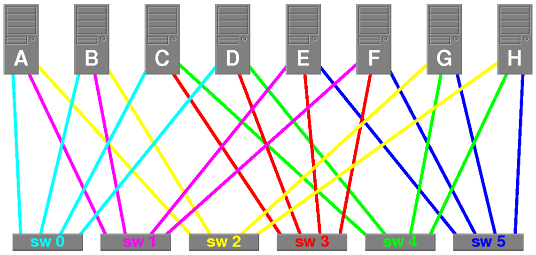
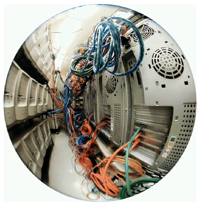
Flat Outperformance

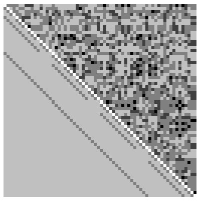
INTERCONNECTION NETWORKS, sometimes called SYSTEM AREA NETWORKS (SANs), play a critical role in all types of parallel computers – be they clusters spanning many racks or multiple cores on the same chip. Although commodity hardware and straightforward topologies are sometimes effective, communications within parallel programs tend to have specific properties that allow a well-engineered network to dramatically outperform the obvious alternatives.

Point-To-Point Communication. By far the most commonly discussed type of communication is that in which each node is sending a message to one other node. For parallel systems, the bandwidth of an individual wire is generally less important than the bandwidth available if all processors are communicating simultaneously – the *bisection bandwidth*. *Latency* also becomes critical; it might be possible to hide some communication delays by performing unrelated computations while waiting, but latency sets the fundamental limit on the smallest grain size for which speedup can be obtained.

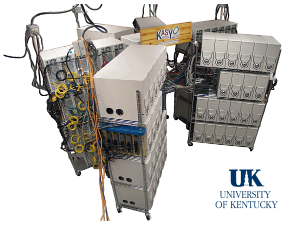


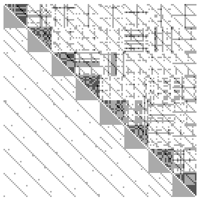
FLAT NEIGHBORHOOD NETWORKS (FNNs). FNNs provide single-switch latency and better bisection bandwidth than a FAT TREE with comparable hardware complexity by using multiple network interfaces per node. As shown above for 4-port switches connecting 8 nodes, an FNN connects nodes to switches such that each node pair has at least one switch in common. The best wiring pattern usually is *asymmetric*; the design is *evolved* from random wiring patterns using a genetic algorithm.



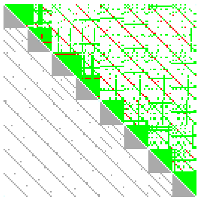


FNN design problems and solution quality are summarized by square maps in which the darkness of each point indicates how many single-switch paths exist between the corresponding pair of nodes; the lower left triangle is the minimum design requirement, the upper right is what the FNN design actually provides. The map above shows the world's first FNN, which in April 2000 connected the 66 nodes of KLAT2 (KENTUCKY LINUX ATHLON TESTBED 2) using 31-port 100Mb/s ETHERNET switches and standard IP to deliver close to 25Gb/s bisection bandwidth, and $30\mu s$ latency, at a network cost of about $8,100.





SPARSE FLAT NEIGHBORHOOD NETWORKS (SFNNs). Surprisingly, very few parallel programs depend on every node talking to every other; usually, each node talks to at most $O(log(N))$ other nodes. This even is true using personalized all-to-all as MPI typically implements it. By ensuring single-switch latency *only for node pairs that are expected to communicate*, SFNNs can provide single-switch latency for all critical communications in a typical application suite using cheap, narrow, switches for systems having many thousands of nodes. The first SFNN was KASY0 (KENTUCKY ASYMMETRIC ZERO), built in 2003, covering 128 nodes using 24-port switches – at about half the network cost of KLAT2's FNN.



FRACTIONAL FLAT NEIGHBORHOOD NETWORKS (FFNNs). Although SFNNs have great price/performance, the search is driven by the performance; FFNNs flip priorities, finding the best coverage possible with a fixed network cost. For example, using far less hardware than KASY0, the above map shows coverage of an FFNN in **green** – only the **red** spots deliver poorer latency.

Note that FNN, SFNN, and FFNN topologies all are fully compatible with ETHERNET, IP, and most other commonly available network technologies and protocols. Design tools, including interactive WWW forms, are freely available online at **`Aggregate.Org`**.

This document should be cited as:

```
@techreport{sc08flat,
author={Henry Dietz and Krishna Prabahala},
title={Flat Outperformance},
institution={University of Kentucky},
address={http://aggregate.org/WHITE/sc08flat.pdf},
month={Nov}, year={2008}}
```

Aggregate.Org
UNBRIDLED COMPUTING
SC08 AUSTIN, TX