HP Auto Port Aggregation Performance and Scalability White Paper

Part number: 5555–1245
published February 2006
Edition: 1.0

Table of Contents

List of Figures

Introduction

HP Auto Port Aggregation is a software product that creates link aggregates, often called “trunks,” which provide a logical grouping of two or more physical ports into a single “fat pipe.” Two primary features are automatic link failure detection and recovery as well as support for load balancing of network traffic across all of the links in the aggregation. This enables you to build large bandwidth logical links into the server that are highly available and completely transparent to the client and server applications.

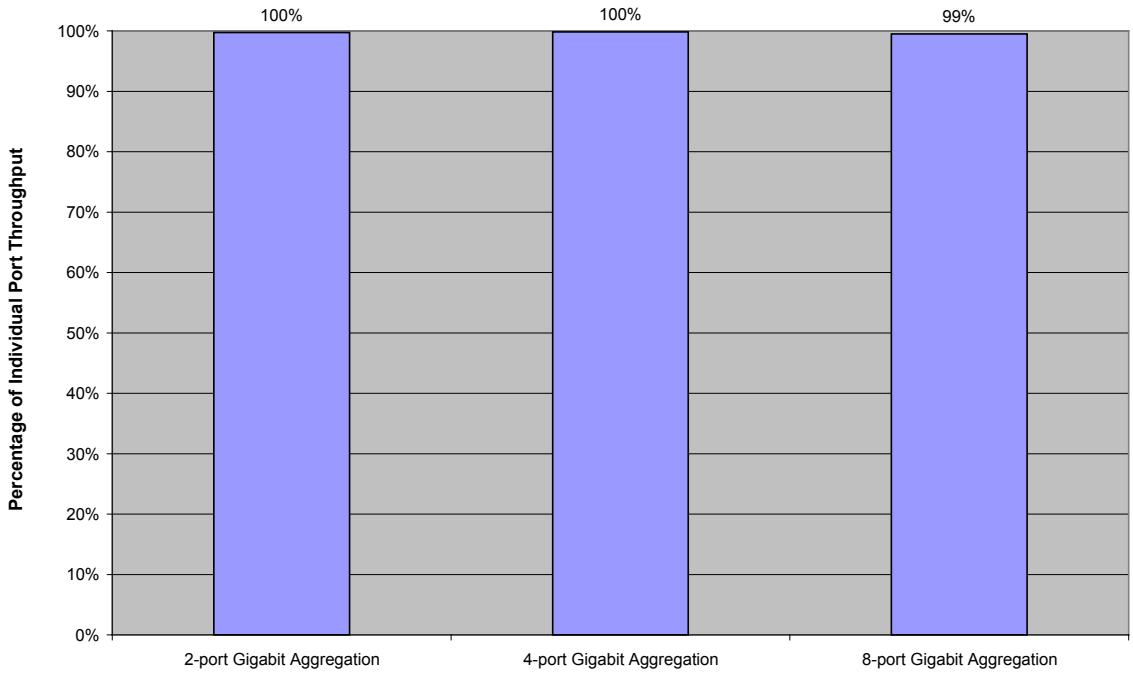
This white paper presents the exceptional performance and scalability testing results for the HP Auto Port Aggregation (HP APA) product. It also provides system setup and configuration recommendations to help you achieve similar results to suit your business needs.

Executive Summary

The HP APA product, when set up and used in the configurations described in this white paper, consistently provided sustained link rate performance for various workloads with two, four, and eight Gigabit Ethernet ports in a single link aggregate.

The exceptional performance results presented in this paper were measured on a 16-processor HP Integrity rx8620 server running HP-UX 11i v2 and other software performance and scalability enhancements to the networking stack. The high-performance HP ProCurve 3400cl-24G switch has the switching capacity needed to sustain the heavy loads generated by these tests.

The following figure shows the sustained link rate throughput of three different link aggregates under a unidirectional workload. Throughput is a measure of how well programs run with a specific workload and how quickly the programs can service user requests.



You can expect to see similar results on other HP servers running HP-UX, including those running HP-UX 11i v1 with the corresponding performance and scalability enhancements to the networking stack.

NOTE In this paper, the unidirectional link rate of a single Gigabit port is 948 Mb/s. When multiple Gigabit ports are mentioned, the unidirectional link rate is taken as n*948 Mb/s, where n is the number of unaggregated individual Gigabit Ethernet ports used in the comparison.

Benchmark Results

The HP APA link aggregates provided outstanding scalable performance for various workloads when used in accordance with the recommendations in "Benchmark Components" (page 11) and "Configuration " (page 9).

Link rate throughput is maintained for link aggregates with 2, 4, and 8 Gigabit Ethernet ports. A 2-port aggregate has a sustained unidirectional throughput equal to 2 individual ports, a 4-port aggregate has a sustained throughput equal to 4 individual ports, and an 8-port aggregate has a sustained throughput equal to 8 individual ports.

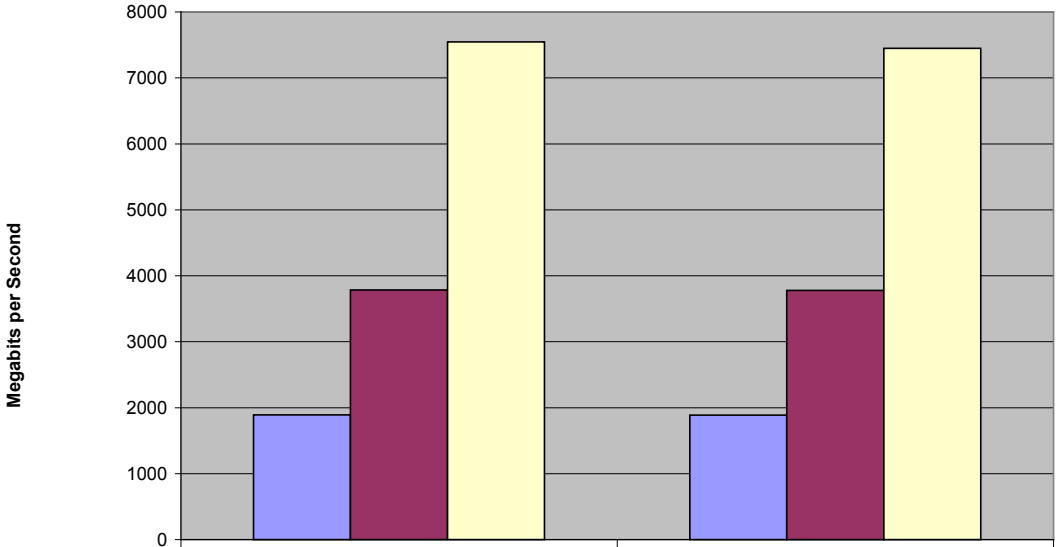
This section presents the following performance data:

- Unidirectional throughput with HP APA
- Bidirectional throughput with HP APA

Unidirectional Throughput Performance

Figure 1 shows HP APA throughput numbers with unidirectional transmit and receive workloads for 2-, 4-, and 8-port link aggregates.

Figure 1 HP APA Throughput for Unidirectional Transmit and Receive Workloads



	Transmit	Receive
2-port HP APA	1891	1887
4-port HP APA	3786	3779
8-port HP APA	7546	7448

HP APA aggregates provide sustained link rate throughput for the unidirectional transmit and receive workloads, for any number of Gigabit links in the aggregate, up to the maximum of 8 ports in an aggregate. With 8 ports in the link aggregate, the throughput is an outstanding 7.5 Gb/s for unidirectional transmit and receive workloads.

During the unidirectional transmit tests, the CPU utilization (normalized across all 16 CPUs on the server) was about 7 percent, 16 percent, and 39 percent for 2-, 4-, and 8-port link aggregates, respectively. The packet rate was approximately 82K packets per second per port.

During the unidirectional receive tests, the CPU utilization (normalized across all the 16 CPUs on the server) was about 7 percent, 16 percent, and 37 percent for 2-, 4-, and 8-port link aggregates, respectively. The average packet rate was approximately 82K packets per second per port.

Bidirectional Throughput Performance

Figure 2 shows HP APA throughput numbers with bidirectional transmit and receive workloads for 2-, 4-, and 8-port link aggregates.

Figure 2 HP APA Throughput for Bidirectional Transmit and Receive Workloads

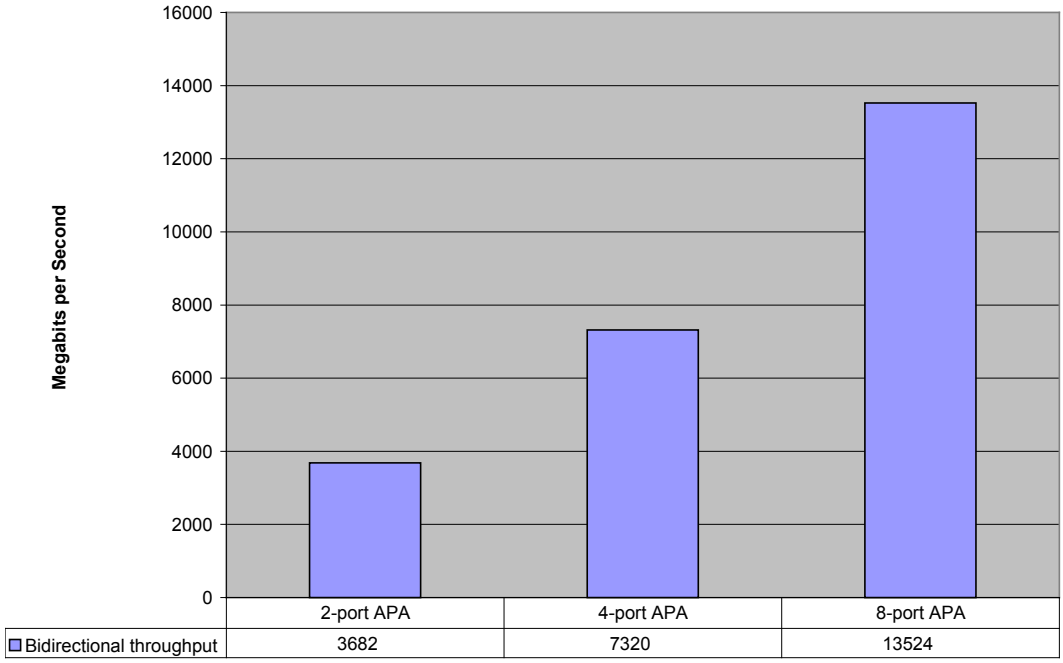


Figure 2 shows that HP APA provides an exceptional sustained data rate of 3.68 Gb/s for a 2-port aggregate and 7.3 Gb/s for a 4-port aggregate, which is 97 percent of link rate. The 8-port aggregate achieved an excellent sustained data rate of 13.5 Gb/s, which is close to 90 percent of link rate.

For these tests, the server CPU utilization (normalized across all 16 CPUs on the server) was approximately 20 percent, 41 percent, and 92 percent for 2-port, 4-port, and 8-port link aggregates, respectively. The average packet rate was around 154K packets per second per port.

Configuration

For the best performance and scalability with HP APA, HP recommends the following:

- Use processor affinity for processes that generate networking traffic over the HP APA link aggregate. For best results, bind a work process to the CPU that handles most of the inbound traffic for that process as follows:
  1. Find the APA link aggregate ports on the server that are used for processing the inbound traffic for a process. The load balancing algorithm for the trunk on the switch determines this. For information on the load balancing algorithms used for trunk configurations on the switch, see the documentation for your switch.
  2. Ensure that the inbound processing for all workloads is evenly distributed across all processors on the server. Be sure that no two ports have the same CPU doing all the inbound processing for both ports. Ideally, assign all ports unique interrupt CPUs. If necessary, change the interrupt CPU configuration for ports from the default configuration. Use the `intctl` command to change the interrupt CPU configuration. See intctl(1M) for more information.
  3. Set the processor affinity when starting the process or change the processor affinity for an already running process, by using the `mpsched` command. See mpsched(1M) for more information.

NOTE If the work processes are not explicitly configured for processor affinity, you might see a noticeable increase in CPU utilization. In addition, you might also see a decrease in the aggregate

throughput or the data rate for the networking work load, especially with link aggregates using four or more Gigabit Ethernet ports.

- Set the load balancing algorithm for the HP APA aggregate on the server to ensure that the outbound traffic server is evenly balanced.

  Use the `lanadmin` command to set the outbound load balancing algorithm used by HP APA. HP APA provides a variety of load balancing algorithms to suit the individual needs of your configuration. See the *HP Auto Port Aggregation (APA) Support Guide* for more information. The link to HP APA documentation are listed in "For More Information" (page 13).

- Configure the switch so that the inbound traffic to the server is evenly balanced across all the ports that form the HP APA link aggregate on the server. The aggregation mode of the HP APA link aggregate must match the aggregation mode of the ports on the switch. See your switch configuration manuals for information on how to setup link aggregation trunks on the switch.
- Install the HP A7012A, PCI-X 2-port 1000Base-T Gigabit Ethernet adapters (used to generate the performance results) in the highest performance PCI-X 133MHz slots on the server.

  Also, when forming 4- or 8-Gigabit port HP APA aggregates, evenly distribute the ports among the two internal I/O bays on the rx8620. See your installation manuals for the location of the two internal I/O bays and the PCI-X 133MHz slots in these I/O bays.

  For the 4-port link aggregate tests, the server system used two PCI-X 2-port 1000Base-T Gigabit Ethernet adapters with one adapter in each of the two internal I/O bays in PCI-X 133MHz slots.

  For the 8-port link aggregate tests, the server system used four PCI-X 2-port 1000Base-T Gigabit Ethernet adapters with two adapters in each of the two internal I/O bays in PCI-X 133MHz slots.

- Install software on the server to enhance performance and scalability of the networking stack.

  For HP-UX 11i v2, install the following:

  - Cumulative LAN patch PHNE_33429 or a superseding patch. You can download this patch from the HP Information Technology Resource Center (HP ITRC) Web site at:

    http://www2.itrc.hp.com/service/patch/mainPage.do

  - Transport Optional Upgrade Release (TOUR) version 3.0 for 11i v2. You can download this software for free from **http://software.hp.com**.
  - Cumulative STREAMS patch PHNE_32277 or a superseding patch. This enables the STREAMS enhancements already in TOUR 3.0 bundle for 11i v2. You can download this patch from the HP Information Technology Resource Center (HP ITRC) Web site at:

    http://www2.itrc.hp.com/service/patch/mainPage.do

  - HP IP Filter product version A.03.05.12. You can download this product for free from **http://software.hp.com** or from the AR0512 CD.
  - HP APA product version B.11.23.10. This product is available on the AR0512 CD. This version (or a succeeding version) of HP APA is required only if you want to aggregate more than 4 ports. The previous versions of HP APA already support up to 4 ports in a link aggregate.
  - IETHER driver patch PHNE_31118 or a superseding patch for the HP A7012A, PCI-X 2-port 1000Base-T Gigabit Ethernet adapter. You can download this patch from the HP Information Technology Resource Center (HP ITRC) Web site:

    http://www2.itrc.hp.com/service/patch/mainPage.do

  For HP-UX 11i v1, install the following:

  - Cumulative LAN patch PHNE_33704 or a superseding patch. You can download this patch from the HP Information Technology Resource Center (HP ITRC) Web site at:

    http://www2.itrc.hp.com/service/patch/mainPage.do

  - Transport Optional Upgrade Release (TOUR) Version 3.0 for 11i v1. You can download this software for free from **http://software.hp.com**.

- Cumulative STREAMS patch PHNE_33313 or a superseding patch. This enables the STREAMS enhancements already in the TOUR 3.0 bundle for 11i v1. You can download this patch from the HP Information Technology Resource Center (HP ITRC) Web site at:

  http://www2.itrc.hp.com/service/patch/mainPage.do

- If you have the HP IP Filter product installed on your system, upgrade to the latest version. You can download the latest version for free at:

  http://software.hp.com

- IETHER driver patch PHNE_31153 or a superseding patch for the HP A7012A, PCI-X 2-port 1000Base-T Gigabit Ethernet adapter. You can download this patch from the HP Information Technology Resource Center (HP ITRC) Web site at:

  http://www2.itrc.hp.com/service/patch/mainPage.do

- Interrupt Migration enhancement release for HP-UX 11i v1. You can download this patch bundle for free at:

  http://software.hp.com

  After you install this patch bundle, use the `intctl` command to change the default interrupt configuration on the server. See intctl(1M) for more information.

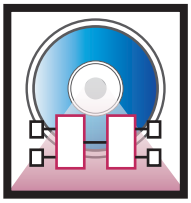
Benchmark Components

This section describes the hardware and software used to achieve the performance data.

Server (System Under Test)

The test was conducted on an HP Integrity rx8620 server with sixteen 1.5 GHz Intel® Itanium® 2 processors with 6 MB L3 unified cache, 32 GB (gigabytes) of memory with 2 internal I/O bays. The server ran HP-UX 11i v2 May 2005 Operating Environment Upgrade release (OEUR) and the performance and scalability enhancements to the networking stack listed in "Configuration " (page 9).

HP APA Software



The HP APA aggregates on the server were the default manual mode aggregates using the default MAC address-based load balancing algorithm.

Clients (Systems That Drove the Test Load)

Eight HP Integrity rx2600 clients, each with two 1.5 GHz Intel® Itanium® 2 processors with 6 MB L3 unified cache and 2GB memory. The clients ran HP-UX 11i v2. Each client used a single Gigabit Ethernet adapter to drive the test workload on the server.

Gigabit Ethernet Adapters

The rx8620 server had four HP A7012A, PCI-X 2-port 1000Base-T Gigabit Ethernet adapters placed in highest performing PCI-X 133 MHz slots. The cards were placed in slots numbered 4 and 5 in each of the two internal I/O bays.

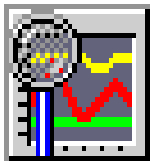
Each rx2600 client used one HP A6825A, PCI 1-port 1000Base-T adapter in PCI slot 4 to drive the tests. All Gigabit Ethernet cards were configured with a maximum transmission unit (MTU) size of 1500 bytes and with both transmit and receive checksum offload capabilities enabled.

NOTE The TCP Segmentation Offload (TSO) feature was disabled on all Gigabit Ethernet cards on the server. Expect better server utilization if you enable TSO.

Switch

An HP ProCurve 3400cl-24G switch connected all 1000Base-T Gigabit Ethernet ports on the server and clients. The switch firmware version was M.08.54. The server ports on the switch were trunked to form 2-, 4-, or 8-port aggregates for the tests.

Benchmarking Software



HP generated test results using the `netperf` Version 2.4.1 benchmarking software and the following options:

`-C local_CPU_rate`	Specifies to gather local CPU utilization for the test. You can calibrate the CPU rates separately using LOC_CPU and REM_CPU tests. HP used `-C 1.5e+09` for our tests because both server and clients used 1.5 GHz CPUs.
`-c remote_CPU_rate`	Specifies to gather remote CPU utilization for the test. The CPU rates can be calibrated separately using LOC_CPU and REM_CPU tests. HP used `-c 1.5e+09` for our tests because both server and clients used 1.5 GHz CPUs.
`-H remoteIP`	Specifies the remote client IP address.
`-I time`	Specifies duration (in seconds) of the test. HP typically used 60 seconds per test run.
`-S socket_buffer_size`	Specifies the remote socket buffer sizes. HP used 128K buffer sizes.
`-s socket_buffer_size`	Specifies the local socket buffer sizes. HP used 128K buffer sizes.
`-T localCPU, remoteCPU`	Allows the binding of `netperf/netserver` processes on the local (server) and remote (client) CPUs.
`-t test_type`	Use TCP_STREAM for transmit tests and TCP_MAERTS for receive tests. Transmit and receive tests were run concurrently to generate bidirectional workloads. The TCP_MAERTS test is new. You can find more information about this test at: **http://www.netperf.org/netperf/NetperfNew.html**
`-m message_size`	Specifies the message size for the test. We used 32K message size.

You can find more information about `netperf`, documentation, and obtain a free copy at:

http://www.netperf.org

For More Information

- HP APA documentation:
  http://docs/hp/com/en/netcom.html#Auto%20Port%20Aggregation%20%28APA%29
- Transport Optional Upgrade Release Version 3.0
  For information:
  http://docs.hp.com/en/5991-4436/index.html
  For a free software download:
  http://h20293.www2.hp.com/portal/swdepot/displayProductInfo.do?productNumber=TOUR
- IP Filter product information and free download:
  http://h20293.www2.hp.com/portal/swdepot/displayProductInfo.do?productNumber=B9901AA
- HP ProCurve Network products:
  http://www.hp.com/rnd/index.htm