

## Recent Advances in Cluster Networks

Charles L. Seitz  
Myricom, Inc.  
chuck@myri.com

### Abstract

*One of the significant advances in cluster networks over the past two years is that it is now practical to connect up to tens of thousands of hosts with networks that have enormous and scalable total capacity, and in which the communication from a host to any other host has the same cost. This same-cost property is desirable because it allows computing processes to be assigned to hosts according to cluster-management or load-balancing considerations, and without regard for the mapping of the communication patterns of the computation to the network topology.*

*The way in which this advance was accomplished is an elegant blend of theory and technology.*

*It is no great trick to connect any number of hosts with a topology such as a mesh or torus. If you examine the capacity of such networks under arbitrary traffic patterns, and for larger and larger networks, you will notice that the number of internal communication links does not grow with the number of hosts. This internal capacity can be formalized in terms of the minimal number of communication links cut by any line that bisects the hosts. This metric is called the minimal bisection.*

*For the scalable Clos networks that Myricom supplies, the minimal bisection is as large as possible, considering that host links may also be included in the cuts of the network. Hence, a Clos network is said to have "full bisection." The Clos network, first described by Charles Clos in a paper published in 1954, has*

*several other properties that make the Clos topology ideal for a cluster network.*

*With this theoretical understanding of Clos networks, how do we make them practical? Clos networks require a lot of switches, and a lot of switch-to-switch communication links. For example, a 1024-host Clos network requires 320 16-port switches, with 80% of the switch ports connecting to another switch, and only 20% connecting to hosts.*

*The simple answer to this question is to start with inexpensive switches, ideally single-chip switches, and to package them in a way that provides most or all of the switch-to-switch links on an inexpensive wiring medium - circuit-board traces, not cables.*

*Of course, there are open research problems. In order to reach the ideal of a network that handles all traffic patterns equally well, one can take advantage of another property of the Clos network. The Clos topology provides multiple routes between hosts, and all shortest routes are deadlock-free. When a host interface can send successive packets to another host along multiple routes, the traffic is dispersed in a way that statistically avoids "hot spots," high utilization of specific network links that can result from single-route mappings of the communication patterns of application programs to the topology. This same dispersive routing technique can provide fault tolerance on a much smaller time scale than by periodic mapping, and can exploit multiple ports on host interfaces. There may be one or two excellent Ph.D.-thesis topics lurking here.*