggregation without a switch by connecting two servers back-to-back, this scenario is not described here.

Jumbo frames are not necessary for port aggregation, but for most network workloads, the use of jumbo frames can reduce server utilization or improve throughput, or both. If you are using jumbo frames, all hardware components (interfaces and switch modules) and software components (switch and server operating systems) in the network must support them. Not all switches support jumbo frames, although the Cisco 6500 series does. Also, if the interfaces used for the workload will need to interoperate with other devices on the same network that do not support jumbo frames, then jumbo frames cannot be used. Ideally, jumbo frames should be used only in self-contained clusters in which all machines in the cluster can be configured for jumbo frames. Jumbo frames are discussed in greater detail later in this document.

4. General guidelines on configuration

The goal when setting up port aggregation is to balance traffic across all interfaces, both between the server and the switch, and between the switch and the server.

On the server, the choice of mode and hash function determines the way in which outgoing traffic is distributed across the server's ports. Incoming traffic is distributed according to the switch configuration.

The most commonly adopted hash functions available on both the server and the switch are based on the IP address and/or MAC address of the source or destination, or both. One factor in configuring port aggregation will therefore be the size of the cluster. If all of the traffic being sent on one port aggregation is destined for a single MAC address or a single IP address, or for a very small number of MAC addresses or IP addresses, as would typically be the case for a very small cluster, then the traffic is likely to be very unbalanced if the choice of interface is based on MAC address or IP address. In other words, traffic will be sent mainly or entirely over one interface. We focus on ensuring balance in this type of small cluster, where balance is crucial for good performance.

5. Server configuration options

Beginning with AIX 5.2, three mutually exclusive modes are available on the server:

- *standard*: This mode indicates that the port aggregation does not participate in IEEE 802.3ad LACP. The distribution of outgoing packets is described below.
- *8023ad*: This mode enables the use of IEEE 802.3ad LACP. The distribution of outgoing packets is described below.
- *round-robin*: This mode indicates that the port aggregation does not participate in IEEE 802.3ad LACP. All outgoing packets are distributed evenly across all ports in the port aggregation, and may be sent out in a slightly different order from the order in which they were sent to the port aggregation.

It is not recommended to set the mode to *round-robin.* Although this mode provides the highest bandwidth utilization, its use leads to the highest probability of out-of-order packets, which is correlated with a lower throughput.[2] *Round-robin* is the typical choice for hosts connected back-to-back, and will not be discussed in this article.

Either *standard* or *8023ad* mode is recommended. These modes configure both the server and switch to choose the interface within the channel in a similar way. In particular, for any one connection (that is, for a specific combination of source and destination port), the network interfaces chosen to send packets on each side of the switch will always be the same. This minimizes the number of out-of-order packets.

When *standard* or *8023ad* mode is chosen, a hash function based on the content of the packet's header is used to determine on which adapter outgoing packets are sent. Which fields of the header are used depends on the hash mode. Four hash modes are available. The AIX Information Center provides the following description of these modes:[3]

- *default*: “The destination IP address of the packet is used to determine the outgoing adapter. For non-IP traffic (such as ARP), the last byte of the destination MAC address is used to do the calculation. This mode guarantees packets are sent out over the EtherChannel in the order they were received, but it may not make full use of the bandwidth.”
- *src_port*: “The source UDP or TCP port value of the packet is used to determine the outgoing adapter. If the packet is not UDP or TCP traffic, the last byte of the destination IP address will be used. If the packet is not IP traffic, the last byte of the destination MAC address will be used.”
- *dst_port*: “The destination UDP or TCP port value of the packet is used to determine the outgoing adapter. If the packet is not UDP or TCP traffic, the last byte of the destination IP will be used. If the packet is not IP traffic, the last byte of the destination MAC address is used.”

[2] See section 9.4 Discussion of results for a discussion of some unexpected results relating to out-of-order packets and throughput.

[3] http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.commadmn/doc/commadmndita/etherchannel_config.htm. Note: ARP stands for Address Resolution Protocol.

- *src_dst_port*: “The source and destination UDP or TCP port values of the packet is used to determine the outgoing adapter (specifically, the source and destination ports are added and then divided by two before being fed into the algorithm). If the packet is not UDP or TCP traffic, the last byte of the destination IP is used. If the packet is not IP traffic, the last byte of the destination MAC address will be used. This mode can give good packet distribution in most situations, both for clients and servers.”

In this article, we recommend setting the hash mode to *src_dst_port* because this is the setting that will most likely lead to balanced performance.

6. Cisco switch configuration options

There are several different kinds and levels of configuration options. The following list gives some relevant options:

- Frame/switch level – load-balancing distribution method
- Module/blade level – aggregation protocol
- Port level – port mode

Note that this list is not exhaustive, and that the default values and methods for specifying options can vary depending on which of two operating systems (CatOS and IOS) is running on the switch. Furthermore, not all options apply to all switch models.

6.1 Frame/switch level - load-balancing distribution method

The available load-balancing distribution methods are:

- MAC - MAC address(es) (Layer 2) – this is the default method on the CatOS operating system
- IP - IP address(es) (Layer 3) – this is the default method on the IOS operating system
- TCP/UDP port number(s) (Layer 4) – this is referred to as “Session” in CatOS.

The load balancing can be based solely on the source address, destination address, or both source and destination addresses.

6.2 Module/blade level - aggregation protocol

The available protocols are PAgP and LACP.

On the CatOS operating system, the aggregation protocol is specified at the module/blade level using the following command:

```
set channelprotocol {pagp | lacp} mod
```

On the IOS operating system, the protocol runs on the switch processor. For this operating system only, it is not necessary to specify the protocol. When the protocol is not specified, it is inferred from other settings, most importantly the port mode. To specify the protocol explicitly, use one of the following commands:

```
      channel-protocol {lacp | pagp}
      channel-group
```

6.3 Port level - port mode

The port mode determines how a port will negotiate and participate in a port aggregation.[4] Note in particular that certain port settings will, in effect, override or suppress the chosen module aggregation protocol for that particular port.

6.3.1 Port modes for PAgP

There are two PAgP-specific port mode settings:

- *auto* - This is the default port mode setting. When it is used, a port will respond to received PAgP packets but will not initiate a PAgP negotiation sequence.
- *desirable* - This is an alternative port mode setting. When it is used, a port will try to initiate a PAgP negotiation sequence with one or more other ports.

Therefore, a LAN port in *desirable* mode can form an EtherChannel with another LAN port that is in either *auto* or *desirable* mode. Two LAN ports in *auto* mode cannot form an EtherChannel because neither port will initiate negotiation. The keywords *silent* and *non-silent* can be used with (and only with) the *auto* and *desirable* modes. Specify *silent* when no traffic is expected from the other device to prevent the link from being reported to the spanning-tree protocol as down. When the other device is an interface on a server that will run the networking application, as is usually the case, always specify *non-silent*.

In addition, there is another port mode ("*on*") that indicates that ports will be aggregated manually (by the administrator) without the use of an automatic channeling protocol. In the case of CatOS, the associated module must be configured with the PAgP protocol and the protocol is then ignored. In the case of IOS, this option is called *manual channeling* and no protocol is selected. This option is described in section 7.1.3 Manual channeling (no protocol) below. Do not specify either *silent* or *non-silent* with this mode.

6.3.2 Port modes for LACP

When configuring switch ports under the LACP protocol, there are two LACP-specific port mode settings:

[4] Consult http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/catos/8.x/configuration/guide/channel.html for more information about the PAgP- and LACP-specific port mode settings.

- *passive* - This is the default port mode setting. When it is used, a port will respond to received LACP packets but will not initiate a LACP negotiation sequence.
- *active* - This is an alternative port mode setting. When it is used, a port will try to initiate a LACP negotiation sequence with one or more other ports.

These modes correspond to the *auto* and *desirable* modes described above. In other words, a LAN port in *active* mode can form a Link Aggregation with another LAN port that is in either *passive* or *active* mode. Two LAN ports in *passive* mode cannot form a Link Aggregation because neither port will initiate negotiation.

7. Recommended configuration options

7.1 Load-balancing distribution method based on Layer 4

The options described in this section use load-balancing distribution methods based on Layer 4, or TCP/UDP port numbers. The load-balancing option you should choose depends on which operating system is running on the switch:

- On the CatOS operating system, set the load-balancing distribution method to *session both.*
- On IOS, set it to *src-dst-port.*[5]

7.1.1 PAgP protocol

- On the server, set the mode to *standard* and the hash mode to *src_dst_port.*
- On the switch, set the frame load-balancing distribution method to the appropriate choice for the operating system as described above; for CatOS, set the module protocol to *PAgP*; and for both switch operating systems, set the port mode to *desirable non-silent.*

7.1.2 LACP protocol

- On the server, set the mode to *8023ad* and the hash mode to *src_dst_port.*
- On the switch, set the frame load-balancing distribution method to the appropriate choice for the operating system as described above; for CatOS, set the module protocol to *IEEE 802.3ad LACP*; and for both switch operating systems, set the port mode to *active.*

7.1.3 Manual channeling (no protocol)

- On the server, set the mode to *standard* and the hash mode to *src_dst_port.*

[5] Newer releases of the IOS operating system offer a mode called *src-dst-mixed-ip-port* that might offer advantages over the *src-dst-port* mode. This mode is not discussed here because of how new it is.

- On the switch, set the frame load-balancing distribution method to the appropriate choice for the operating system as described above; for CatOS, set the module protocol to *PAgP* (note that this setting will be ignored); and for both switch operating systems, set the port mode to *on*.
- On the switch, manually create channel groups, each one containing the ports in that channel. It is important to know which ports on the switch are connected to which interfaces on the server.

If the channeling is controlled by manually assigning switch ports to groups, it is possible to set the port setting to *on*, which forces the port to channel without PAgP. That is, channeling is effectively controlled by the physical cabling and port groups, without any recourse to the aggregation protocol. This configuration provides an example of what is referred to as manual channeling, in which a port setting can override the module protocol for that port. This setting can, in some cases, provide a higher maximum throughput than other port settings that use the PAgP protocol.

However, the disadvantage of this configuration is that it can lead to black-holing traffic. That is, in the event of a hardware failure on one interface, the switch might not mark the failing interface as down, which would result in a high rate of errors and retransmissions until the administrator manually reconfigures the channel. For this reason, it is recommended that the PAgP or LACP protocol be used when possible.

7.2 Load-balancing distribution method based on Layers 2 or 3

The options described in the preceding section are recommended where they are supported. However, there are several reasons why it might not be possible to use a configuration in which the frame load-balancing distribution method is set to Layer 4. For example:

- This configuration might not be supported by the particular combination of operating system and hardware. For example, on switches other than the Cisco Catalyst 6513 described in this article, Layer 4 load-balancing might not be available.
- Setting the option to this value might pose problems for the installation because this setting applies to all ports on all modules on the switch. If the switch is used by other unrelated machines, it might not be desirable to change the setting for all machines. Even if the decision is made to modify this setting for all machines, the system should be tested extensively to verify that the setting does not adversely affect the unrelated machines in the environment.

When Layer 4 load-balancing is not possible, you should use one of the following configurations instead. Again, the load-balancing option you should choose depends on which operating system is running on the switch, but will be either Layer 2 (MAC) or Layer 3 (IP).

7.2.1 PAgP protocol

- On the server, set the mode to *standard* and the hash mode to *src_dst_port.*
- On the switch, set the frame load-balancing distribution method to an appropriate choice (not Layer 4). On CatOS, set the module protocol to *PAgP.* Set the port mode to *desirable non-silent.*

7.2.2 LACP protocol

- On the server, set the mode to *8023ad* and the hash mode to *src_dst_port.*
- On the switch, set the frame load-balancing distribution method to an appropriate choice (not Layer 4). On CatOS, set the module protocol to *IEEE 802.3ad LACP.* Set the port mode to *active.*

7.3 Hardware considerations

In order to obtain the desired network throughput, it is important to take several hardware configuration factors into account, including the choice of Ethernet adapters in the servers and the choice of modules and other equipment in the Cisco switch. This document does not address all such considerations, but it should be noted that, for a configuration in which two or more servers, each with a port aggregation of two or more Gigabit adapter ports, are interconnected using a Cisco model 6509 or 6513 switch, the recommended Ethernet interface module is as follows:

48 port 10/100/1000mb Ethernet   WS-X6748-GE-TX

This module has certain prerequisites concerning Supervisor Engine hardware.

8. Sample instructions for configuring port aggregation

This section provides sample instructions for setting up three different port aggregation configurations. In all cases, the switch runs the IOS operating system.

The following examples assume that:

- The name of the AIX port aggregation is `ent6`, and it consists of two physical interfaces, `ent2` and `ent4`, and has the logical interface `en6`
- The switch Gigabit Ethernet interfaces on module 8 are named 8/4 8/5 (first group) and 8/26 8/27 (second group), and these two groups are formed into port channels that are numbered 3 and 4, respectively. The network is assigned to a VLAN named 129. Examples of configuring port channel 3 are shown. The equivalent instructions are not shown for port channel 4 because these are identical except for parameters such as names and numbers.

All examples assume that there are no other mutually conflicting, pre-existing configuration options already in effect. If there are, these must first be removed or undone. Also, some of the sample parameter settings such as IP addresses might be unnecessary or might need to be changed for a specific environment. Other parameter settings not included below might need to be specified if the configuration is different.

When configuring a switch running IOS, it is important, when verifying the configuration using **`show`** commands, to keep in mind that that are two kinds of configuration sets for each interface: the running configuration and the active or operating configuration. Roughly speaking, the running configuration is the accumulation of options set by commands entered by the administrator, whereas the operating configuration is the current state of the interface. For any interface (Gigabit Ethernet or port channel), the **`show interface`** command displays its operating mode, and the **`show running-config`** command displays its configured mode. Both of these configurations must be viewed to obtain a complete picture of the interface.

8.1 Configuring LACP active aggregation and 8023ad mode with Layer 4 load-balancing

Set up the AIX port aggregation:

Notes:

- The following steps assume support for jumbo frames. Refer to section 9.2.7 Measurements comparing standard and jumbo frames for more information.
- AIX commands are shown in explicit command-line format for ease of exposition. However, it might instead be easier to use the smit user interface. For example, the command “smitty etherchannel” can be used to configure the port aggregation.

1. Set jumbo frames in the AIX physical network interface, which is assumed to exist already:

```
chdev -l ent2 -a jumbo_frames=yes
chdev -l ent4 -a jumbo_frames=yes
```

2. Create the AIX physical port aggregation from the physical interfaces, with jumbo-frame support:

```
mkdev -c adapter -s pseudo -t ibm_ech -a adapter_names="ent2,ent4" \
  -a mode=8023ad -a hash_mode=src_dst_port -a use_jumbo_frame=yes \
  -a num_retries=3 -a retry_time=1
```

3. Create the AIX logical port aggregation from the AIX physical port aggregation. Setting the MTU (maximum transmission unit) to 9000 enables jumbo-frame support in the logical port aggregation:

```
mkdev -c if -s EN -t en -a netaddr="253.125.129.9" -a \
  netmask="255.255.255.0" -w "en6" -a state=up -a arp=on -a mtu=9000
```

4. Verify the attributes:

```
lsattr -H -E -l ent2
lsattr -H -E -l ent4
lsattr -H -E -l ent6
lsattr -H -E -l en6
ifconfig en6
```

Configure the switch Gigabit Ethernet ports:

5. Perform one (not both) of the following steps:

- Check to discover if any port channels are already defined and available:
  ```
  show etherchannel summary
  ```
  and if so pick an available one.

Or:

- Remove any existing channel group if it exists. For example, for channel 3:
  ```
  configure terminal
  no interface port-channel 3
  end
  ```

6. Create the LACP channel group from Gigabit Ethernet interfaces 8/4-5. At the same time, ensure that MTU 9216 is enabled on each physical interface. (Note: This can be done in the scope of the range only if the ports to be configured are consecutive. If they are not consecutive, you must set this value for each port individually, as shown in steps 9 and 10.)

```
configure terminal
interface range gigabitethernet 8/4-5
mtu 9216
channel-group 3 mode active
end
```

7. Enter the port channel submenu and add the relevant configuration, setting the access VLAN, description, and MTU, and enabling the port channel:

```
configure terminal
interface port-channel 3
switchport access vlan 129
description ??????
mtu 9216
no shut
end
```

8. Repeat steps 5 through 7 for channel group 4.

9. [Perform this step only if the ports are not consecutive.] Enter the interface submenu and ensure that MTU 9216 is enabled in each port.

```
configure terminal
interface gigabitethernet 8/4
mtu 9216
no shut
end
```

10. [Perform this step only if the ports are not consecutive.] Repeat step 9 for each of the three remaining Gigabit Ethernet interfaces.

11. Configure load-balancing:
```
configure terminal
port-channel load-balance src-dst-port
end
```

12. Show the running and operational configurations for all interfaces, and show the port aggregations including load-balancing. When you view the output from the **`show`** command, you should verify that the following conditions are all true:

- The channels are shown in the port aggregation summary, and with the correct protocol
- The channels consist of the correct Gigabit Ethernet interfaces
- The channels are shown as being switchports with no IP address
- The channels all have the correct MTU
- The ports are enabled and have the correct MTU

```
show run interface port 3
show interface port 3
show running interface gigabitethernet 8/4
show interface gigabitethernet 8/4
show interface status
show etherchannel summary
show etherchannel port
show etherchannel port-channel
show etherchannel lo
```

8.2 Configuring manual channeling aggregation and standard mode with Layer 4 load-balancing

Most commands are identical to those given in section 8.1 Configuring LACP active aggregation and 8023ad mode with Layer 4 load-balancing. Only steps 2 and 6 are different.

1. Set jumbo frames in the AIX physical network interface, which is assumed to exist already:
```
chdev -l ent2 -a jumbo_frames=yes
chdev -l ent4 -a jumbo_frames=yes
```

2. Create the AIX physical port aggregation from the physical interfaces, with jumbo-frame support:
```
mkdev -c adapter -s pseudo -t ibm_ech -a adapter_names="ent2,ent4" \
  -a mode=standard -a hash_mode=src_dst_port -a use_jumbo_frame=yes \
  -a num_retries=3 -a retry_time=1
```

3. Create the AIX logical port aggregation from the AIX physical port aggregation. Setting the MTU (maximum transmission unit) to 9000 enables jumbo-frame support in the logical port aggregation:

```
mkdev -c if -s EN -t en -a netaddr="253.125.129.9" -a \
  netmask="255.255.255.0" -w "en6" -a state=up -a arp=on -a mtu=9000
```

4. Verify the attributes:

```
lsattr -H -E -l ent2
lsattr -H -E -l ent4
lsattr -H -E -l ent6
lsattr -H -E -l en6
ifconfig en6
```

Configure the switch Gigabit Ethernet ports:

5. Perform one (not both) of the following steps:

- Check to discover if any port channels are already defined and available:
  ```
  show etherchannel summary
  ```
  and if so pick an available one.

Or:

- Remove any existing channel group if it exists. For example, for channel 3:
  ```
  configure terminal
  no interface port-channel 3
  end
  ```

6. Create the manual channel group from Gigabit Ethernet interfaces 8/4-5:

```
configure terminal
interface range gigabitethernet 8/4-5
channel-group 3 mode on
end
```

7. Enter the port channel submenu and add the relevant configuration, setting the access VLAN, description, and MTU, and enabling the port channel:

```
configure terminal
interface port-channel 3
switchport access vlan 129
description ??????
mtu 9216
no shut
end
```

8. Repeat steps 5 through 7 for channel group 4.

9. [Perform this step only if the ports are not consecutive.] Enter the interface submenu and ensure that MTU 9216 is enabled in each port.

```
configure terminal
interface gigabitethernet 8/4
mtu 9216
no shut
end
```

10. [Perform this step only if the ports are not consecutive.] Repeat step 9 for each of the three remaining Gigabit Ethernet interfaces.

11. Configure load-balancing:

```
configure terminal
port-channel load-balance src-dst-port
end
```

12. Show the running and operational configurations for all interfaces, and show the port aggregations including load-balancing. When you view the output from the **`show`** command, you should verify that the following conditions are all true:

- The channels are shown in the port aggregation summary, and with the correct protocol
- The channels consist of the correct Gigabit Ethernet interfaces
- The channels are shown as being switchports with no IP address
- The channels all have the correct MTU
- The ports are enabled and have the correct MTU

```
show run interface port 3
show interface port 3
show running interface gigabitethernet 8/4
show interface gigabitethernet 8/4
show interface status
show etherchannel summary
show etherchannel port
show etherchannel port-channel
show etherchannel lo
```

8.3 Configuring PAgP desirable aggregation with standard mode

Most commands are identical to those given in section 8.1 Configuring LACP active aggregation and 8023ad mode with Layer 4 load-balancing. Only steps 2 and 6 are different. The instructions for configuring load-balancing (step 11) are not given since they are presumed to be pre-set or unalterable, or both.

1. Set jumbo frames in the AIX physical network interface, which is assumed to exist already:

```
chdev -l ent2 -a jumbo_frames=yes
chdev -l ent4 -a jumbo_frames=yes
```

2. Create the AIX physical port aggregation from the physical interfaces, with jumbo-frame support:

```
mkdev -c adapter -s pseudo -t ibm_ech -a adapter_names="ent2,ent4" \
  -a mode=standard -a hash_mode=src_dst_port -a use_jumbo_frame=yes \
  -a num_retries=3 -a retry_time=1
```

3. Create the AIX logical port aggregation from the AIX physical port aggregation. Setting the MTU (maximum transmission unit) to 9000 enables jumbo-frame support in the logical port aggregation:

```
mkdev -c if -s EN -t en -a netaddr="253.125.129.9" -a \
  netmask="255.255.255.0" -w "en6" -a state=up -a arp=on -a mtu=9000
```

4. Verify the attributes:

```
lsattr -H -E -l ent2
lsattr -H -E -l ent4
lsattr -H -E -l ent6
lsattr -H -E -l en6
ifconfig en6
```

Configure the switch Gigabit Ethernet ports:

5. Perform one (not both) of the following steps:

- Check to discover if any port channels are already defined and available:

  ```
  show etherchannel summary
  ```

  and if so pick an available one.

Or:

- Remove any existing channel group if it exists. For example, for channel 3:

  ```
  configure terminal
  no interface port-channel 3
  end
  ```

6. Create a manual channel group from Gigabit Ethernet interfaces 8/4-5:

```
configure terminal
interface range gigabitethernet 8/4-5
channel-group 3 mode desirable non-silent
end
```

7. Enter the port channel submenu and add the relevant configuration, setting the access VLAN, description, and MTU, and enabling the port channel:

```
configure terminal
interface port-channel 3
switchport access vlan 129
description ??????
mtu 9216
no shut
end
```

8. Repeat steps 5 through 7 for channel group 4.

9. [Perform this step only if the ports are not consecutive.] Enter the interface submenu and ensure that MTU 9216 is enabled in each port.

```
configure terminal
interface gigabitethernet 8/4
mtu 9216
no shut
end
```

10. [Perform this step only if the ports are not consecutive.] Repeat step 9 for each of the three remaining Gigabit Ethernet interfaces.

11. Show the running and operational configurations for all interfaces, and show the port aggregations including load-balancing. When you view the output from the **`show`** command, you should verify that the following conditions are all true:

- The channels are shown in the port aggregation summary, and with the correct protocol
- The channels consist of the correct Gigabit Ethernet interfaces
- The channels are shown as being switchports with no IP address
- The channels all have the correct MTU
- The ports are enabled and have the correct MTU

```
show run interface port 3
show interface port 3
show running interface gigabitethernet 8/4
show interface gigabitethernet 8/4
show interface status
show etherchannel summary
show etherchannel port
show etherchannel port-channel
show etherchannel lo
```

9. Experimental results

Network throughput and degree of load-balancing were measured for a number of different combinations of server and switch configuration options. This section summarizes these results.

9.1 System and application

Two test systems were used.

The first test system consisted of two IBM® Power 570 (formerly called System p® 570) servers, each with a pair of dual-port 10/100/1000 Base-TX PCI-Express adapters, with one port from each adapter aggregated into a port aggregation. The servers were installed with AIX version 5.3 and were interconnected by a Cisco 6513 network switch containing a 48-port 10/100/1000mb Ethernet WS-X6748-GE-TX module and running the IOS operating system.

The second test system consisted of two IBM System p5 575 servers, each with a pair of Gigabit Ethernet-SX PCI-X adapters aggregated into a port aggregation. The servers were installed with AIX version 5.3 and were interconnected by a Cisco 6513 network switch. The 8-port 1000BaseX Ethernet WS-X6408A-GBIC module was used in the 6513 switch. Since the measurements were taken, Cisco has withdrawn this module from marketing. A more recent and recommended module is the WS-X6748-GE-TX module mentioned above. On this system the switch ran the CatOS operating system.

The application used to test various configurations is called *socketprocs*. In operation, one set of processes (the “client” processes) run on one machine and another set of processes (the “server” processes) run on the same or another machine. On each machine, N processes each connect to each of N processes on the other, for a total of N*N connections, using the TCP/IP protocol. On the server machine, all processes act as servers and listen, accept, and then *recv* and reply. On the client machine, all processes connect and send, and *recv* the reply, until the specified number of messages has been sent. All messages are a fixed size, with a default size of 4K.

Note an important aspect of this workload as regards port aggregation: all N*N connections have the same source MAC address (that of the sending port aggregation) and the same destination MAC address (that of the receiving port aggregation - the Layer 2 information), as well as the same source and destination IP addresses (those of the port aggregation - the Layer 3 information), but every connection has a different combination of source and destination TCP port number (the Layer 4 information).

9.2 Configurations and results -- Test system 1

9.2.1 Configuration settings

Forty-two runs were completed, each with a different configuration. Each run is identified by a unique number. Columns 2-6 contain an encoded description of the setup and configuration:

App meth. The method used by the application for waiting for TCP/IP events:

- s select - formerly the most common method in use in applications; used in versions of the DB2® database product prior to Version 9.5.0.1.
- e pollset - using edge-triggered polling: has the potential for higher throughput owing to use of non-blocking sockets and less waiting, but with possibly higher CPU utilization. This alternative was used in most runs with optimal configuration in order to demonstrate the maximum achievable throughput. DB2 Version 9.5.0.1 and later versions use this method.
- p pollset - using simple level-triggered polling: an evolution of traditional poll offering lower CPU utilization.

Num procs. The number of processes sending and receiving on each server.

Switch config. Three configuration values, separated by hyphens, are given.

The first value is the frame load-balancing policy (frame distribution). Two policies were assessed:

- p : src_dest_port (Layer 4)
- i : src_dest_ip (Layer 3)

The second value is the port aggregation protocol:

- P : PAgP
- L : LACP
- M : manual. This means that channel groups were defined manually but with no aggregation protocol - that is, port mode on
- N : none. This means that channel groups were not defined

The third value is the port mode. The port mode depends on, and has different meanings, depending on whether the PAgP or LACP protocol is used:

- d : PAgP desirable
- v : LACP active
- o : on (manual channeling)
- n : none - corresponding to protocol N, meaning that channel groups were not defined

Server config. Two configuration values, separated by hyphens, are given. These are the two principal configuration options in the AIX port aggregation. The first value is the mode, which specifies both the aggregation mode and the method by which the port on which to send a packet is chosen:

- 8d : IEEE 802.3ad Link Aggregation Control Protocol (LACP)
- sd : standard
- rn : round-robin

The second value is the hash mode:

- st : src_dst_port - balance based on TCP port (Layer 4)
- dt : default - required for round-robin mode

MTU. The maximum transmission unit is set in the port aggregation logical interface (en6) on the server:

- S indicates standard, which is MTU 1500
- J indicates jumbo frames, which is MTU 9000

Note that, in the switch, for jumbo frames, the corresponding MTU setting is 9216. As noted in the results, in run 18, the MTU was set to 9216 in the ports but 1500 in the channels.

Flw ctl. Two characters, the first indicating switch flow control setting and the second indicating whether flow control is enabled or disabled in the port aggregation NICs. For the switch:

- S : send = desired and receive = off (default)
- B : send = desired and receive = desired
- N : send = off and receive = off

For the port aggregation NICs on the server:

- D : Disabled
- E : Enabled

At the switch, channel groups were defined for each pair of ports corresponding to (connected to) server ports included in the port aggregation. For example, here is the set of channel groups for runs with LACP channeling:

```
show etherchannel summary
Flags:      D - down    P - in port-channel
      I - stand-alone   s - suspended
      H - Hot-standby (LACP only)
      R - Layer3        S - Layer2
      U - in use        f - failed to allocate aggregator
      u - unsuitable for bundling
Number of channel-groups in use: 4
Number of aggregators: 4
Group Port-channel Protocol Ports
------+-------------+-----------+--------------------------------------
---------
1 [N/A]
2 [N/A]
3 Po3(SD) LACP Gi8/4(D) Gi8/5(D)
4 Po4(SD) LACP Gi8/26(D) Gi8/27(D)
```

The configuration of the AIX port aggregation was set consistently with the port aggregation protocol on the switch:

- for PAgP , manual and none: mode = *standard* and hash_mode = *src_dst_port* except for one run with mode = *round_robin* and hash_mode = *default*
- for LACP: mode = *8023ad* and hash_mode = *src_dst_port*

Thus, at all times, load should be balanced across the interfaces in the sending port aggregation (outbound traffic).

9.2.2 Measurements

Measurements from the 42 runs are shown below. Measured data included:

CPU utilization. CPU utilization on the client and on the server. Both user time and system time are shown.

Balance. Ratio of number of packets through the first network port in each port aggregation. This measures degree of load-balancing within the channel. A value of 0.5 indicates that the first port carried 50% of the traffic, and therefore that the workload was fully load-balanced across the two ports. A value of 1.00 indicates that the first port carried 100% of the traffic, and therefore that there was no load-balancing across the two ports.

Mbits / sec. The throughput, in Mbits per second aggregate (transmit plus receive) through each port aggregation. This is the primary measurement.

Measurements are presented once in full, and then as several subsets, each grouped by certain configuration attributes, to show the effect of varying one other attribute. For example, the first subset shows the effect of the different settings of the load balancing algorithm in the switch, taking as examples three pairs of runs in which, within each pair, load balancing was the only variation. To simplify the later tables, CPU utilization statistics are included only if they are discussed.

After each subset brief observations are made. Further analysis is provided in section 9.4 Discussion of results.

9.2.3 Measurements in full

	Configuration						Measurements						
							CPU utilization				Balance		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	cli usr	cli sys	serv usr	serv sys	cli xmit ratio	cli recv ratio	Mbits / sec
1	s	8	i - L - v	8d - st	J	SE	4.8	8.9	4.9	9.1	0.91	1.00	1904
2	s	8	p - L - v	8d - st	J	SE	8.8	18.4	9.5	18.9	0.50	0.50	3791
3	s	8	p - P - d	sd - st	J	SE	8.4	18.1	8.7	18.6	0.49	0.49	3252
4	s	8	p - M - o	sd - st	J	SE	8.9	18.4	9.3	18.8	0.50	0.50	3779
5	s	8	p - P - d	sd - st	J	SE	6.6	14.4	7.0	14.7	0.49	0.49	2342
6	s	8	p - P - d	sd - st	J	SE	7.6	16.5	8.8	16.8	0.50	0.50	2874
7	s	8	p - M - o	8d - st	J	SE	10.7	23.0	11.0	23.4	0.51	0.51	3257
8	p	8	p - M - o	8d - st	J	SE	10.5	19.6	10.7	20.0	0.50	0.50	3203
9	e	8	p - M - o	8d - st	J	SE	12.2	24.3	12.6	25.2	0.51	0.51	3341
10	e	8	p - M - o	sd - st	J	SE	12.4	24.7	12.5	25.1	0.50	0.50	3398
11	e	12	p - M - o	sd - st	J	SE	15.6	35.1	16.4	35.7	0.50	0.51	3947
12	e	16	p - M - o	sd - st	J	SE	14.4	65.1	13.9	72.5	0.50	0.50	3951

13	e	12	p - M - o	rn - dt	J	SE	14.1	31.1	14.1	30.6	0.50	0.50	3485
14	e	12	p - M - o	sd - st	J	SE	15.6	35.2	16.4	36.3	0.50	0.52	3947
15	e	12	p - M - o	sd - st	S	SE	14.8	43.6	14.9	45.3	0.50	0.50	3686
16	e	12	p - M - o	8d - st	S	SE	14.8	43.8	14.9	45.3	0.50	0.50	3686
17	e	12	p - L - v	8d - st	J	SE	15.6	35.0	16.4	35.9	0.50	0.51	3946
18	e	12	p - P - d	sd - st	J	SE	12.0	26.2	12.1	26.1	0.50	0.50	2806[1]
19	e	12	p - P - d	sd - st	J	SE	13.8	31.6	14.2	32.2	0.49	0.49	3406
20	e	12	p - P - d	sd - st	S	SE	13.3	66.6	13.3	67.7	0.50	0.50	3330
21	e	12	p - L - v	8d - st	S	SE	14.8	43.7	14.9	44.5	0.50	0.50	3686
22	e	12	i - L - v	8d - st	J	SE	9.3	15.8	9.3	16.6	0.02	1.00	1969
23	e	12	i - P - d	sd - st	J	SE	12.0	27.4	12.4	28.0	0.50	0.51	2915
24	e	12	p - N - n	8d - st	J	SE	12.6	28.4	12.8	28.1	0.45	0.43	3006
25	e	12	p - N - n	sd - st	J	SE	12.5	27.3	12.8	27.6	0.47	0.46	2954
26	e	12	p - N - n	8d - st	J	SE	13.9	31.9	14.3	32.5	0.49	0.50	3462[2]
27	e	12	p - N - n	sd - st	S	SE	13.2	66.0	13.3	67.6	0.50	0.50	3346
28	e	16	p - N - n	sd - st	S	SE	11.8	78.2	11.9	79.1	0.50	0.50	3328
29	e	12	p - M - o	sd - st	J	SE	15.6	35.4	16.4	36.4	0.50	0.51	3948
30	e	16	p - M - o	sd - st	J	SE	13.9	69.0	13.7	73.1	0.50	0.50	3952
31	e	12	p - L - v	8d - st	S	SE	14.9	43.8	15.0	45.5	0.50	0.50	3686
32	e	16	p - L - v	8d - st	S	SE	15.3	52.0	16.3	53.8	0.50	0.50	3686
33	e	12	p - M - o	sd - st	S	SE	14.8	43.7	15.0	45.3	0.50	0.50	3686
34	e	16	p - M - o	sd - st	S	SE	15.3	51.8	16.3	53.7	0.50	0.50	3687
35	e	12	p - M - o	sd - st	J	SE	15.6	35.1	16.4	36.0	0.50	0.51	3948
36	e	12	p - M - o	sd - st	J	BE	15.6	35.3	16.3	35.8	0.50	0.51	3947
37	e	16	p - M - o	sd - st	J	BE	13.8	68.9	13.7	73.1	0.50	0.50	3952
38	e	12	p - M - o	sd - st	J	ND	15.6	35.3	16.4	36.5	0.50	0.51	3950
39	e	12	p - L - v	8d - st	S	ND	14.8	43.8	14.9	45.3	0.50	0.50	3686
40	e	16	p - L - v	8d - st	S	ND	15.3	51.6	16.1	53.5	0.50	0.50	3687

41	e	12	p - L - v	8d - st	S	BE	14.8	43.9	15.0	45.5	0.50	0.50	3686
42	e	16	p - L - v	8d - st	S	BE	15.3	52.0	16.2	53.3	0.50	0.50	3685

[1] MTU 9216 set in the Gigabit Ethernet ports but not set in the channels. In all other runs with MTU 9000, MTU 9216 was set in the Gigabit Ethernet ports and in the channels.
[2] LACP aggregation mode set in Gigabit Ethernet ports with no channel defined. In all other runs, the Gigabit Ethernet ports had no explicit aggregation mode.

Table 1: Measurements in full for test system 1

9.2.4 Measurements comparing switch load balancing

	Configuration						Measurements		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	Balance		Mbits / sec
							cli xmit ratio	cli recv ratio	
1	s	8	i - L - v	8d - st	J	SE	0.91	1.00	1904
2	s	8	p - L - v	8d - st	J	SE	0.50	0.50	3791
22	e	12	i - L - v	8d - st	J	SE	0.02	1.00	1969
17	e	12	p - L - v	8d - st	J	SE	0.50	0.51	3946
23	e	12	i - P - d	sd - st	J	SE	0.50	0.51	2915
19	e	12	p - P - d	sd - st	J	SE	0.49	0.49	3406

Table 2: Measurements comparing switch load balancing for test system 1

Clearly, the use of *src_dest_port* (Layer 4) load balancing improves throughput greatly for this configuration and workload. In the case of the LACP protocol, the improvement is 100%.[6] This is simply the effect of using both interfaces for switch-to-server traffic. Recall that this configuration has a total of two IP addresses, one for the port aggregation on each server, and therefore hashing on IP addresses cannot balance across the interfaces inside the port aggregation. Note: In order to measure packet balance in a port aggregation, use the command:

```
entstat –d ent<nn>
```

where ent<nn> is the port aggregation device.
For example

```
  entstat –d ent6
```

9.2.5 Measurements comparing switch protocol and port mode

Note that switch protocol and port mode are strongly inter-related with the port aggregation mode on the server, and although some observations can be drawn about each separately, they should be viewed as a single tunable.

	Configuration						Measurements		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	Balance		Mbits / sec
							cli xmit	cli recv	

[6] A much smaller improvement is seen for PAgP because results for PAgP with Layer 3 load balancing and port mode *desirable non-silent* (run 23) were much better than expected. The reason for this result is not known.

							ratio	ratio	
2	s	8	p - L - v	8d - st	J	SE	0.50	0.50	3791
7	s	8	p - M - o	8d - st	J	SE	0.51	0.51	3257
4	s	8	p - M - o	sd - st	J	SE	0.50	0.50	3779
3	s	8	p - P - d	sd - st	J	SE	0.49	0.49	3252
22	e	12	i - L - v	8d - st	J	SE	0.02	1.00	1969
23	e	12	i - P - d	sd - st	J	SE	0.50	0.51	2915
17	e	12	p - L - v	8d - st	J	SE	0.50	0.51	3946
11	e	12	p - M - o	sd - st	J	SE	0.50	0.51	3947
24	e	12	p - N - n	8d - st	J	SE	0.45	0.43	3006
26	e	12	p - N - n	8d - st	J	SE	0.49	0.50	3462[2]
25	e	12	p - N - n	sd - st	J	SE	0.47	0.46	2954
19	e	12	p - P - d	sd - st	J	SE	0.49	0.49	3406
41	e	12	p - L - v	8d - st	S	BE	0.50	0.50	3686
16	e	12	p - M - o	8d - st	S	SE	0.50	0.50	3686
20	e	12	p - P - d	sd - st	S	SE	0.50	0.50	3330
40	e	16	p - L - v	8d - st	S	ND	0.50	0.50	3687
34	e	16	p - M - o	sd - st	S	SE	0.50	0.50	3687
28	e	16	p - N - n	sd - st	S	SE	0.50	0.50	3328

[2] LACP aggregation mode set in Gigabit Ethernet ports with no channel defined. In all other runs, the Gigabit Ethernet ports had no explicit aggregation mode.

Table 3: Measurements comparing switch protocol and port mode for test system 1

The first group shows that, when the load-balancing policy is *src_dest_ip*, (Layer 3), then the PAgP protocol produces a throughput higher than that of LACP and also, remarkably, higher than that of a single interface. In other words, to some extent, it compensates for the limitations of balancing on IP address.

The remaining groups all show, for different configurations, that the LACP-active-with-8023ad-in-the-port-aggregation and manual-on-with-standard-port-aggregation combinations are superior to other combinations. In particular, it is always preferable to create channel groups explicitly, as the variations in which no channel groups were created all performed worse. This is the case even when the port aggregation mode is 802.3ad. In some documents, it is asserted that use of 802.3ad aggregation mode avoids the need to configure the switch to specify which ports belong to the same aggregation; this is true, but doing so carried a performance penalty in the case of the system tested here.

9.2.6 Measurements comparing port aggregation mode on the server

	Configuration						Measurements		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	Balance		Mbits / sec
							cli xmit ratio	cli recv ratio	
13	e	12	p - M - o	rn - dt	J	SE	0.50	0.50	3485
11	e	12	p - M - o	sd - st	J	SE	0.50	0.51	3947
16	e	12	p - M - o	8d - st	S	SE	0.50	0.50	3686
15	e	12	p - M - o	sd - st	S	SE	0.50	0.50	3686
32	e	16	p - L - v	8d - st	S	SE	0.50	0.50	3686

34	e	16	p - M - o	sd - st	S	SE	0.50	0.50	3687

Table 4: Measurements comparing port aggregation mode on the server for test system 1

The first group demonstrates that the round-robin mode is inferior. Otherwise, choice of port aggregation mode on the server and choice of switch protocol and port mode go hand-in-hand. In these comparisons, the combinations of LACP-active-with-8023ad-in-the-port-aggregation and manual-on-with-standard-port-aggregation perform identically.

9.2.7 Measurements comparing standard and jumbo frames

	Configuration						Measurements		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	Balance		Mbits / sec
							cli xmit ratio	cli recv ratio	
17	e	12	p - L - v	8d - st	J	SE	0.50	0.51	3946
21	e	12	p - L - v	8d - st	S	SE	0.50	0.50	3686
11	e	12	p - M - o	sd - st	J	SE	0.50	0.51	3947
15	e	12	p - M - o	sd - st	S	SE	0.50	0.50	3686
25	e	12	p - N - n	sd - st	J	SE	0.47	0.46	2954
27	e	12	p - N - n	sd - st	S	SE	0.50	0.50	3346
18	e	12	p - P - d	sd - st	J	SE	0.50	0.50	2806[1]
19	e	12	p - P - d	sd - st	J	SE	0.49	0.49	3406
20	e	12	p - P - d	sd - st	S	SE	0.50	0.50	3330
[1] MTU 9216 set in the Gigabit Ethernet ports but not set in the channels. In all other runs with MTU 9000, MTU 9216 was set in the Gigabit Ethernet ports and in the channels.									

Table 5: Measurements comparing standard and jumbo frames for test system 1

Use of jumbo frames yields better throughput except for the case of no port channels defined in the switch, when non-jumbo is preferable.

The lower throughout of the run 18 in the last group demonstrates the importance of setting the correct MTU in all five required interfaces, both physical and logical: the AIX physical network interface, the AIX physical port-aggregation, the AIX logical port-aggregation, the switch Gigabit Ethernet port, and the switch port channel. Setting them in some but not all interfaces can result in errors and/or poor performance, which can be difficult to diagnose. Beginning with AIX 5.2, when a port aggregation is created, the “jumbo_frames” attribute of the underlying adapters will automatically be set to “yes” when jumbo frames are enabled on the port aggregation (configuration will fail if any of the underlying adapters does not have this attribute); and the “mtu” attribute of the adapters will also be automatically set to “9000”. However, it is important to verify that all the devices and interfaces are set correctly.

Note that UDP-based networking applications, such as name servers, will fail silently with packets dropped if jumbo frames are enabled at a sender but not at a receiver.

9.2.8 Measurements comparing flow-control

	Configuration						Measurements		
							Balance		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	cli xmit ratio	cli recv ratio	Mbits / sec
36	e	12	p - M - o	sd - st	J	BE	0.50	0.51	3947
38	e	12	p - M - o	sd - st	J	ND	0.50	0.51	3950
11	e	12	p - M - o	sd - st	J	SE	0.50	0.51	3947
41	e	12	p - L - v	8d - st	S	BE	0.50	0.50	3686
39	e	12	p - L - v	8d - st	S	ND	0.50	0.50	3686
31	e	12	p - L - v	8d - st	S	SE	0.50	0.50	3686

Table 6: Measurements comparing flow-control for test system 1

These results show that flow control had no effect for this workload, probably because both the server and the switch are well-matched and both have very high processing speeds relative to the network traffic. CPU utilization on the server was never higher than around 85%. It would be interesting to re-assess flow control with a 10 Gigabit network. Another reason that flow control had no effect was the use of larger full size frames. Flow control can become more important with smaller packets, where the packet rate can be much higher. The impact of flow control also depends on what other adapters are on the same PHB (PCI host bridge) and what other adapters may be sharing the same DMA path into memory. Likewise, dual-port or 4-port adapters are more likely to benefit from flow control because there are more ports competing for the same path to memory.

9.2.9 Measurements comparing application method of waiting for TCP/IP events

	Configuration						Measurements		
							Balance		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	cli xmit ratio	cli recv ratio	Mbits / sec
9	e	8	p - M - o	8d - st	J	SE	0.51	0.51	3341
8	p	8	p - M - o	8d - st	J	SE	0.50	0.50	3203
7	s	8	p - M - o	8d - st	J	SE	0.51	0.51	3257

Table 7: Measurements comparing application method of waiting for TCP/IP events for test system 1

The pollset method with edge-triggered events (non-blocking sockets) performed slightly better than the other methods. As for the comparison of flow-control, the power of these servers renders such efficiencies in the application less important for this workload. It would have been interesting to compare the application method at the higher number of processes but there was insufficient time.

9.2.10 Measurements comparing application number of processes

	Configuration						Measurements						
							CPU utilization				Balance		
Run num	App meth	Num procs	Switch config	Server config	MTU	Flw ctl	cli usr	cli sys	serv usr	serv sys	cli xmit ratio	cli recv ratio	Mbits / sec
10	e	8	p - M - o	sd - st	J	SE	12.4	24.7	12.5	25.1	0.50	0.50	3398
11	e	12	p - M - o	sd - st	J	SE	15.6	35.1	16.4	35.7	0.50	0.51	3947
12	e	16	p - M - o	sd - st	J	SE	14.4	65.1	13.9	72.5	0.50	0.50	3951

Table 8: Measurements comparing application number of processes for test system 1

The optimal throughput was achieved with 12 processes. Above that number, throughput remained unchanged while CPU utilization increased.

9.3 Configurations and results -- Test system 2

The following 12 runs were completed. In cases where the configuration is described earlier in this paper, the section number in which the configuration is described is given in parentheses. In these runs, the columns have the same meanings as in the preceding tables, with the following exceptions:

Switch config. Three configuration values, separated by hyphens, are given.
The first value is the frame load-balancing policy (frame distribution). Two policies were assessed:

- s session both (Layer 4)
- M MAC both (Layer 2)

The second value is the port aggregation protocol:

- P PAgP
- L LACP
  When the port mode is configured as *on*, the switch ignores the port aggregation protocol and channeling is specified manually by the user by means of administrative groups. For this case, the port aggregation protocol is shown in lower case.

The third value is the port mode. The port mode depends on, and has different meanings, depending on whether the PAgP or LACP protocol is used:

- u : PAgP auto
- d : PAgP desirable
- v : LACP active

- o : on (manual channeling)

One additional metric is shown for test system 2.

Out of order packets / sec. This metric indicates how well the channel provides routing consistency for successive packets of any one connection. That is, for a given connection, if successive packets follow the same route in terms of choice of interface within each port aggregation, then their order is likely to be preserved. If they do not follow the same route, their order may change, resulting in an out-of-order packet.

	Configuration					Measurements			
Run numb	App meth	Num procs	Switch config	Server config	MTU	Out of order packets / sec	Balance		Mbits / sec
							client xmit ratio	client recv ratio	
1	s	8	s - P - u	sd - st	J	21283.4	0.5	0.5	2843.3
2 (7.1.1)	s	8	s - P - d	sd - st	J	21620.2	0.5	0.5	2778.8
3 (7.1.3)	s	8	s - p - o	sd - st	J	1894.9	0.5	0.5	3548.7
4	s	8	s - P - d	sd - st	J	22012.4	0.5	0.5	2814.8
5 (7.2.1)	s	8	M - P - d	sd - st	J	19781.6	0.6	0.6	2689.6
6	s	8	M - p - o	sd - st	J	270.2	0.0	1.0	1824.6
7 (7.2.2)	s	8	M - L - v	8d - st	J	264.1	0.0	1.0	1820.5
8	s	8	M - l - o	8d - st	J	271.2	1.0	1.0	1837.9
9	s	8	s - l - o	8d - st	J	1811.8	0.5	0.5	3603.7
10 (7.1.2)	s	8	s - L - v	8d - st	J	1890.3	0.5	0.5	3611.9
11	e	8	s - L - v	8d - st	J	2220.8	0.5	0.5	3874.8
12	e	8	s - p - o	sd - st	J	2088.4	0.5	0.5	3883.6

Table 9: Measurements for test system 2

9.4 Discussion of results

1. With Layer 4 load balancing, traffic was balanced across both interfaces of the port aggregation on both send and receive, which contributed to a high throughput.
2. For test system 1, the highest throughput was obtained when the switch load-balancing policy was *src_dest_port* (Layer 4) and either one of the preferred combinations of LACP-active-with-8023ad-in-the-port-aggregation or manual-on-with-standard-port-aggregation was set. For test system 2, the highest throughput was obtained when the load balancing policy was *session both* and when either the protocol was set to LACP with port mode *active* or was overridden using port mode *on*. With these settings, maximum throughput was excellent, and extremely close to

the maximum rated capabilities of all the hardware components (2 Gbits/sec concurrently on each of transmit and receive).

3. For test system 1, setting the load-balancing policy to *src_dest_ip*, (Layer 3), the default IOS policy, resulted in total imbalance of traffic in the two interfaces in each port aggregation, with the exception of the case of *PAgP desirable*, where balance was reasonable but not perfect. For test system 2, similar results were obtained when the load-balancing policy was set to MAC both (Layer 2), the default CatOS policy, again with the exception of *PAgP desirable*. Although these are not ideal configurations, in some cases they might be the best configuration available if it is not possible to set the switch load-balancing policy to Layer 4. This could be because either this functionality is not available, or because although available, it applies to the entire switch, and other systems connected to the same switch might be incompatible with Layer 4 load-balancing.
4. Our results on test system 2, in which the switch uses the CatOS operating system, show a strong correlation between a high number of out-of-order packets and a lower throughput, which is expected. To our surprise, we did not see the same correlation on test system 1, in which the switch uses the IOS operating system. For this reason, out-of-order packets are not shown for test system 1. One possible explanation for this result is that the correlation between out-of-order packets and throughput on IOS is more complex than with CatOS and depends on other unknown variables; another possibility is that the **`entstat`** tool reported incorrect data about out-of-order packets.

For a spreadsheet giving additional information about the configurations and results discussed in this article, contact the authors.

9.5 Recommendations

1. Whenever possible, use Layer 4 load-balancing. Assuming that this load-balancing policy can be used, use LACP with port mode *active*.
2. If Layer 4 load-balancing cannot be used because of one of the reasons outlined in the preceding section, use the default load-balancing policy with port mode *PAgP desirable*.
3. Use of jumbo frames is recommended provided that all equipment in the LAN or VLAN supports them. However, as previously discussed, it is important to ensure that the high MTU is set everywhere for jumbo-frame use. Enabling jumbo frames in only some of the relevant interfaces results in lower throughput, but no warning.

10. References

AIX Information Center:
http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp
See in particular:
http://publib.boulder.ibm.com/infocenter/pseries/v5r3/topic/com.ibm.aix.commadmn/doc/commadmndita/etherchannel_intro.htm

CatOS:
Configuration guides:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/catos/6.x/configuration/guide/confg_gd.html
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/catos/8.x/configuration/guide/confg_gd.html
Command references:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/catos/6.x/command/reference/cmd_ref.html
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/catos/8.x/command/reference/cmd_ref.html

IOS:
Configuration guide:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/book.html
Command reference:
Up to IOS release 12.1, all commands are contained in a single book. This book is therefore a handy single point of reference, although possibly out of date relative to later releases for some functionality:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.1E/native/command/reference/comref.html
Beginning with release 12.2, commands are divided into several books, which include:
http://www.cisco.com/en/US/docs/ios/interface/command/reference/ir_book.html
http://www.cisco.com/en/US/docs/ios/12_2/interface/command/reference/finter_r.html

Comparison of CatOS with IOS:
http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/prod_white_paper09186a00800c8441.html

Notices

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM’s current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this publication to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM’s sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, savings or other results.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All performance data contained in this publication was obtained in the specific operating environment and under the conditions described above and is presented as an illustration only. Performance obtained in other operating environments may vary and customers should conduct their own testing. The actual throughput or performance that any user will experience will vary depending upon many factors, including 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 results similar to those stated here.

The information in this document concerning non-IBM products was obtained from the supplier(s) of those products. IBM has not tested such products and cannot confirm the accuracy of the performance compatibility or any other claims related to non-IBM products. Questions about the capabilities of non-IBM products or services should be addressed to the supplier(s) of those products or services.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites.

Trademarks