

White Paper



*Link Aggregation according to
IEEE Standard 802.3ad*

Link Aggregation according to IEEE 802.3ad

Copyright © SysKonnnect GmbH, 2002. All rights reserved.

All brand and product names are trademarks or registered trademarks of their respective holders. Contents are subject to change.

English

(v1.10 10-Oct-2002)

About SysKonnnect and Marvell:

SysKonnnect focuses on the worldwide development, manufacture, and marketing of high-end adapter boards for computer and telecom networks. SysKonnnect products include both high-performance fiber- and copper-based network interface cards (NICs) for Gigabit-Ethernet systems, as well as FDDI/CDDI and FDDI concentrators for strategic networks. The Company's comprehensive line of SK-NET NICs is ideally suited to secure-server-based public- and private-sector computing environments, such as electronic commerce, finance, healthcare, imaging, and other bandwidth-intensive applications, as well as to enterprise systems developed by SAP and Baan.

SysKonnnect maintains offices in Great Britain and the USA.

Visit www.syskonnnect.com.

SysKonnnect is a subsidiary of Marvell Technology Group Ltd. (NASDAQ: MRVL). Marvell is the leading global semiconductor provider of complete broadband communications solutions for the data communications and storage markets. The Company's diverse product portfolio includes switching, transceiver, communications controller, wireless, and storage solutions that power the entire communications infrastructure, including enterprise, metro, home, and storage networking. As used in this release, the terms "Company" and "Marvell" refer to Marvell Technology Group Ltd. and its subsidiaries, including Marvell Semiconductor, Inc. (MSI), Marvell Asia Pte Ltd (MAPL), Marvell Japan K.K., Marvell Taiwan Ltd., Marvell International Ltd. (MIL), Galileo Technology Ltd., and SysKonnnect GmbH. MSI is headquartered in Sunnyvale, CA and designs, develops, and markets products on behalf of MIL and MAPL. MSI may be contacted at 408-222-2500 or at www.marvell.com.

Table of Contents

1 Why Link Aggregation?	5
Higher Link Availability	5
Increased Link Capacity	5
Aggregating replaces Upgrading	5
2 Types of Link Aggregation	7
Switch-to-Switch Connections	7
Switch-to-Station (Server or Router) Connections	8
Station-to-Station Connections	9
3 The IEEE Standard 802.3ad	11
4 Configuration	13
Physical issues in Link Aggregation	13
Addressing	13
Distribution of Frames	13
Technology Constraints	13
SysKconnect solution for Link Aggregation with Gigabit Ethernet	14
Link Aggregation Control Protocol (LACP)	14
Creation of aggregators and teams	14
Redundant Switch Failover	15
SysKconnect Network Control for Windows 2000	17
5 Conclusion	21

1 *Why Link Aggregation?*

Link Aggregation or trunking is a method of combining physical network links into a single logical link for increased bandwidth. With Link aggregation we are able to increase the capacity and availability of the communications channel between devices (both switches and end stations) using existing Fast Ethernet and Gigabit Ethernet technology. Two or more Gigabit Ethernet connections are combined in order to increase the bandwidth capability and to create resilient and redundant links. A set of multiple parallel physical links between two devices is grouped together to form a single logical link.

Link Aggregation also provides load balancing where the processing and communications activity is distributed across several links in a trunk so that no single link is overwhelmed.

By taking multiple LAN connections and treating them as a unified, aggregated link, we can achieve practical benefits in many applications.

Link Aggregation provides the following important benefits:

- Higher link availability
- Increased link capacity
- Improvements are obtained using existing hardware (no upgrading to higher-capacity link technology is necessary)

Higher Link Availability

Link aggregation prevents the failure of any single component link from leading to a disruption of the communications between the interconnected devices. The loss of a link within an aggregation reduces the available capacity but the connection is maintained and the data flow is not interrupted.

Increased Link Capacity

The performance is improved because the capacity of an aggregated link is higher than each individual link alone.

Standard LAN technology provides data rates of 10 Mb/s, 100 Mb/s, and 1000 Mb/s. Link Aggregation can fill the gaps of these available data rates when an intermediate performance level is more appropriate; a factor of 10 increase may be overkill in some environments.

If a higher capacity than 1000 Mb/s is needed, the user can group several SysKconnect 1000 Mb/s adapters together to form a high speed connection and additionally benefit from the failover function the SysKconnect driver for Link Aggregation supports. This provides migration to 10 Gigabit Ethernet solutions which are not yet available.

Aggregating replaces Upgrading

If the link capacity is to be increased, there are usually two possibilities: either upgrade the native link capacity or use an aggregate of two or more lower-speed links (if provided by the card's manufacturer). Upgrades typically occur in factors of 10. In many cases, however, the device cannot take advantage of this increase. A performance improvement of 1:10 is not achieved, moreover the bottleneck is just moved from the network link to some other element within the device. Thus, the performance will always be limited by the weakest link, the end-to-end connection.

Link aggregation may be less expensive than a native speed upgrade and yet achieve a similar performance level. Both the hardware costs for a higher speed link and the equivalent number of lower speed connections have to be balanced to decide which approach is the most advantageous.

Sometimes link aggregation may even be the only means to improve performance when the highest data rate available on the market is not sufficient.

Many network administrators have experienced that upgrading the network hardware (e.g., switching from 10 Mb/s network adapters to 100 Mb/s network adapters), in the end, led to a performance improvement much less than the 10:1 ratio implied by the hardware change (or perhaps no improvement at all!).

Whether an aggregated link actually yields a performance improvement commensurate with the number of links provided depends to a great extent on network traffic patterns and the algorithm used by the devices to distribute frames among aggregated links. To the extent that traffic can be distributed uniformly across the links, the effective capacity will increase as desired. If the traffic and distribution algorithm is such that a few links carry the bulk of the traffic while others go nearly idle, the improvement will be less than anticipated.

2 Types of Link Aggregation

There are a number of situations where Link Aggregation is commonly deployed:

- Switch-to-switch connections
- Switch-to-station (server or router) connections
- Station-to-station connections

The following figure shows the different uses.

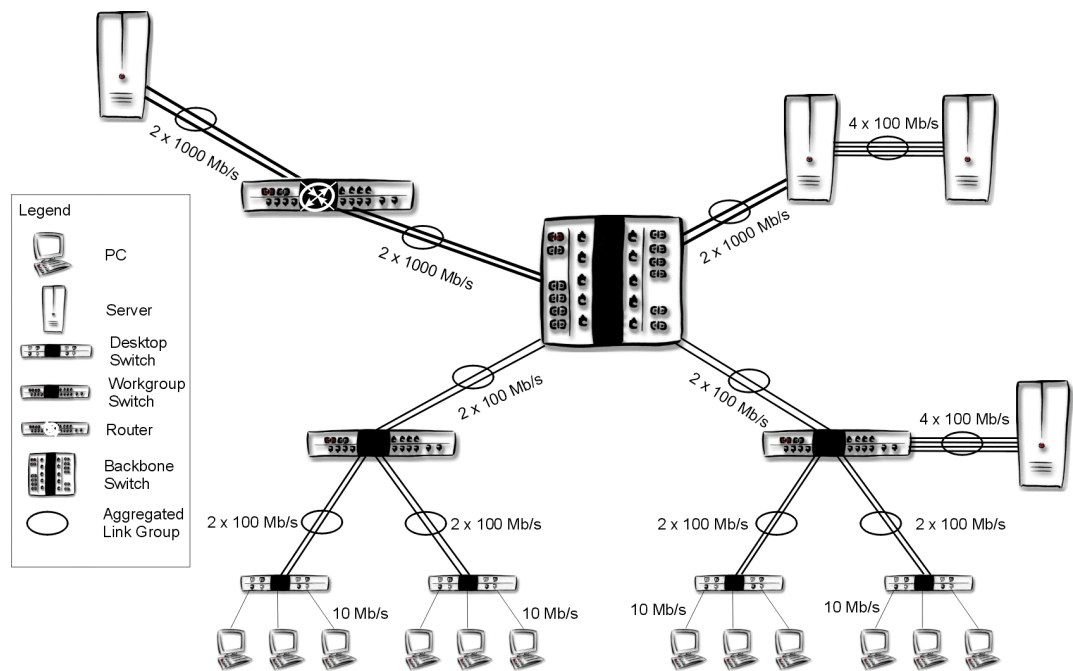


Figure 1. Aggregated links

Switch-to-Switch Connections

In this scenario, multiple workgroups are joined to form one aggregated link. By aggregating multiple links, the higher speed connections can be achieved without hardware upgrade.

Example

In Figure 2, two switches are shown which are connected using four 1000 Mb/s links. If one link fails between these two switches, the other links in the link aggregation group take over the traffic and the connection is maintained.

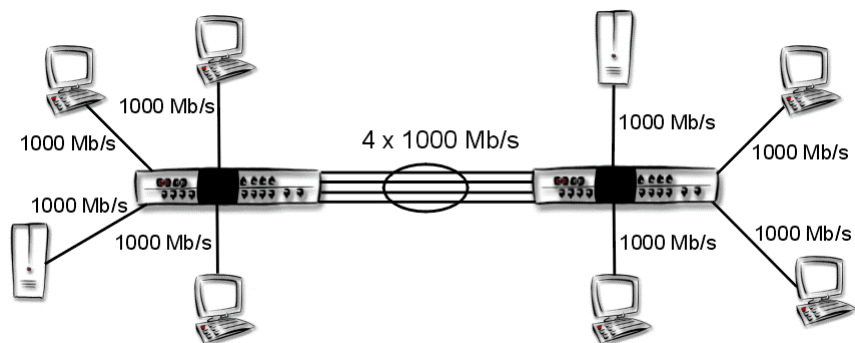


Figure 2. Switch-to-switch connection

This configuration reduces the number of ports available for connection to external devices. Aggregation thus implies a trade-off between port usage and additional capacity for a given device pair.

Switch-to-Station (Server or Router) Connections

Most server platforms can saturate a single 100 Mb/s link with many of the applications available today. Thus, link capacity becomes the limiting factor for overall system performance.

Example

In Figure 3, two servers are shown, each connected to a switch using four 100 Mb/s links. In this application, link aggregation is used to improve performance for the link-constrained station. By aggregating multiple links, better performance is achieved without requiring a hardware upgrade to either the server or the switch. Aggregation on the server side can generally be achieved through software changes in the device driver for the LAN interface(s).

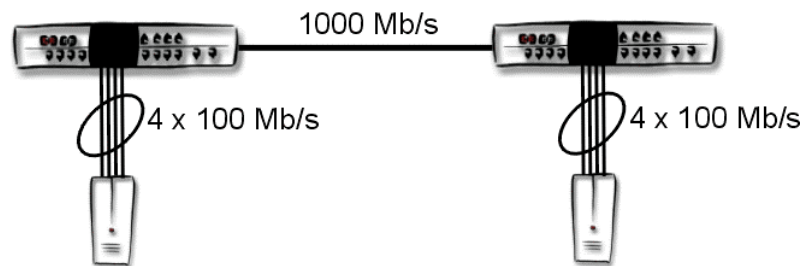


Figure 3. Switch-to-station connection

Link Aggregation trades off port usage for effective link capacity. While it is common for high port-density switches to have some number of excess ports, it is rare for a server to have unused network interface cards. In addition, traditional single-port network adapters use a server backplane slot for each interface; often a server configuration will have only a limited number of slots available for network peripherals. In response to this problem, a number of manufacturers offer multiport network adapters specifically for use in servers, e.g. SysKonnnect's Dual Link Gigabit Ethernet Adapter.

Figure 1 depicts multiple 1000 Mb/s links being aggregated between the backbone switch and a high-performance enterprise backbone router. From the perspective of the switch, a network layer router is simply an end station – not much different from a server. As such, we can aggregate links between a switch and a router for the same reasons as in the switch-to-server case. One important difference arises regarding the choice of algorithm used to distribute frames among the links.

Station-to-Station Connections

In the case of aggregation directly between a pair of end stations, no switches are involved at all. As in the station-to-switch case, the higher performance channel is created without having to upgrade to higher-speed LAN hardware. In some cases, higher-speed NICs may not even be available for a particular server platform, making link aggregation the only practical choice for improved performance.

Example Figure 4 shows two servers interconnected by an aggregation of four 1000 Mb/s links.

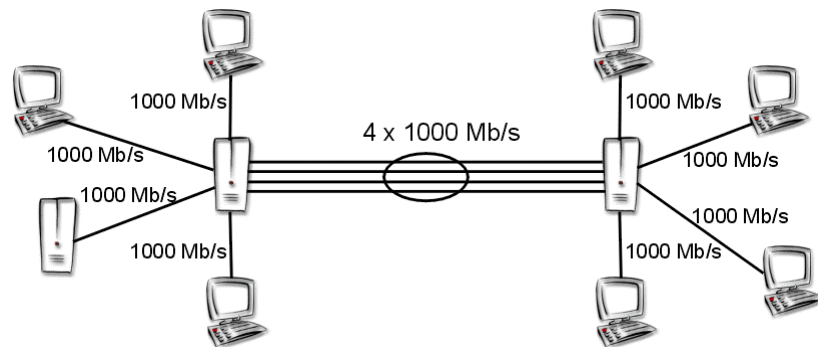


Figure 4. Station-to-station connection

This high-speed connection may be useful for multi-processing or server redundancy applications where high performance is needed to maintain real-time server coherence (this configuration is sometimes called *back-end network*).

This page left blank to accommodate double-sided printing.

3 The IEEE Standard 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 mis-ordering of frames
During both steady-state operation and link (re-) configuration, there is a high probability that frames are neither duplicated nor mis-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).

4 Configuration

Physical issues in Link Aggregation

Addressing

Each network interface controller is assigned a unique MAC address. Usually this address is programmed into the ROM during manufacturing. During initialization, the device driver reads the contents of the ROM and transfers the address to a register within the MAC controller. In most cases, this address is used as source and destination address during the transmission of packets. Aggregated links are to appear as a single link with a single logical network interface and therefore only have one “virtual” MAC address. The MAC address of one of the interfaces belonging to the aggregated link provides the “virtual” address of the logical link.

Frame Distribution

When applying WAN technologies, frames are sometimes broken into smaller units to accelerate transmission (such as in the bonding of B-channel ISDN lines). LAN communications channels, however, do not support sub-frame transfers. The complete frame has to be sent through the same physical link. Using aggregated links, the task is to select the link on which to transmit a given frame. Sending one long frame may take longer than sending several short ones, so the short frames may be received earlier than one long frame. The order has to be restored at the receiver side. Thus, an agreement has been made: all frames belonging to one conversation must be transmitted through the same physical link, which guarantees correct ordering at the receiving end station. For this reason no sequencing information may be added to the frames. Traffic belonging to separate conversations can be sent through various links in a random order. The algorithm for assigning frames to a conversation depends on the application environment and the kind of devices used at each end of the link.

When a conversation is to be transferred to another link because the originally mapped link is out of service (failed or configured out of the aggregation) or a new link has become available relieving the existing ones, precautions have to be taken to avoid mis-ordering of frames at the receiver. This can be realized either by means of a delay time the distributor must determine somehow or through an explicit marker protocol that searches for a marker identifying the last frame of a conversation. The distributor inserts a “marker message” behind the last frame of a conversation. After the collector receives this “marker message” it sends a response to the distributor, which then knows, that all frames of the conversation have been delivered. Now the distributor can send frames of these types of conversations via a new link without delay. If the conversation is to be transferred to a new link, because the originally mapped link failed, this method will not work. There is no path on which the message marker can be transferred, i.e. the distributor has to employ the timeout method.

Technology Constraints

In principle, the devices applied in the aggregation restrict the throughput. Using an aggregation of four 100 Mb/s links instead of one 100 Mb/s link will increase the capacity but the throughput on each link remains the same.

SysKonnnect solution for Link Aggregation with Gigabit Ethernet

All SysKonnnect Gigabit Ethernet adapters support link aggregation according to the IEEE standard 802.3ad. At the moment SysKonnnect provides a Link Aggregation driver for Windows 2000. In the future, Link Aggregation support according to IEEE 802.3ad will be implemented in drivers for other operating systems such as Linux, Sun Solaris, HP-UX and IBM AIX. The SysKonnnect Link Aggregation intermediate driver for Windows 2000, which is part of the “SysKonnnect Network Driver Installation Package”, supports the possibility of combining all ports in a system into port groups. A port group consists of one or more ports, which are connected to the same non-segmented network. One port can not belong to more than one port group. Every port group will behave like a single network interface to the operating system. If a port group consists of only one port, it will behave like a single link adapter. This functional aggregation of ports enables Link Aggregation.

Link Aggregation means that multiple connections between two network instances are treated like a single connection of higher bandwidth. It is implemented according to the standard 802.3ad. It will increase the bandwidth (assuming the system has enough resources to process additional data) and fault tolerance of the connection.

Link Aggregation offers an efficient and low-cost solution to increase bandwidth between server and switch. Another advantage it provides is that if a connection fails completely the remaining links can take over the traffic and thus replace the broken line.

The SysKonnnect Link Aggregation driver helps to increase the network performance by distributing the network traffic among the network adapters belonging to the same group (load balancing). As soon as the server is contacted the driver assigns links to the diverse applications according to the network load. This way, bandwidth can be extended proportionally to the network adapters.

All ports that are configured for Link Aggregation (at least one) can be used to transmit and/or receive frames, depending on their configuration. If one connection fails the aggregated connection will lose bandwidth but remain stable as long as at least one connection of the group is working.

In addition, SysKonnnect offers a Redundant Switch Failover mechanism (RSF). If a switch fails completely, RSF can move the link to a different switch, which then takes over the traffic.

Link Aggregation Control Protocol (LACP)

The LACP is required by the IEEE standard 802.3ad for dynamically exchanging configuration information among cooperating systems in order to automatically configure and maintain link aggregation groups. The protocol is able to automatically detect the presence and capabilities of other aggregation capable devices, i.e. with LACP it is possible to specify which links in a system can be aggregated.

Creation of aggregators and teams

The SysKonnnect Link Aggregation driver uses LACP to establish link aggregation groups between two connected devices. Each device considers itself as being an “actor” and the device at the other end of the link as “partner”, i.e. both devices are actors and both are partners depending on the point of view. The ports send their actor information (LACP frames) to the other device in order to find a suitable partner to form an aggregation. LACP

compares the actor and partner information and then decides which ports can be trunked to establish an aggregator. Aggregators are built automatically from those ports for which partner information is available. Using the SysKonnnect Network Control running on Windows 2000 (see below) the user can combine two or more ports to create a team. For drivers of other operating systems, similar tools for the configuration of Link Aggregation will be available.

Teams and aggregators have the following main features:

- There is a virtual MAC address for the whole team (the MAC address of one adapter of the team).
- The aggregator with the most active links is the active aggregator.
- Every other aggregator is in hot standby.
- The aggregators are configured automatically (see the link aggregation standard).

Below the team level, LACP automatically creates aggregators as described above. If all ports have the same partner information, only one aggregator is established. If there are “n” types of partner information, “n” aggregators will be created. With more than one aggregator within a team, the redundant switch failover mechanism RSF provided by SysKonnnect can be employed. This unique feature is described in the next chapter “Redundant Switch Failover”.

Redundant Switch Failover

Beyond the features required for Link Aggregation in the IEEE 802.3ad standard, SysKonnnect drivers support an additional failover feature, Redundant Switch Failover (RSF). At the moment this feature is only implemented in the driver for Windows 2000, but will be available for other operating systems in the future. The standard requires that all links of a link aggregation group are connected to the same partner. With SysKonnnect’s Link Aggregation driver one team can comprise more than one aggregator. If a team has more than one aggregator, which are connected to separate switches, the failover feature is utilized.

The decision on which link (aggregator) data is transferred is based on the bandwidth (link speed multiplied with number of ports). The link with the largest bandwidth is used for data transmission. If the bandwidth of the aggregators is equal, the first aggregator of a team is used.

In the following figure a typical system employing Link Aggregation and SysKonnnect’s RSF is depicted. Two scenarios are described in the example.

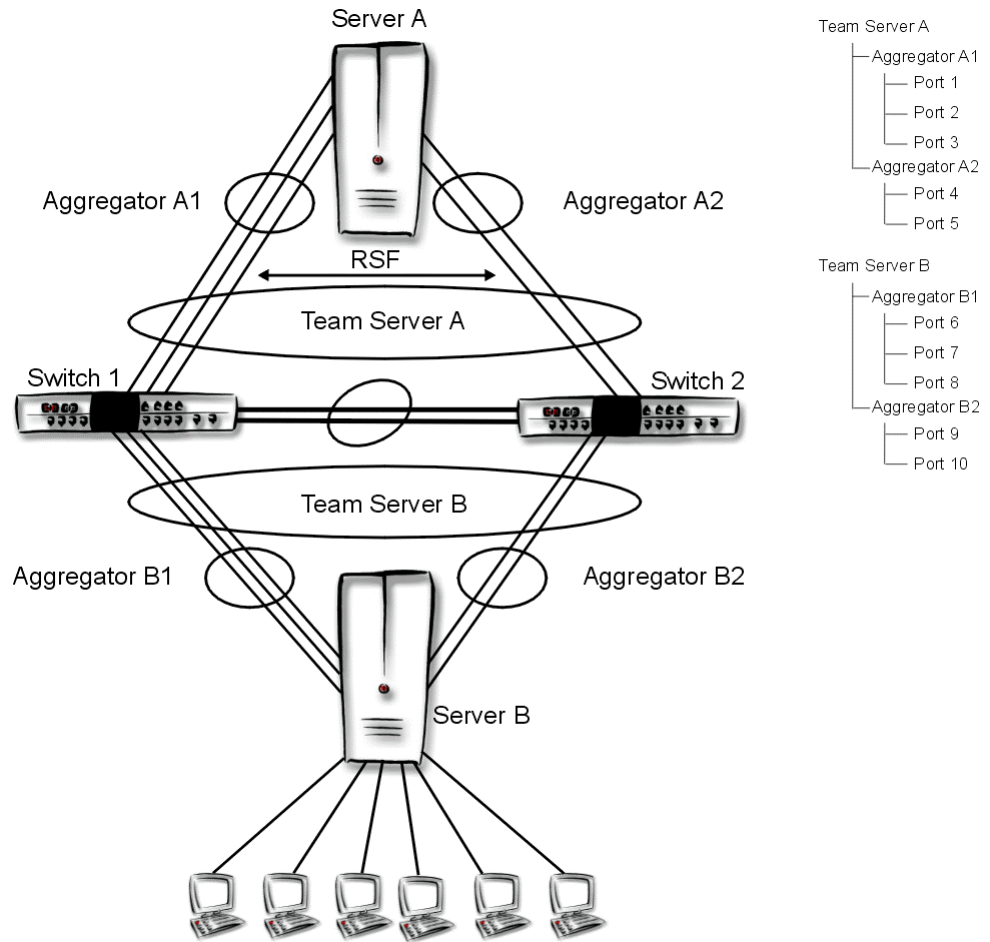


Figure 5. Link Aggregation and RSF

Example In the above system data is normally transferred from Server A to Server B via Aggregator A1, Switch 1, and Aggregator B1 and vice versa, i.e. Aggregators A1 and B1 are the active links, Aggregators A2 and B2 are in hot standby.

Scenario 1: Port failure

Assume two links of Aggregator A1 fail: RSF switches the data flow of Server A to Aggregator A2, because data is always transferred on the link with the larger bandwidth. Data from Server A to Server B is then transmitted via Aggregator A2, Switch 2, Switch 1, and Aggregator B1 (is still the aggregator with the greatest bandwidth in Team of Server B) and vice versa. In case the two links of Aggregator A1 become active again, RSF switches data transfer back to Aggregator A1 due to the larger bandwidth.

Scenario 2: Switch failure

Assume Switch 1 fails: RSF switches the data flow from Server A to Aggregator A2 and the data flow from Server B to Aggregator B2 because both Aggregator A1 and B1 do not have an active link anymore. Data from Server A to Server B is then transmitted via Aggregator A2, Switch 2, and Aggregator B2 and vice versa. When switch 1 is functional again, RSF switches data transmission back to Aggregator A1 and B1 due to the larger bandwidth.

SysKconnect Network Control for Windows 2000

The link aggregation functionality in SysKconnect drivers is configured and controlled via the SysKconnect Network Control, a utility program running on Windows 2000 systems. The SysKconnect Network Control can be reached via “Start” --> “Settings” --> “Control Panel” --> “SysKconnect Network Control”. This utility program enables the user to configure all SysKconnect SK-98xx Gigabit Ethernet adapters installed in a system. The various tabs contain tree views showing the installed adapters and their configuration. The tab “Adapter” displays the network adapters available in the system with their corresponding ports. It shows the ports which have been configured as VLANs, for RLMT (Redundant Link Management Technology) and the ones which have been combined to form a team.

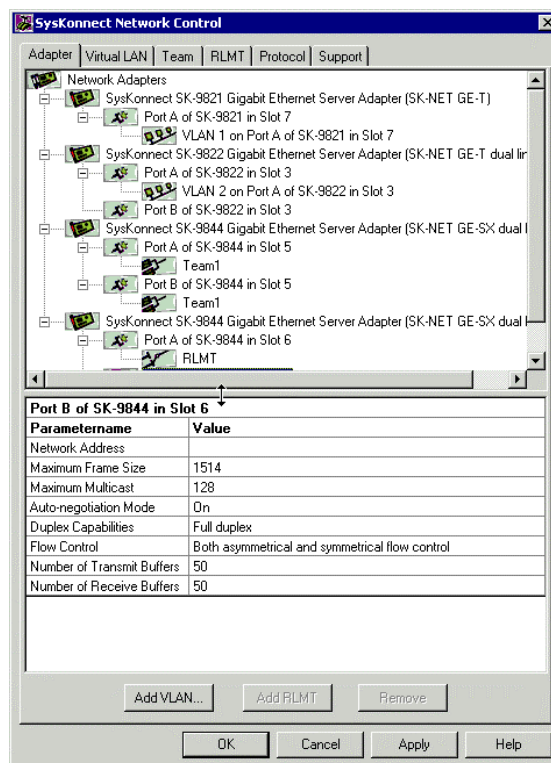


Figure 6. Adapter Overview in SysKconnect Network Control

The corresponding tab in the SysKconnect Network Control where the user is able to configure the link aggregation features is called “Team”. The “Team” tab shows all links or ports of SysKconnect Gigabit Ethernet adapters which are available for teaming or have been already grouped to form a team.

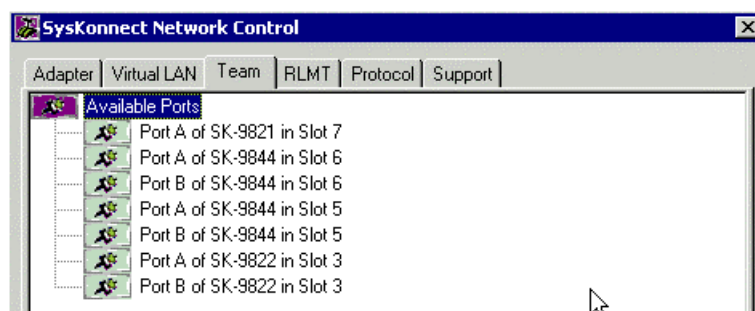


Figure 7. Available ports for teaming

In this tab the user is able to create teams, add ports to teams, remove teams, and rename teams. After a new team has been created the user can add ports to this team. If a port from this team has found a partner on the other end of the connection, which is suitable for link aggregation, a message is displayed next to the corresponding port (see Figure 8).

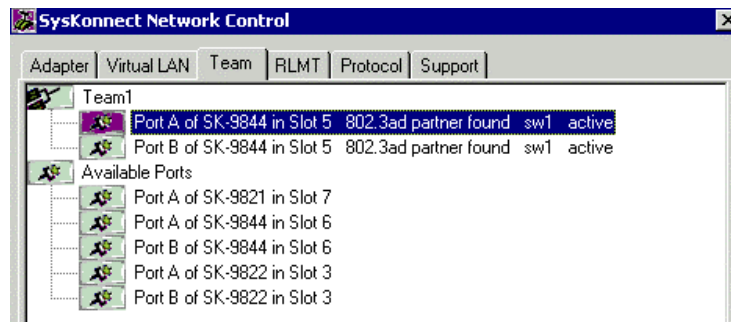


Figure 8. 802.3ad partner found

If a team is selected, the corresponding parameters are displayed below the tree view. The following parameters can be viewed:

Parameter	Values	Description
IP Address	xxx.xxx.xxx.xxx (decimal)	Unique 32-bit (4 bytes) address of an end station within a TCP/IP network. The IP address can not be defined in the SysKconnect Network Control.
Maximum Frame Size	12..9014	This parameter specifies the maximum frame size in bytes the driver will support. The performance of the network usually increases if a large packet size is used. Do not use values larger than 1514 if you are not sure whether or not your network supports jumbo frames. If VLAN is configured the actual frame size on the port is always 4 Bytes larger than the configured frame size of the VLAN because the VLAN tag is inserted into the frame. The Maximum Frame Size can be defined in the SysKconnect Network Control.
Maximum Multicast	0..10000	This option specifies the maximum number of multicast addresses the driver accepts. The Maximum Multicast can be defined in the SysKconnect Network Control.

If a port, which is part of a team, is selected, the following additional link aggregation parameters are displayed among others (see Figure 9):

Parameter	Values	Description
Link Status	Link up Link down	This parameter indicates whether the port has an active link (up) or is inactive (down). The status of this parameter is delivered by the network driver.
Link Aggregation Status	802.3ad partner found no partner found	This parameter indicates whether the port has found an 802.3ad partner with which it can form a team. The status of this parameter is delivered by the network driver.
Switch membership	e.g. sw1, sw2, sw3, ...	This parameter shows to which switch the port is connected physically. The information for this parameter is delivered by the network driver. It can only be changed, if you physically connect the port to a different switch.
Failover Status	active standby	This parameter indicates whether the port is active or in standby mode to take over in case the active link fails.

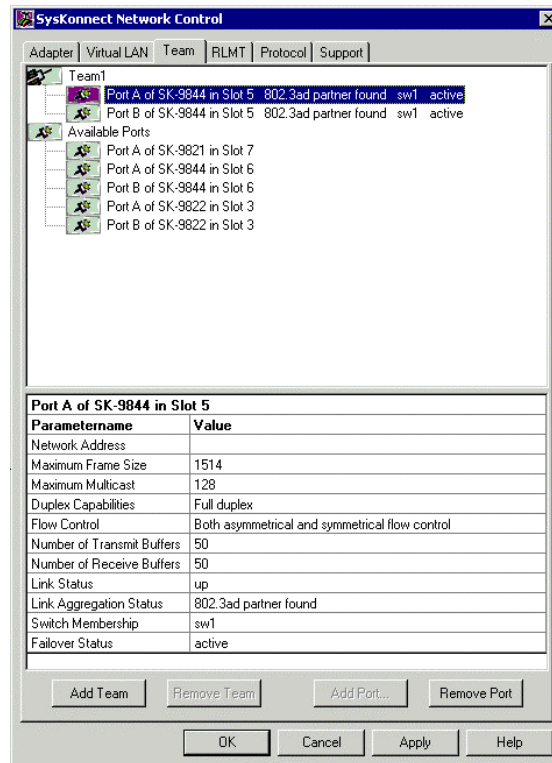


Figure 9. Port parameters of port belonging to a team

This page left blank to accommodate double-sided printing.

5 Conclusion

SysKonnnect' s solution for Link Aggregation offers two main features which are essential for every network administrator: it provides increased capacity and a fail safe system. By employing Link Aggregation the costs for upgrading the performance and the resiliency of a system can be kept reasonable because both benefits can be attained using existing hardware. By using the automatic configuration protocol LACP we can provide redundancy with automatic switching to the standby link in case the active link fails.

The SysKonnnect driver enables load balancing not only on the basis of MAC address information but also on the basis of IP, TCP, and UDP information.

Higher throughput by aggregating multiple links is possible with existing hardware. No additional network adapters have to be purchased. The benefits of Link Aggregation can be reached with the SysKonnnect Network Driver Installation Package for Windows 2000. The package contains the Miniport driver, the Virtual LAN (VLAN) intermediate driver, the Link Aggregation (LAGG) intermediate driver and the configuration utility SysKonnnect Network Control and is available for free download for SysKonnnect customers on our web site: www.syskonnnect.com.

Demanding applications running in high-performance environments like servers in enterprises, web servers, and intranet servers gain particularly from the high-bandwidth and duplex capabilities of Link Aggregation.

The SysKonnnect implementation of Link Aggregation also provides a perfect solution on the road to the migration to 10 Gigabit Ethernet which will be integrated in the future. The user can fill the capacity gap by employing for example four 1000 Mb/s adapters in a team and also gets the benefit of a failsafe system by making use of the Redundant Switch Failover mechanism the SysKonnnect driver provides.



Headquarters

SysKonnect GmbH
A Marvell@Company
Siemensstrasse 23
D-76275 Ettlingen

Germany

Phone: + 49 7243 502 100
Support: + 49 7243 502 330
Fax: + 49 7243 502 989

E-mail: sales@syskonnect.de

Americas, Canada,
and Pacific Rim

SysKonnect, Inc.
A Marvell@Company
1922 Zanker Road
San Jose, CA 95112

USA

Sales: + 1 800 716 3334
+ 1 408 222 0809
Support: + 1 866 782 2507
+ 1 408 222 0666
Fax: + 1 408 752 9040

E-mail: info@syskonnect.com

Europe, Middle East,
and Africa

SysKonnect Ltd.
A Marvell@Company
55 Henley Drive
Frimley Green, Camberley
Surrey, Gu16 6NF

United Kingdom

Phone: + 44 1 252 836 467
Fax: + 44 1 252 836 537

E-mail: sales-uk@syskonnect.de