Small Tree Communications

Link Aggregation White Paper
(Based on IEEE 802.3ad)

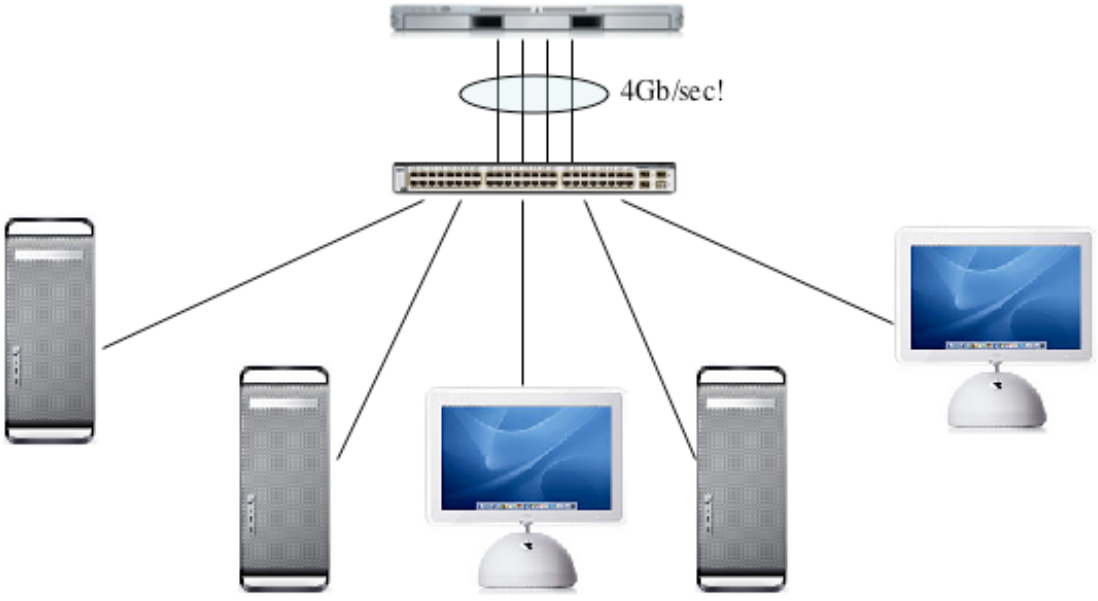
Link Aggregation (also known as trunking or bonding) is a software only mechanism that allows an administrator to combine multiple Ethernet ports into one logical port. When combined in this way, the logical port that is created provides several benefits:

Higher Bandwidth – The capacity of multiple links is combined into one logical link.

Higher Availability – If one physical port is lost for any reason (a cable is unplugged, a switch port fails, or the card itself fails) the logical port will transparently continue to function over the remaining physical ports.

More Efficient Utilization – All traffic to or from the logical port is load balanced over all of the available physical ports.

Using Small Tree Communication's Link Aggregation software, an administrator can increase the available bandwidth and reliability of the network link between an Apple G5 server and the switch. This allows for larger client loads per system with a longer mean time between failures. One can even remove a port from the link aggregation group in order that the machine can be put onto a new subnet. In this way, the network configuration of the system can be altered without the machine being taken off the network.



About Small Tree Communications

Small Tree Communications LLC was founded by a talented group of high performance networking and kernel engineers that recognized a unique opportunity in the Apple G5 platform.

We decided that rather than work for proprietary Unix vendors who are steadily losing market share, or arguing with open source Linux developers over whether it was important to have zero-copy networking for sockets, we found in Apple a company that's focused on its customers, provides the features they want, and has struck a balance between making their source available to the community while maintaining enough control to assure they could react to the business needs of their customers.

Small Tree Communications has created a range of solutions for your G5 and Xserve platforms including 10Gb Ethernet, Dual port Gigabit Ethernet and 802.3ad Link Aggregation. All designed so that you (our customers) can finally buy an enterprise server that works like you want it to work, is easy to use, and is kind of stylish too!

Small Tree Communications has offices in California, Minnesota and Wisconsin. You can find out more about us by visiting http://www.small-tree.com

Goals and Objectives of 802.3ad

"Link Aggregation allows one or more links to be aggregated together to form a Link Aggregation Group, such that a MAC client can treat the Link Aggregation Group as if it were a single link" (from IEEE Standard 802.3, 2000 Edition, page 1215).

The standard lists the following main goals and objectives for Link Aggregation (from IEEE Standard 802.3, 2000 Edition, page 1215):

- Increased bandwidth
  - The capacity of multiple links is combined into one logical link.
- Increased availability
  - The failure or replacement of a single link within a Link Aggregation Group need not cause failure from the perspective of a MAC Client.
- Linearly incremental bandwidth
  - Bandwidth can be increased in unit multiples as opposed to the order-of-magnitude increase available through Physical Layer technology options (10 Mb/s, 100 Mb/s, 1000 Mb/s, etc.).
- Load sharing
  - MAC Client traffic may be distributed across multiple links. Automatic configuration In the absence of manual overrides, an appropriate set of Link Aggregation Groups is automatically configured, and individual links are allocated to those groups.
- Rapid configuration and reconfiguration
  - In the event of changes in physical connectivity, Link Aggregation will quickly converge to a new configuration, typically on the order of 1 second or less.
- Deterministic behavior
  - Depending on the selection algorithm chosen, the configuration can be made to resolve deterministically; i.e. the resulting aggregation can be made independent of the order in which events occur, and be completely determined by the capabilities of the individual links and their physical connectivity.
- Low risk of duplication or incorrect ordering of frames
  - During both steady-state operation and link (re-) configuration, there is a high probability that frames are neither duplicated nor re-ordered.
- Support of existing IEEE 802.3 MAC Clients (frames transmitted are ordinary MAC frames)
  - No change is required to existing higher-layer protocols or application to use Link Aggregation.
- Backwards compatibility with aggregation-unaware devices
  - Links that cannot take part in Link Aggregation - either because of their inherent capabilities, management configuration, or the capabilities of the devices to which they attach – operate as normal, individual IEEE 802.3 links.
- Accommodation of differing capabilities and constraints

  - Devices with differing hardware and software constraints on Link Aggregation are, to the extent possible, accommodated.
- No change to the IEEE 802.3 frame format
  - Link Aggregation neither adds to, nor changes the contents of frames exchanged between MAC Clients.
- Network Management Support
  - The standard specifies appropriate management objects for configuration, monitoring, and control of Link Aggregation.

Link Aggregation, according to IEEE 802.3, does not support the following:

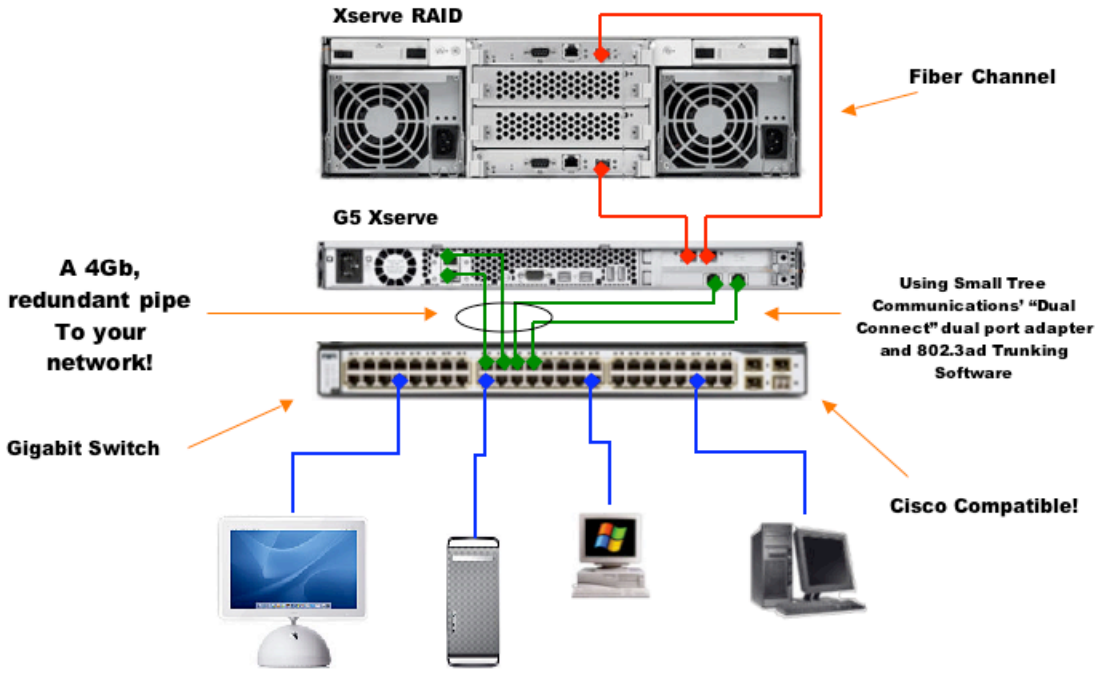
- Multipoint Aggregations
  - The mechanisms specified in this clause do not support aggregations among more than two systems.
- Dissimilar MACs
  - Link Aggregation is supported only on links using the IEEE 802.3 MAC (Gigabit Ethernet and FDDI are not supported in parallel but dissimilar PHYs such as copper and fiber are supported)
- Half duplex operation
  - Link Aggregation is supported only on point-to-point links with MACs operating in full duplex mode.
- Operation across multiple data rates
  - All links in a Link Aggregation Group operate at the same data rate (e.g. 10 Mb/s, 100 Mb/s, or 1000 Mb/s).

Sample Applications

Fileserver:

A small company maintains its database of stock photographs on a G5 Xserve using an Xserve RAID. During certain periods of the quarter when sales are high, and when groups of employees start work, or return from lunch, the server's responsiveness is slow. After some investigation by their local Apple consultant, it is determined that the Gigabit Ethernet port serving the artists is being overrun when all 6 artists attempt to download files to their local systems for processing. The other Gigabit Ethernet port on the G5 server is not being utilized at all. The Xserve RAID is not being used to its maximum bandwidth.

The Apple consultant installs a Small Tree Communications' "Dual Connect" dual port Gigabit Ethernet card and connects the four Gigabit Ethernet ports to the customer's switch. After installing Small Tree Communications' Link Aggregation software, he binds all four ports together into one logical interface. Now, even during peak periods, the traffic to and from the server is evenly balanced between the four ports and the Xserve RAID is fully utilized.



Administrative flexibility:

A small ISP is upgrading his local network to make use of a larger switch. However during this process, he wants to avoid any downtime for his customers. At any time of the day or night there are interactive games and downloads taking place.

The servers he needs to move to the new switch are all connected via 3 aggregated physical links using Small Tree Communication's Link Aggregation software and Gigabit Ethernet ports. The admin simply removes one of the 3 ports from the group and connects it physically to the new switch; He is then able to create a new network connection. Once this is complete and the network is verified to be functional, the admin can modify the DNS entries pointing to his system and customers' connections will begin migrating to the new switch. As things stabilize and traffic moves to the new network, the admin removes the last ports from the old aggregation group and creates a new aggregation group around the new network.

Load Balancing:

A large fileserver is serving home directories to several hundred users on 5 networks. During certain hours, the engineering department will flood their interface because of compiler runs and source code downloads. The result is that the system becomes sluggish for all users as the networking stack attempts to handle retransmits and timeouts on the flooded Gigabit Ethernet port.

The System Administrator corrects this problem by installing Small Tree Communication's Link Aggregation software and trunking all of the 5 networks together as one. He now has a much simpler network topology with less network to network routing and when his programmers go through peak demand periods, their traffic seamlessly spreads over all 5 interfaces. As a result, they get better filesystem utilization and they do not see the server's networking stack bog down with unnecessary timeout and retransmit handling.

Network Backups:

An administrator is tasked with implementing a backup strategy for her network of MAC and Windows clients. She purchases a tape drive for her G5 Xserve and a popular network backup utility. However during the first night's run, she discovers a problem. The backup utility started save processes on several systems connected to the same subnet and that subnet became saturated. This caused the data rate to the tape drive to slow down and reduced the amount of data she was able to pack onto the drive. Additionally, because each subnet became saturated in sequence as the backup continued, her evening backups were not complete by the time her users began to arrive the next day. They all saw very poor performance on their systems and she eventually had to kill the backup so people could work effectively.

To correct this problem, the administrator purchased a copy of Small Tree Communication's Link Aggregation software. This allowed her to combine several of her subnets into a single aggregation group and the backup was able to load balance across all four of her available interfaces. The result was much better backup performance to her tape drive and the backup was able to complete well before employees began arriving in the morning.

High Availability Operation:

A contractor wishes to install a fault tolerant QuickTime™ server for his client. They have several G5 systems serving data off a large fiber channel raid using a popular clustered filesystem and the internal dual port Gigabit Ethernet interfaces. However, when a port is taken out of service for maintenance, users connected to that network experience a "hang" and must restart their video from the beginning. This has caused several problems for customers, especially when the video is actually coming from a live event such as a news conference.

As a solution, the contractor installs 2 Small Tree Communications Dual-Connect Gigabit Ethernet adapters into each machine and trunks them together with the internal ports to create a virtual 6 Gigabit link to the switch. Even in the case where the system cannot drive all six ports at full line rate, the admin now has no single point of failure on that servers network. As a result, he can continue to serve traffic from each server even while replacing cables or switch ports that have been lost due to hardware failure.

Implementation

Our link aggregation implementation currently consists of three binary components:

FILTER: The "Filter" module is a Network Kernel Extension (NKE) that attaches itself to the output path of the virtual interface and the input path of the physical links, in order to redirect those packets. Packets being sent out the virtual interface are redirected to the physical link mapped to that TCP or UDP connection based upon a hashing scheme. All non-UDP or TCP traffic is unconditionally redirected to the first physical link. When redirecting outbound packets, the destination MAC address is changed to the MAC address of the peer of the physical link. The filter also snoops packets received on the physical links, and redirects them to the virtual interface's input queue. When redirecting inbound packets, the destination MAC address is changed to the fake MAC address of the virtual link.

CONTROLLER: The "Controller" module is an IOKit Ethernet driver which instantiates the virtual interfaces. Since there is no underlying hardware, this driver implementation is minimal.

DAEMON: The "Daemon" is a user program called lacpd, which runs in the background as root. This code runs automatically each time the system is restarted. It looks for the configuration file /etc/lacpd.conf and parses it to determine the organization of physical links into virtual interfaces. It controls the two kernel components by sending User Client commands to the Controller. It also communicates with its link aggregation peers on other systems on each of its physical links by sending and receiving Link Aggregation Control Protocol (LACP) frames, which are always sent to a specified multicast address. The LACP frames are sent and received via the Berkeley Packet Filter (BPF) interface for raw I/O.

Conclusion

Small Tree Communication's Link Aggregation software provides three benefits that no network administrator can do without: Increased bandwidth to the network, fault tolerance and better utilization of existing connections, all without the need to purchase additional expensive hardware!

Additionally, Small Tree Communication's Link Aggregation software has a very flexible licensing model that allows you to purchase only what you need. You can begin using the software on your G5 Xserve's existing Gigabit Ethernet connectors for as little as $199. If you later find yourself in need of additional bandwidth to that system, ports can be added to the configuration one at a time, or by purchasing our unlimited upgrade.