MULTI-SWITCH CHASSIS

Inventors: Bjorn Dag Johnsen, Oslo (NO);
           Ola Torudbakken, Oslo (NO);
           Andreas Bechtleheim, Menlo
           Park, CA (US)

Assignee: Sun Microsystems, Inc., Santa
           Clara, CA (US)

Abstract

In a switch system \( L \) groups of the line switch elements are
connectable to cables that include \( L \) links such that each of the
\( L \) links within a cable connect to a switch element of a respec-
tive one of the \( L \) groups. Fabric switch elements are con-
ected such that a fabric switch element is connected to the
line switch elements of one of the group of line switch ele-
ments.
FIG. 4B
 FIG. 6
FIG. 12
MULTI-SWITCH CHASSIS
RELATED APPLICATIONS


BACKGROUND

[0002] The invention relates to communication systems, for example to a switch system.

[0003] An example of a switch system is a Clos Network. A Clos network, first described by Charles Clos in 1954, is a multi-stage fabric built from smaller individual switch elements that provides full-bisected bandwidth for all end points, assuming effective dispersive routing.

[0004] When constructing large Clos based fabric topologies, it is desirable to minimize the number of cables as well as the number of individual switch chassis instances that has to be involved. For those reasons, it is desirable to have a single large switch chassis with high enough radix (number of ports) to enable connectivity to all relevant nodes in the set of possible target network configurations.

[0005] However, due to various physical constraints (e.g., maximum board size, number of individual switching elements, stages and associated connectivity, power and cooling capacity, connector sizes and cable management) it may not be possible to construct a single switch with sufficient radix. Also, the entry level cost of the largest possible switch may make it less attractive or unattractive for configurations that do not utilize the full radix.

[0006] The present invention seeks to at least mitigate these concerns.

SUMMARY

[0007] An aspect of the invention provides a switch system that includes line switch elements and fabric switch elements, wherein L groups of line switch elements are connectable to cables that include L links such that each of the L links within a cable connect to a switch element of a respective one of the L groups. A fabric switch element is arranged to connect to line switch elements of one of the group of line switch elements.

[0008] An example embodiment comprises a switch system that includes one or more line cards connectable to cables that comprises up to L links. L groups of line switch elements are configured such that each link within the cables connect to a line switch element of a respective one of the L groups of line switch elements. One or more fabric cards are connected to the line card(s). Fabric switch elements are provided on the fabric card(s). Connectivity to the fabric switch elements is such that a fabric switch element connects only to line switch elements that belong to a group of line switch elements.

[0009] The configuration of an example embodiment of the invention is such that the radix of the complete switch can be observed as the sum of the radices of the individual fabric switches, whereby the imposed management model can be that of a single core switch rather than multiple core switches.

[0010] Although various aspects of the invention are set out in the accompanying independent and dependent claims, other aspects of the invention include any combination of features from the described embodiments and/or the accompanying dependent claims, possibly with the features of the independent claims, and not solely the combinations explicitly set out in the accompanying claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Specific embodiments are described by way of example only with reference to the accompanying Figures in which:

[0012] FIG. 1 is a schematic representation of a switch chassis;

[0013] FIG. 2 is a schematic representation of the front of an example switch chassis;

[0014] FIG. 3 is a schematic representation of a midplane illustrating the logical connectivity through the midplane between cards at the rear and cards at the front orientationally with respect to each other;

[0015] FIG. 4A is a schematic diagram of an example management infrastructure;

[0016] FIG. 4B continues the schematic diagram of FIG. 4A;

[0017] FIGS. 5 to 11 are views of an example of a switch chassis;

[0018] FIG. 12 is a first isometric view of an example of a midplane;

[0019] FIG. 13 is a further isometric view of an example of a midplane;

[0020] FIG. 14 is an isometric view of an example of a fabric card;

[0021] FIG. 15 is an isometric view of an example of a fabric card;

[0022] FIG. 16 is schematic representations of part of a switch chassis;

[0023] FIG. 17 is a further schematic representation of part of a switch chassis;

[0024] FIG. 18 is a schematic representation of the connections of two cards orthogonally with respect to each other;

[0025] FIG. 19 is a schematic representation of an example of orthogonally arranged connectors;

[0026] FIG. 20 is a schematic side view of one of the connectors of FIG. 19;

[0027] FIG. 21 is a plan view of an example configuration of vias for the orthogonal connector pairing of FIG. 19;

[0028] FIG. 22 is a cross-section through of a via;

[0029] FIG. 23 is a schematic side view of example of an alternative to the connector of FIG. 20;

[0030] FIG. 24 is a schematic end view of an example cable connector;

[0031] FIG. 25 is a schematic side view of the example cable connector;

[0032] FIG. 26 represents a footprint of the cable connector;

[0033] FIGS. 27 and 28 illustrates example of signal routing for a cable connector;

[0034] FIG. 29 illustrates an example of a power supply for the cable connector;

[0035] FIG. 30 illustrates an example of cable status sense detection circuitry;

[0036] FIG. 31 illustrates an example of hot plug control circuitry;

[0037] FIG. 32 is a schematic representation of airflow though a switch chassis; and

[0038] FIG. 33 is a schematic representation of an example of a switch infrastructure.
While the invention is susceptible to various modifications and alternative forms, specific embodiments are shown by way of example in the drawings and are herein described in detail. It should be understood, however, that drawings and detailed description thereto are not intended to limit the invention to the particular form disclosed, but on the contrary, the invention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the present invention.

DETAILED DESCRIPTION

An example embodiment of a 3456-port InfiniBand 4x DDR switch in a custom rack chassis is described, with the switch architecture being based upon a 5-stage CLOS fabric. The rack chassis can form a switch enclosure.

The CLOS network, first described by Charles Clos in 1954, is a multi-stage fabric built from smaller individual switch elements that provides full-bisectional bandwidth for all end points, assuming effective dispersive routing.

Given that an external connection (copper or fiber) costs several times more per port than the silicon cost, the key to make large CLOS networks practical is to minimize the number of external cables required and to maximize the number of internal interconnections. This reduces the cost and increases the reliability. For example, a 5-stage fabric constructed with switching elements of size (n) supports (n^2/2^n) edge points, using (5*n^2/2^n) switch elements with a total of (3*n^2/2^n) connections. The ratio of total to external connections is 5:1, i.e. 80% of all connections can be kept internal. The switch elements (switch chips) in the described example can be implemented using a device with 24 4x DDR ports.

An example switch uses a connector that support 3 4x ports per connector, which can further to minimize a number of cables required. This can provide a further 3:1 reduction in the number of cables. In a described example, only 1152 cables (1/3*n^2/2^n) are required.

In contrast if prior commercially available 288-port switches and 24-port switches were used to create a 3456-port fabric a total of 6912 cables (2*n^2/2^n) would be required.

The example switch can provide a single chassis that can implement a 5-stage CLOS fabric with 3456 4x DDR ports. High density external interfaces can be provided, including fiber, shielded copper, fiber and twisted pair copper. The amount of cabling can be reduced by 84.4% when compared to building a 3456-port fabric with commercially available 24-port and 288-port switches. In the present example, an orthogonal midplane design can be provided that is capable of DDR data rates.

An example switch can address a full range of HPC cluster computing from a few hundred to many thousand of nodes with a reliable and cost-effective solution that uses fewer chassis and cables than prior solutions.

FIGS. 1 and 2 are schematic diagrams of an example of a switch chassis as viewed from the rear (FIG. 1) and front (FIG. 2), respectively. This example comprises a custom rack chassis that is 60" high, 47" wide, and 36" deep, not including a cable management system. The present example provides a passive orthogonal midplane design (not shown in FIGS. 1 and 2) that provides a direct interface between Line Cards (LC) and Fabric Cards (FC). The line cards provide connections to external lines and the fabric card form switch fabric cards for providing switching functions.

In the present example, up to 18 fabric cards (FC0 to FC17) are provided. Each fabric card plugs vertically into the midplane from the rear.

In the present example, up to 24 line cards (LC0 to LC23) can be provided. Each line card provides 144 4x ports (24 stacked 168-circuit cable connectors). Each line card plugs horizontally into the midplane from the front.

Up to 16 hot-pluggable power supply units (PS0-PS16) are provided. Each midplane card plugs into a power distribution board (PDB), which is not shown in FIGS. 1 and 2. Two bushs (not shown in FIGS. 1 and 2), one per group of 8 power supply units, distribute direct current (DC) supply to the line cards and the fabric cards.

Two hot-pluggable Chassis Management Controllers (CMCs) plug into the power distribution board from the front. Each chassis management controller comprises a mezzanine card.

The power distribution board is a passive power distribution board that supports up to 16 power supply units DC connectors and 2 chassis management controller slot connectors. The power distribution board connects to the midplane through ribbon cables that carry low-speed signals.

In the present example, up to 144 fan modules (Fan0-Fan143) are provided, with 8 fan modules per fabric card in the present instance. Cooling air flow in controlled to be from the front to the rear, using redundant fans on each fabric card to pull the air from the line cards through openings (not shown in FIGS. 1 and 2), in the midplane. The power supply units have their own fans for cooling with the air exiting through the rear of the chassis. The power supply units are also used to cool the chassis management controllers.

FIG. 3 is a schematic representation of a printed circuit board 30, which is configured as a midplane 30 in the switch chassis 10. The midplane 30 is configured in an orthogonal manner such that each fabric card 12 can connect to each of the line cards 14 without requiring any signal traces on the midplane 30. The orthogonal midplane design can provide excellent signal integrity in excess of 10 Gbps per differential pair.

The midplane 30 is represented schematically to show an array of midplane connector pairs 32 as black squares with ventilation openings shown as white rectangles. Each midplane connector pair comprises a pair of connectors (to be explained in more detail later) with one connector on a first face the midplane and a second connector on the other face of the midplane, the first and second connectors being electrically interconnected by way of pass-through vias (not shown in FIG. 3) formed in the midplane 30. As will be explained later, the first and second connectors of a midplane connector pair 32 are each multipath connectors. They are arranged orthogonally with respect to one another such that a first midplane connector of a midplane connector pair 32 is connectable to a fabric card 12 on a first side of the plane 30 in a first orientation and a second midplane connector of the midplane connector pair 32 is connectable to a line card on a second side of the plane 30 in a second orientation substantially orthogonally to the first orientation.

In an example described herein, each of the first connectors of the respective midplane connector pairs 32 of a column 31 of midplane connector pairs 32 can be connected
to one fabric card 12. This can be repeated column by column for successive fabric cards 12. In an example described herein, each of the second connectors of the respective midplane connector pairs 32 of a row 33 of midplane connector pairs 32 can be connected to one line card 14. This can be repeated row by row for successive line cards 14. As a result, the midplane can be populated by vertically oriented fabric cards 12 on the first side of the midplane and horizontally oriented line cards 12 on the second side of the midplane 30.

In the present example the midplane 30 provides orthogonal connectivity between fabric cards 12 and the line cards 14 using orthogonal connector pairs. Each orthogonal connector pair provides 64 differential signal pairs, which is sufficient to carry the high-speed signals needed as well as a number of low-speed signals. The orthogonal connector pairs are not shown in FIG. 3, but are described later.

The midplane 30 is also configured to provide 3.3 VDC standby power distribution to all cards and to provide 12C/ System Management Bus connections for all fabric cards 12 and line cards 14.

Another function of the midplane 30 is to provide thermal openings for a front-to-rear airflow. The white holes in FIG. 3 (e.g. hole 34) form openings in the midplane for airflow. In this example the midplane is approximately 50% open for airflow.

The fabric cards 12 each support 24 connectors and the line cards 14 each support 18 connectors.

FIG. 3 also illustrates an example of how the fabric cards 12, the midplane 20 and the line cards 14 interconnect. In this example there are 24 switch chips on a line card 14 and 8 chips on each of the 18 fabric cards 12.

As previously mentioned a 5-stage Clos fabric has a size n^n/2m/2 in which n is the size of the switch element. The example switch element in FIG. 3 has n equal to 24 ports. Each line card 14 has 24 chips in 2 rows with 12 chips in each row. Each of 12 ports of each switch chip 35 in a first row 36 of the line card 14 is connected to 2 cable connectors 42, with 6 ports per connector. There are a total of 24 cable connectors per line card 14. Each cable connector can accommodate two physical independent cables that each carries 3 ports (links). Each cable connector 42 can accommodate 6 ports. The remaining 12 ports of each switch chip 35 in the first row 26 is connected to one chip 35 each in a second row 38 of chips 35.

There are 18 midplane connectors 32 per line card 14. Each midplane connector 32 provides one physical connection to one fabric card 14. Each midplane connector 32 can accommodate 8 x 4 links (there are 8 differential pairs per 4 x link and a total of 64 differential pairs provided by the orthogonal connector)

12 ports of each of the switch chips 35 in the second row 38 of the line card 14 are connected to 2 line card connectors 40 that are used to connect the line card 14 to the midplane connectors 32 and thereby with the fabric cards 12 through the orthogonally oriented midplane connector pair. Of the 12 ports per switch chip 35, eight ports are connected to one line card connector 40, and the remaining four ports are connected to another line card connector 40 as represented by the numbers 8 and 4 adjacent the two left hand switch chips 35 in the second row 38. 2 switch chips are thereby connected to a group of 5 line card connectors 40 and hence to a group of three midplane connectors pairs 32.

The remaining 12 ports of each switch chip 35 in the second row 38 of the line card 14 are connected to each of the 12 switch chips 35 in the first row 36 of the line card 14.

At the fabric card 12 all links through an orthogonally oriented midplane connector pair 32 are connected to one line card 14. A single orthogonal connector 46 carries 8 links. These links are connected to one switch element 44 each at the fabric card 12.

Also shown in FIG. 3 are power connectors 37 on the midplane and power connectors 39 on the fabric cards 12.

There has been described a system with 24 line cards with 144 ports each, realized through 48 physical cable connectors that each carry 3 links. The switch fabric structure of each line card 14 is fully connected, so the line card 14 itself can be viewed upon as a fully non-blocking 144 port switch. In addition each line card 14 has 144 links that are connected to 18 fabric cards. The 18 fabric cards then connect all the line cards 14 together in a 5-stage non-blocking Clos topology.

FIGS. 4A and 4B are schematic diagrams of an example management infrastructure. This example provides redundant chassis management controllers 18. In addition each fabric card 12 and line card 14 supports an management controller. There are redundant management connections from each chassis management controller 18 to each of the fabric card and line card management controllers. In addition there are 12C connections to each of the power supply units 16. The management connections pass between the fabric cards 12, the line cards 14, the power supply units 16 and the chassis management cards 18 via the midplane and the power distribution board 22 in the present example.

FIGS. 5 to 11 provide various schematic views of an example of a switch chassis in accordance with the invention.

FIG. 5 is a front view of the switch chassis 10 showing cable management structures 50. FIG. 6 is a rear view of the switch chassis 10 showing the fabric cards 12, the power supply units 16 and cable management structures 50. FIG. 6 is a side view of the switch chassis 10 further showing the cable management structures 50. FIG. 8 is a side view of the switch chassis 10 further showing the cable management structures 50. FIG. 10 is an isometric view of the switch chassis 10 from the line card 14 (front) side showing the cable management structures 50. FIG. 10 is an isometric view of the switch chassis 10 from the line card 14 (front) side showing the cable management structures 50.

FIGS. 12 and 13 provide various schematic views of an example of a midplane 30 in accordance with the invention. FIG. 12 is an isometric view of the midplane 30 from the line card 14 (front) side and FIG. 13 is an isometric view of the midplane 30 from the line card 12 (rear) side. FIG. 12 shows the array formed from rows and columns of the second connectors 64 of the midplane connectors pairs 32 described with reference to FIG. 3. FIG. 13 shows the array formed from rows and columns of the first connectors 62 of the midplane connectors pairs 32 described with reference to FIG. 3.

FIG. 14 is an isometric view of an example of a line card 14. This shows the first and second rows 36 and 38 of switch chips 35, the line board connectors 40 and the cable connectors 42. As can be seen in FIG. 14, the cable connectors
42 are stacked double connectors such each cable connector can connect to two cables 52 and 54.

[0074] FIG. 15 is an isometric view of an example of a fabric card 12. This shows the fabric card connectors 46 and the switch elements 44.

[0075] FIG. 16 is a schematic representation of an example of two chassis management controllers 18 plugged into one side of a power distribution board 22 and 16 power supply units 16 plugged into the other side of the power distribution board 22. In the present example, the chassis management controllers 18 are plugged into the front side of the power distribution board 22 and the power supply units 16 are plugged into the rear side of the power distribution board 22 as mounted in the switch chassis. FIG. 17 illustrates bus bars 24 for a 3.3V standby supply.

[0076] In the present example the midplane 30 is a passive printed circuit board that has dimensions of 1066.8 mm (42")x908.05 mm (35.75")x7.1 mm (0.280"). The active area is 40"x34". 864 8x8 midplane connectors (432 midplane connectors per side) are provided. There is a ribbon cable connection the power distribution board 22 and a 3.3V standby copper bar to the power distribution board 22.

[0077] In the present example a fabric card 12 comprises a printed circuit board with dimensions of 254 mm (10")x1016 mm (40")x4.5 mm (177"). It comprises 24 8x8 fabric card connectors 46, one power connector 39, 8 fan module connectors and 8 switch chips 44.

[0078] In the present example a line card 14 comprises a printed circuit board with dimensions of 317.5 mm (12.5")x965.2 mm (38")x4.5 mm (177"). It comprises 24 stacked cable 168-circuit connectors 42, 18 8x8 card connectors 40, 1 busbar connector and 24 switch chips 35.

[0079] In the present example a power distribution board 22 comprises a printed circuit board, 16 power supply DC connectors, 14 6x6 card connectors (7 connectors per chassis management card 18, ribbon cable connectors for low-speed connectivity to the midplane 30, and a 3.3V standby copper bar to the midplane 30.

[0080] In the present example a chassis management card 18 comprises 14 6x6 card connectors (7 connectors per chassis management card), two RJ45 connectors for Ethernet available on a chassis management card panel, two RJ45 connectors for serial available at the chassis management card panel, three RJ45s for line card/fabric card debug console access at the chassis management card panel, three 16X rotary switches used to switch between which line card/fabric card debug console is connected to the other RJ45s above, and a 220-pin connector for the mezzanine.

[0081] In the present example a mezzanine has dimensions: 92.0 mmx50.8 mm and comprises 4 mounting holes screw with either 5 mm or 8 mm standoff from the chassis management card board, a 220-pin connector for connectivity to chassis management board.

[0082] FIG. 18 is a schematic isometric view of an example of a midplane connector pair 32. As can be seen in FIG. 18, the connector comprises a first, fabric side, connector 62 and a second, line card side, connector 64. In this example, each of the connector 62 and 64 is substantially U-shaped and comprises an 8x8 array of contact pins.

[0083] It will be noted that the second connector 64 of the midplane connector pair 32 is rotated through substantially 90 degrees with respect to the first connector 62. The first connector 62 is configured to connect to a corresponding fabric card connector 46 of a fabric card 12. The second connector 62 is configured to connect to a corresponding fabric card connector 46 of a line card 14. Through the orientation of the second connector 64 of the midplane connector pair 32 substantially orthogonally to the orientation of the first connector 62, it can be seen that the line card 14 is mounted substantially orthogonally to the fabric card 12. In the present example the line card 14 is mounted substantially horizontally and the fabric card is mounted substantially vertically 12.

[0084] Each of the contact pins on the connector 62 is electrically connectable to a corresponding contact of the fabric card connector 46. Each of the contact pins on the connector 64 is electrically connectable to a corresponding contact of the line card connector 40. The connector pins of the respective connectors 62 and 64 are connected by means of pass-through vias in the midplane 30 as will now be described in more detail.

[0085] FIG. 19 illustrates an example of the configuration of a first midplane connector 62 and a second midplane connector 64 of a midplane connector pair 32 in more detail. In the example shown in FIG. 19 that second connector 64 (the line card side connector) comprises a substantially U-shaped frame 70 including a substantially planar base 71 and first and second substantially planar walls 72 and 74 that extend at substantially at 90 degrees from the base 71. The inside edges of the first and second substantially planar sides 72 and 74 are provided with ridges 76 and grooves 78 that provide guides for the line card connector 40.

[0086] As can be seen in FIG. 18, the line card connector 40 has a structure that comprises a plurality of contact planes 63 that are aligned side by side, such that it has a generally planar construction that extends up from the line card 14. Line card connector planes comprise printed circuit boards carrying traces leading to contacts. The traces and contacts can be provided on both sides of the printed circuit boards of the line card connector planes.

[0087] By comparing FIGS. 18 and 19, it can be seen that each contact plane 63 of the line card connector 40 can be entered into a respective one of the grooves 78 so that connectors of the line card connector 40 can then engage with contact pins 80 of the second connector 64. In the case of the line card side connector portion 64, the orientation of second connector 64 and the grooves 78 therein means that the line card 12 is supported in a substantially horizontal orientation. In the example shown in FIG. 19, an 8x8 array of connector pins 80 is provided.

[0088] The first midplane connector 62 (fabric card side connector) of the midplane connector pair 32 has substantially the same form as the second midplane connector 62 of the midplane connector pair 32, except that it is oriented at substantially 90 degrees to the second midplane connector 64. In this example the second midplane connector 62 comprises a substantially U-shaped support frame 75 including a substantially planar base and first and second substantially walls and that extend at substantially at 90 degrees from the base. The inside edges of the first and second substantially planar sides are provided with ridges and grooves that provide guides for the fabric card connector 46. The fabric card connector 46 has the same basic structure as that of the line card connector 40 in the present instance. Thus, in the same way as for the line card connector, each of a plurality of contact planes of the fabric card connector 46 can be entered into a respective one of the grooves so that connectors of the fabric card connector 46 can then engage with contact pins of the
first connector 62. The orientation of the first connector 62 and the grooves therein means that the fabric card 12 is supported in a substantially vertical orientation.

[0089] In the example illustrated in FIG. 19, the orthogonal connector 60 provides an 8x8 array of connector pins 80 provided that can support supports 64 differential pairs or 32 bi-directional serial channels (two wires per direction) in a footprint of 32.2x32.2 mm.

[0090] As mentioned above, the contact pins of the first and second midplane connectors 62 and 64 of a midplane connector pair 32 are connected by means of pass through vias in the midplane.

[0091] FIG. 20 illustrates a side view of an example of a midplane connector, for example the midplane connector 62 mounted on the midplane. In the example shown in FIG. 20 the midplane connector 64 comprises a substantially U-shaped frame 70 including a substantially planar base 71 and first and second substantially planar walls 72 and 74 that extend at substantially 90 degrees from the base 71. The contact pins 80 are each connected to pairs of contact tails 81 that are arranged in sprung pairs that are arranged to be push fitted into pass through vias 83 in the midplane 30.

[0092] In use, the other midplane connector (e.g., the first midplane 62) of the midplane connector pair would be inserted into the pass through vias in the other side of the midplane 30 in the orthogonal orientation as discussed previously.

[0093] FIG. 21 is a schematic representation of an area of the midplane for receiving the midplane connectors 62 and 64 of the midplane connector pair 32. This shows the array of vias 83. FIG. 22 is a schematic cross-section though such a via 83 in the showing the conductive wall 85 of the via 83. The conductive wall 85 can be formed by metal plating the wall of the via, for example.

[0094] The examples of the midplane connectors described with reference to FIGS. 18 and 20 had a generally U-shape. However, other configurations for the midplane connectors are possible. For example FIG. 23 illustrates another example of a midplane connector pair 32, where the first and second midplane connectors 62 and 64 are generally the same as the first and second midplane connectors 62 and 64 described with reference to FIG. 19 except that, in addition to the first and second walls 72 and 74, third and fourth walls 73 and 75 are provided. The additional walls provide a generally box-shaped configuration that can facilitate the insertion and support for the cards to be connected thereto.

[0095] It will be appreciated that in other examples the first and second midplane connectors could have different shapes and/or configurations appropriate for the connections for the cards to be connected thereto.

[0096] The array of midplane connector pairs 32 as described above provides outstanding performance in excess of 10 Gbps over a conventional FR4 midplane because the orthogonal connector arrangements allow signals to pass directly from the line card to the fabric card without requiring any signal traces on the midplane itself. The orthogonal arrangements of the cards that can result from the use of the array of orthogonally arranged connector pairs also avoids the problem of needing to route a large number of signals on the midplane to interconnect line and fabric cards, minimizing the number of layers required. This provides a major simplification compared to existing fabric switches. Thus, by providing an array of such orthogonal connectors, each of a set of horizontally arranged line cards 12 can be connected to each of a set of vertically aligned fabric cards without needing intermediate wiring.

[0097] FIGS. 24 and 25 provide an end view and a side view, respectively, of an example of a cable connector 42 as mentioned with reference to FIGS. 3 and 14. As shown in FIGS. 24 and 25, the cable connectors 24 and 25 include first and second cable connections 92 and 94 stacked within a single housing 90. This provides for a very compact design. Board contacts 96 are provided for connecting the connector to a line card 14. FIG. 26 is a plan view of the connector footprint for the board contacts 96 of the cable connector 42. The stacked arrangement facilitates the providing of line cards that are high density line cards supporting a 12X cable providing 24 line pairs with 3X4 links aggregated into a single cable. The cable connectors provide 12X cable connectors that are smaller than a conventional 4X connector, 3X denser than a standard InfiniBand 4X connector and electrically and mechanically superior. Using 12X cable (24 pairs) can be almost 50% more area efficient than three 4X cables and requires three times fewer cables to install and manage.

[0098] FIGS. 27 and 28 illustrate an example of the routing of signals from each of two 12X port sections 92 and 94 of a cable connector 42 to the equalizers and to a switch chip on a line card 14. FIG. 27 shows an example of routing from a first 12X port section. FIG. 28 shows an example of the routing from a second 12X port section. The transmit (Tx) lines are equalized, and can be connected directly from the switch chip to the cable connector. The can be routed on lower layers in order to minimize via stub effects.

[0099] FIG. 29 illustrates an example of a power supply for the cable connector and FIG. 30 illustrates an example of a cable status sense detection circuitry. The cable sense detection circuitry is operable to test from each end whether the other end is plugged or not, and, if plugged, to see if power from the power supply is on. Provisions are made such that “leaking” power from a powered to un-powered end is avoided. A valid status assumes that an active end is plugged. FIG. 31 is a schematic diagram of an example of a hot plug control circuit that enables hot plugging of cables. The switch chassis can thereby provide active cable support for providing active signal restoration at a cable connector. Active cable support can provides benefits of increased distances for copper cables as a result of active signal restoration at the connector, increased maximum cable distance by over 50%, using thinner and more flexible cables (e.g., reducing a cable diameter by up to 30%, which facilitates good cable management. A cable to connector interface can provide one, more or all of local and remote cable insertion detection, cable length indication, remote node power-on detection, remote power, a serial number and a management interface.

[0100] FIG. 32 is a schematic representation of the airflow through an example switch chassis. As illustrated by the arrows, the airflow is from the front to the rear, being drawn through by fans 20 in the fabric cards 12 and the power supplies 18.

[0101] The air inlet is via perforations at the line card 14 front panel. Fans 20 at the fabric cards 12 pull air across the line cards, though the openings 34 in the vertical midplane 30 and across the fabric cards 12.

[0102] Line card cooling is naturally redundant since the fabric cards are orientate orthogonally to the line cards. In other words, cooling air over each line card is as a result of the contribution of the effect of the fans of the fabric cards along
the line card due to the respective orthogonal alignment. In the case that a fabric card fails or is removed, a portion of the cooling capacity is lost. However, as the cooling is naturally redundant the line cards will continue to operate and be cooled by the remaining fabric cards. Each fan is internally redundant and the fans on the fabric cards 12 can be individually hot swappable without removing the fabric card 12 itself. The fabric card 12 and line card 14 slots can be provided with blockers to inhibit reverse airflow when a card is removed. Empty line card 14 and fabric card 12 slots can be loaded with filler panels that prevent air bypass.

[0103] Each power supply has an internal fan that provides cooling for each power supply. Fans at the power supplies pull air through chassis perforations at the rear, across the chassis management cards 18, and through the power supply units 16. Chassis management card cooling is naturally redundant as multiple power supply units cool a single the chassis management card.

[0104] It will be appreciated that changes and modifications to the above described examples are possible. For example, although in the present example cooling is provided by drawing air from the front to the rear, in another example cooling could be from the rear to the front.

[0105] Also, although in the above described examples the fabric cards and the switch cards are described as being orthogonal to each other, they do not need to be exactly orthogonal to each other. Indeed, in an alternative example they could be angled with respect to each other but need not be exactly orthogonal to each other.

[0106] Also, in the above described examples the midplane connector pairs 32 are configured as first and second connectors 62 and 64, in another example they could be configured as a single connector that is assembled in the midplane. For example, through connectors could be provided that extend through the midplane vias. The through connectors could be manufactured to be integral with a first connector frame (e.g., a U-shaped frame or a box-shaped frame as in FIGS. 19 and 23, respectively) and the contacts inserted through the vias from a first side of the midplane 30. Then a second connector frame could be inserted over the connectors on the second side of the midplane 30 in a mutually orthogonal orientation to the first connector frame.

[0107] An example cable-based switch chassis can provide a very large switch having, for example, one or more of the following advantages, namely a 3456 ports non-blocking Clos (or Fat Tree) fabric, a 110 Terabit/sec bandwidth, major improvements in reliability, a 6:1 reduction in interconnect cables versus leaf and core switches, a new connector with superior mechanical design, major improvement in manageability, a single centralized switch with known topology that provides a 300:1 reduction in entities that need to be managed.

[0108] When constructing large Clos based fabric topologies, as described above, it is desirable to minimize the number of cables as well as the number of individual switch chassis instances that has to be involved. For those reasons, it is desirable to have a single large switch chassis with high enough radix (number of ports) to enable connectivity to all relevant number of end-nodes in the set of possible target cluster configurations.

[0109] As mentioned in the introduction, due to various physical constraints (e.g., maximum board size, number of individual switching elements, stages and associated connectivity, power and cooling capacity, connector sizes and cable management) it may not be possible to construct a single switch with sufficient radix. Also, the entry level cost of the largest possible switch may make it less attractive or unattractive for configurations that do not utilize the full radix.

[0110] In order to provide non-blocking bi-section bandwidth, a higher radix implies more switching stages and higher end-to-end latencies. However, for blade based end-node implementations, it is desirable to have a built-in switch as part of the blade chassis configuration since this provides internal connectivity without any cabling requirement, and therefore higher reliability. Also, in the case where connectivity to external nodes is required (i.e. the typical case for more than trivial-sized configurations), the existence of an internal switch implies redundancy of up-link connectivity and therefore inherently higher reliability in that no individual cable represents a single point of failure. A built-in, or internal switch also provides a first (leaf) switching tier (and initial stage) and therefore enables a second (core) switching tier to be implemented using multiple independent subsystems, e.g., multiple independent switch chassis instances.

[0111] As long as a single cable is used to connect each required pair of ports in a two tier Clos topology, the number of cables is the same in this configuration as in the case of a single large switch. Since the required radix of the core switch is much less in the two tier configuration than in the single switch configuration, it is typically possible to reduce the number of stages in the core switches so that the total number of stages in the fabric configuration is also the same in both cases.

[0112] However, the cabling complexity of the two tier case is significantly higher than in the single switch case in that the up-link connectivity of the leaf (tier 1) switches is evenly distributed among all the core switches to provide dispersive routing.

[0113] Routing may be dispersive within a single switch, whereby there may be no dispersive cabling requirement. In such a case, it is may be possible to aggregate multiple links within a single cable, which can further significantly reduce the number of cables in the system relative to a default case.

[0114] In a two tier configuration, such link aggregation can be used if the radix of the core switches is high enough to implement a Clos topology with the required number of end-nodes. Hence, there is an inherent conflict between minimizing the number of cables and cabling complexity, and minimizing the total number of switching stages in the fabric as well as the number of individual switching elements (and stages) in the core switch implementation.

[0115] One aspect of an invention described herein provides a solution to the dilemma outlined above, whereby different approaches can be integrated in a single solution. In order to reduce the number of cables, link aggregation within a cable can be used. In one example, the number of cables is minimized and the highest practical link aggregation per cable is used.

[0116] By introducing a dependency on an integrated leaf switch implementation on the end-node side, it is possible to use core switch chassis configurations where each cable connector with L individual links connects to L individual (parallel) switch configurations internally in the switch chassis. In this way, the value of L and the radix N of each individual internal parallel switch infrastructure can be selected in order to minimize number of cables and cabling complexity, while
maintaining the lowest possible total number of switching stages and individual switching elements in each core switch chassis.

[0117] Where the leaf switches have a number of up-links that is divisible by L, any configuration that could have been supported based on N×1 radix core switches can be supported with switch chassis instances representing L individual N radix switches. The L×N approach can typically be implemented with lower end-to-end latency and less cost and power requirements due to fewer stages and fewer switching elements.

[0118] An embodiment of the invention can thus provide a non-blocking Clos based interconnect fabric that takes advantage of multi-pathing aspects of embedded leaf switches in computer blade systems in order to increase system reliability but without imposing multiple core switch chassis. Also, this can be achieved with the same (minimal) number of switching stages (and individual switching elements) that a configuration based on a single large switch and no leaf-switches would imply. Hence, the advantages are achieved without inherent additional cost, or power-consumption, nor any additional worst case end-to-end latency.

[0119] An example embodiment of this aspect of the invention can provide a single core switch chassis 200 with multiple independent internal switches such that the existence of individual independent internal switches is transparent to a system administrator when combined with embedded leaf-switch based host implementations.

[0120] This transparency means that the radix of the complete switch is observed as the sum of the radices of the individual internal switches, whereby the imposed management model is that of a single core switch rather than multiple core switches.

[0121] In one example, illustrated in FIG. 33, a multi-switch chassis 200 can be provided in which line-cards 202 and fabric cards 204, 206, 208 are used, whereby the fabric cards represent a single switching stage and have independent switch elements 226, 228 and 230 (i.e., the switch elements are not interconnected on the fabric card). For ease of illustration in FIG. 33, only one line card is shown, although it should be appreciated that a plurality of line cards can be provided in an example implementation. Also, in FIG. 33, three fabric cards are shown, although two or more than three fabric cards may be provided.

[0122] In this example, only one switch tier is provided on a line-card, whereby there is no connectivity between switch elements on the line-cards (i.e., all connectivity is defined by and is via the fabric-cards).

[0123] In this example, assuming L×1 links per cable 210 and connector 212 and L individual switch fabrics within the chassis, the switch elements on the line cards are divided into L groups (e.g., in FIG. 33, L=3). The L individual links within the cable 210 and connector 212 on a line-card 202 are each connected to individual switch-elements 214, 216 and 218, each belonging to a separate group. The “up-links” 220, 222 and 224 from each line-card 202 to the fabric-cards 204, 206 and 208, and connectivity to the switch elements 226, 228 and 230, respectively, on the fabric cards are such that each switch element 226, 2278 and 230 on the fabric card only connect to switch elements 214, 216 and 218 on the line cards belonging to the same group. This implies that for L groups (individual switching fabrics), it is possible to organize the connectivity through the mid-plane so that some fabric cards are dedicated to a single group or that each fabric card has (independent) connectivity for all the groups.

[0124] Accordingly, there has been described a switch system in which L groups of the line switch elements are connectable to cables that include L links such that each of the L links within a cable connect to a switch element of a respective one of the L groups. Fabric switch elements are connected such that a fabric switch element is connected to the line switch elements of one of the group of line switch elements. In an example embodiment of the invention the radix of the complete switch can appear as the sum of the radices of the individual fabric switches.

[0125] It will be appreciated that changes and modifications to the above described embodiments are possible with the scope of the claimed invention.

[0126] For example, in the example illustrated in FIG. 33 three individual switch chips 214, 216 and 218 form three parallel switching units, each connected to one link of a 3-link cable. It will be appreciated that in other embodiments a different number of parallel switching units can be provided connected to a different number of links in a multi-link cable. Also, in another example, a single switching stage may be provided, whereby a single switching stage on a line card may interconnect individual links of a multi-link cable to respective fabric switch elements.

[0127] An example embodiment can facilitate the provision of a very large single switch chassis that can provide, for example one or more of the following advantages, namely a 3456 ports non-blocking Clos (or Fat Tree) fabric, a 110 Terabit/sec bandwidth, major improvements in reliability, a 6:1 reduction in interconnect cables versus leaf and core switches, a new connector with superior mechanical design, major improvement in manageability, a single centralized switch with known topology that provides a 300:1 reduction in entities that need to be managed.

[0128] Although the embodiments above have been described in considerable detail, numerous variations and modifications will become apparent to those skilled in the art once the above disclosure is fully appreciated.

What is claimed is:

1. A switch system comprises plurality of line switch elements and a plurality of fabric switch elements, wherein L groups of the line switch elements are connectable to cables that include L links such that each of the L links within a cable connect to a switch element of a respective one of the L groups and wherein one of the plurality of fabric switch elements connects to line switch elements of one of the group of line switch elements.

2. The switch system of claim 1, wherein each of the plurality of fabric switch elements is connected to the line switch elements of a respective one of the group of line switch elements.

3. The system of claim 1, wherein any interconnectivity between the line switch elements is via the fabric switch elements.

4. The switch system of claim 1, wherein the fabric switch elements form independent switches.

5. The system of claim 4, wherein each fabric switch element has N ports, and the system has N×L switch ports.

6. The system of claim 1, forming a non-blocking Clos-based interconnect fabric.

7. A switch system comprising:
   - at least one line card connectable to at least one cable that comprises L links, the at least one line card comprising
L groups of line switch elements configured such that each of the L links within the at least one cable connect to a line switch element of a respective one of the L groups of line switch elements; and at least one fabric card connected to the at least one line card and comprising fabric switch elements connected such that a fabric switch element connects to line switch elements that belong to one of the L groups.

8. The system of claim 7, wherein the at least one fabric card represents a single switch tier and the fabric switch elements of the at least one fabric card form independent switches.

9. The system of claim 7, wherein interconnectivity between the line switch elements is via the fabric switch elements.

10. The system of claim 7, wherein the line cards comprise cable connectors for the attachment of line cables.

11. The system of claim 7, comprising a plurality of fabric cards, wherein connectivity is provided such that at least one of the fabric cards is dedicated to a single one of the L groups.

12. The system of claim 7, comprising a plurality of fabric cards, wherein connectivity is provided such that each fabric card has connectivity for all the L groups.

13. The system of claim 7, comprising L fabric cards.

14. The system of claim 7, wherein each switch element on the at least one fabric card has N ports, whereby the system of claim 1 has an N*L switch ports.

15. The system of claim 7, comprising a chassis that includes the at least one fabric card and the at least one line card interconnected via a midplane.

16. The system of claim 7, wherein a number of line cables is minimized and a highest practical link aggregation per cable is used.

17. The system of claim 7, forming a non-blocking Clos-based interconnect fabric.

18. A method of implementing a switch system in a switch chassis, the method comprising:

connecting at least one cable that comprises L links to at least one line card that comprises L groups of switch elements configured such that each of the L links within the at least one cable connect to a switch element of a respective one of the L groups of switch elements of the at least one line card; and connecting the at least one line card to at least one fabric card that comprises switch elements such that links from the at least one line-card to the at least one fabric-card and connectivity to the switch elements of the at least one fabric card are such that each switch element on the at least one fabric card only connects to switch elements on the line cards that belong to the same group.

* * * * *