

SCI and High Availability

This document gives an overview of the High Availability (HA) features of the SCI interconnect. It shows that the 2D and 3D Torus architecture provides unique HA features within a single network.

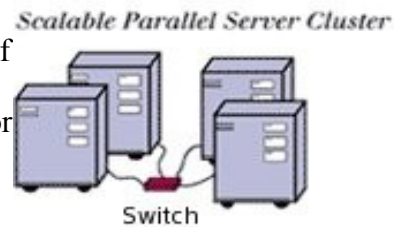
Ring topology

The basic SCI communication structure is the ring topology. Large number of compute nodes can be connected in a single ring, just by connecting the output port on one card to the input port on another card. Data is passed on the SCI ring with very low latency, each node typically adds 30 nanoseconds (ns) to the packet latency. The topology is cheap and scalable since you always can add just another node and don't have to worry about available switch ports. All you need is just another card and a cable. The major drawback with the ring topology is that the ring communication requires all nodes to be powered on and the ring to be fully connected. The configuration is widely used within embedded applications where all nodes are treated as a single machine and will protected within a box/rack, but is useless to build a HA cluster.

SCI Switch topology

The 8 port SCI switch (D535) has been the most common way to improve availability in SCI systems.

The switch guarantees full connectivity between nodes if a one node is down. The HA issue with this topology is that no nodes can communicate if the switch is broken or powered off. This property is shared with all other communication equipments relying on a single centralized. The common way around this problem is to



duplicate the network by installing two SCI cards in every system, and connecting each card to a separate switch. This is the traditional powerful HA configuration. It has no single point of failures, full connectivity is guaranteed independent of node or network failures. There are no fail-over delays since the other network immediately is available. SCI is specially well suited to build such networks since a SCI failure is detected in the range of microseconds and immediately fail over can be established. Dolphin SuperSockets takes advantage of the duplicated network and provides higher throughput while both networks are available. The drawback is the price since you need to duplicate all network components.

SCI 2D / 3D Torus Clusters

2 to hundreds of nodes can be interconnected using the 2 dimensional (2D) SCI Torus topology using D334 or D352 (two port SCI-PCI or SCI-PCI Express) adapter cards or the 3 dimensional (3D) topology using D346 (Tree port PCI-SCI) adapter cards.

Each adapter card in the cluster implements a small link level distributed¹ SCI switch. The traffic in the cluster is efficient routed through multiple redundant paths to the destination. The [SCI network manager](#) manages the routing tables and will automatically reroute the network in the case of failing nodes or cables and full connectivity is

¹ The switch functionality is implemented in the low level SCI Link Controller and does not affect or require any host resources.

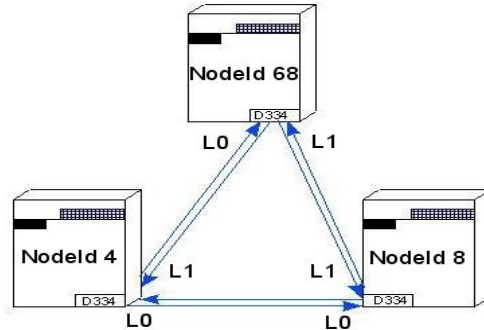
guaranteed as long as there is at least one operational SCI cable between the systems.

2 Node cluster

Two nodes can be interconnected using any kind of SCI cards, one in each machine and two SCI cables. The D350 card includes two separate SCI channels and will provide transparent channel bonding and internal fail over. Channel bonding will also be enabled if two regular cards are installed in each machine.

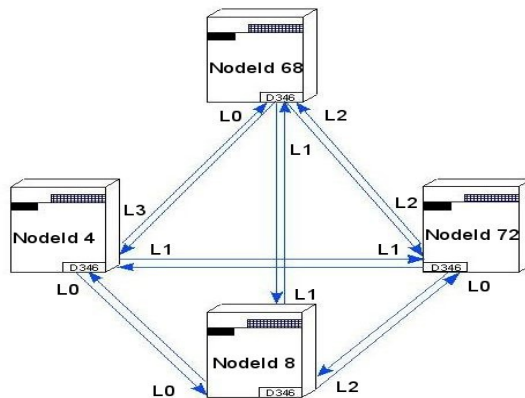
3 Node cluster – 2D Direct HA

By using one D334 or D352 (two port SCI-PCI or SCI-PCI Express card) in each server, you can build a 3 node cluster where full connectivity is guaranteed if a single SCI connection is broken or if one of the nodes is dead. Each node is connected directly to each of the other two nodes using separate SCI rings. Data will be transmitted the shortest path as long as all cables are operational. When a cable failure is detected, all communication will be switched over to the remaining path. A dead node failure will not cause any delays in communication. A cable problem will be detected and handled locally on the nodes in less than 200 millisecond. The L0, L1 and L2 indicates link connectivity.



4 Node cluster – 3D Direct HA

By using one D346 (three port SCI-PCI card) in each server, you can build a 4 node cluster where full connectivity is guaranteed if a single SCI connection is broken or of if one of the nodes is dead. Each node is connected directly to each of the other tree nodes using separate SCI rings. Fail over properties are identical as for the 3 Node cluster HA description above.



Typical failures

Electronics in modern computers are normally very reliable, failures are often traced back to electro-mechanical components like hard-drives, fans and power-supplies. These failures either brings down the system immediately or the next morning when the system administrator wants to do some corrective actions. Clustering hardware have the same behavior, electronics runs forever, switches may fail since they have fans and power

supplies. The HA properties of a computer or a switch can be improved by including hot swappable power supplies and fans, but this will also increase the cost. Cables are very reliable given they are properly protected from external incidents (Cables in a HA system should be strain relieved).

We will therefore argue that it is possible to build a HA system without duplicating the network as long as the clustering components used does not contain power supplies or fans and that the network are able to handle node failures. All SCI configurations can be duplicated if maximum HA is required.

The degree of HA required for a given application must be determined comparing costs versus benefits and requirements. Some life critical applications will always be built using a maximum HA configuration. Some other applications will be well off enabling HA for the most common failures. The fail over time is also an important property.

If you have any questions please don't hesitate to email pci-support@dolphinics.com