Multicasts for Faster Science Applications on Beowulf Clusters

Peter Tamblyn[1,2], Hal Levison[1], Erik Asphaug[3]

[1] Southwest Research Institute

[2] Binary Astronomy, LLC

[3] University of California, Santa Cruz

Supported by NASA Applied Information Systems Research Program

`http://www.boulder.swri.edu/˜ptamblyn/ais/`

American Astronomical Society; January 9, 2003

Beowulf “Super Computers”

Networked set of cheap “off-the shelf” computers working together on a problem.

All we need to know:

- Cheap
- Popular
- Commonly Ethernet based
- Cross Disciplinary

The Problem

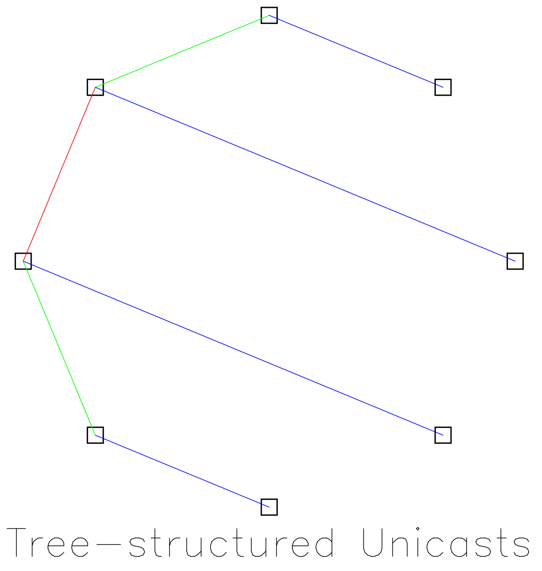
Communications are **much** slower than calculations

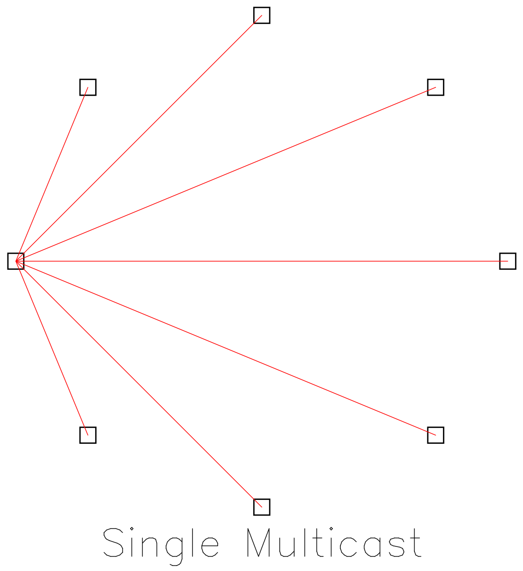
Beowulfs are

- *excellent* for compute-bound simulations
- *adequate* for many simulations
- *dreadful* for communication-bound simulations

Our motivating astronomy problem is a **worst case** for Beowulfs: *every* node needs to know about *every* particle at *every* timestep. Broadcast bound.

Our Solution



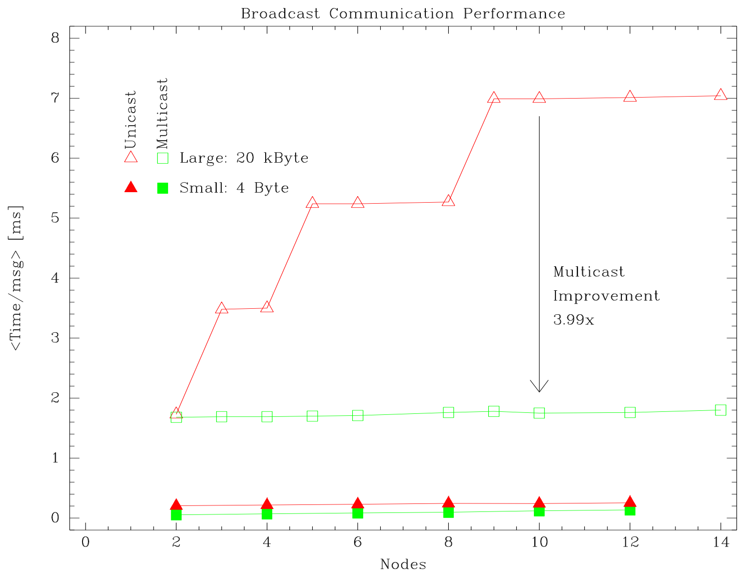


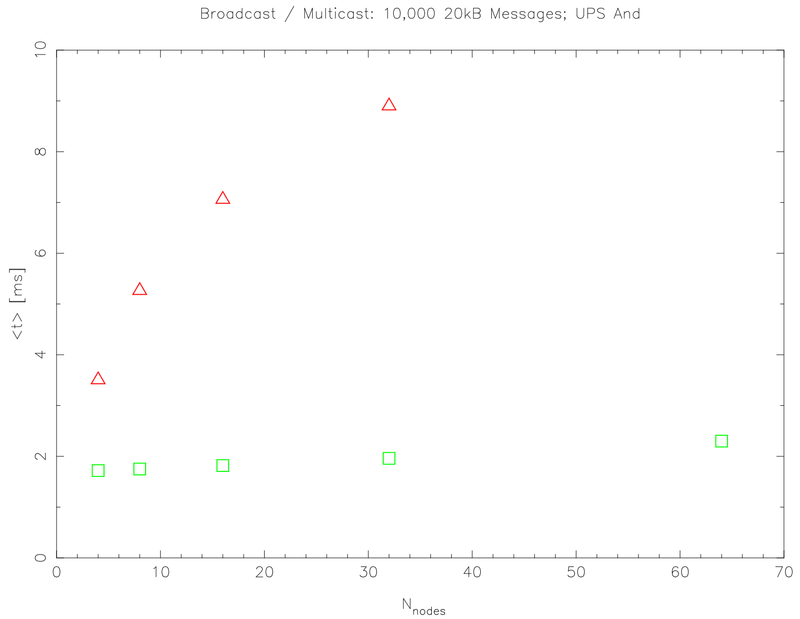
Send message to **all** nodes simultaneously instead of node-to-node.

Support **reliable** multicasts seamlessly under MPI.

Easy to use: no kernel, OS, or application changes required.

Raw Communication Results



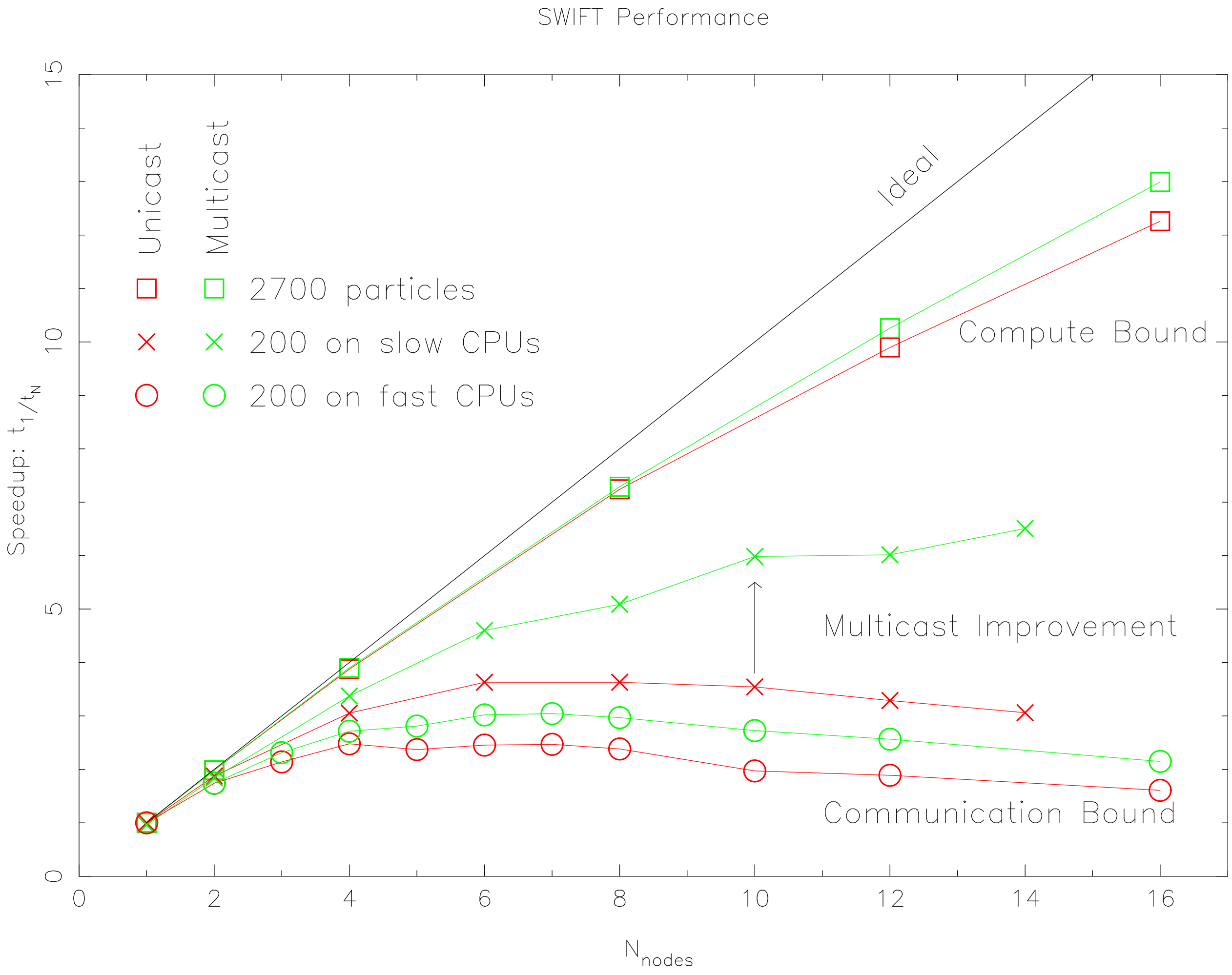


Standard MPI broadcasts require at least $\log_2(N_{nodes})$ communication cycles

Multicasts have no significant additional delay for larger clusters

Nearly perfect scaling: 5% cost at 64 nodes

Science Application Results



Summary

Reliable multicasts provide efficient, scalable alternative to TCP broadcast trees over common Ethernet hardware.

Emphasis on trivial use with existing message passing applications. No changes to hardware, operating system, or application code required.

Implications

- Not important for compute-bound or domain-isolated parallel programs
- We can make some global domain problems much faster
- Broadens the class of problems appropriate for Beowulfs
- Easier to create adequate parallel programs:
  - Global messages with same cost as node-to-node messages
  - Domain decomosition still useful but not vital
  - Broadens the class of potential Beowulf programmers