Specialized Network Topologies for Efficient Communication in Computer Clusters

Urban Borštnik, Milan Hodošček, and Dušanka Janežič

urban@cmm.ki.si

National Institute of Chemistry

Ljubljana, Slovenia

Why Clusters?

Use in computational methods

- Replace traditional supercomputers.
- Low cost.
- (Off-the-shelf) availability.

Precursors are networked workstations.

Developed in the last couple of years

- Personal Computers attained supercomputer performance.
- Cheap & fast networking equipment.

Networking Clusters

Networking speed and latency problems.

Switches

- widely-used;
- easy implementation;
- everyone communicates with everyone.

However

- costly;
- not expandable.

Networking Clusters: Special Topologies

Special topologies

- Point-to-point connections form topology.
- Various types (mesh, hypercube, full graph, . . .).
- Routing: practically needs a switch.
- Special software to take advantage of topology (software conforms to hardware).

Or, we can design a topology to effi ciently perform some types of data transfer. (Hardware conforms to software.)

Topologies

Logical topology

- Describes the software communication pattern.

Physical topology

- Describes the physical connection pattern.

The physical topology should cover at least the logical topology.

CHARMM and CMPI

CHARMM for molecular dynamics simulation of macromolecules

- Primarily uses 2 data-transfer operations.
- Uses own CMPI library for collective operations.

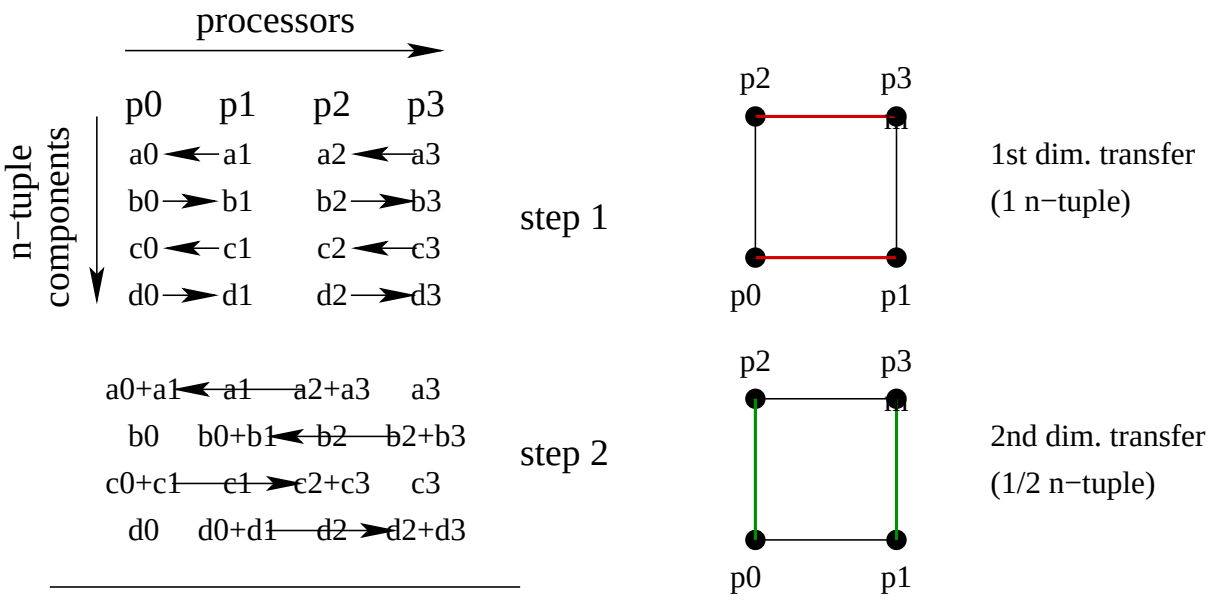
Distributed Vector Global Sum, Distributed Vector Global Broadcast

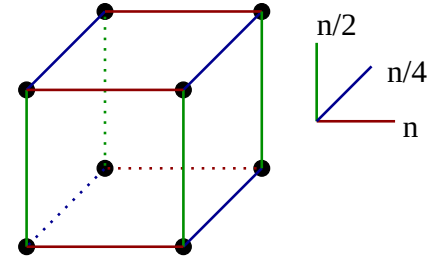
- DVG Sum: Vectors added, resulting vector left scattered across nodes.
- DVG Broadcast: Scattered vector broadcast to all nodes.

CMPI Collective Operations

Effi cient hypercube implementation

- Different amount of data is transferred in each dimension.





Hierarchical Hypercube

A Hierarchical Hypercube topology is a hypercube with links of different speeds for different dimensions.

- DVG Sum & Broadcast transfer $^1/_2$ data in each successive dimension.
- Hierarchical hypercube has the fastest links for the 1^{st} dimension, and progressively slower for the other dimensions.
- Ideally no link saturated, no link with spare capacity.

CROW5

Uses hierarchical hypercube topology to connect 16 SMP PCs

- Dual Athlon MP-1600+
- Gigabit ethernet cards
- Fast ethernet switch

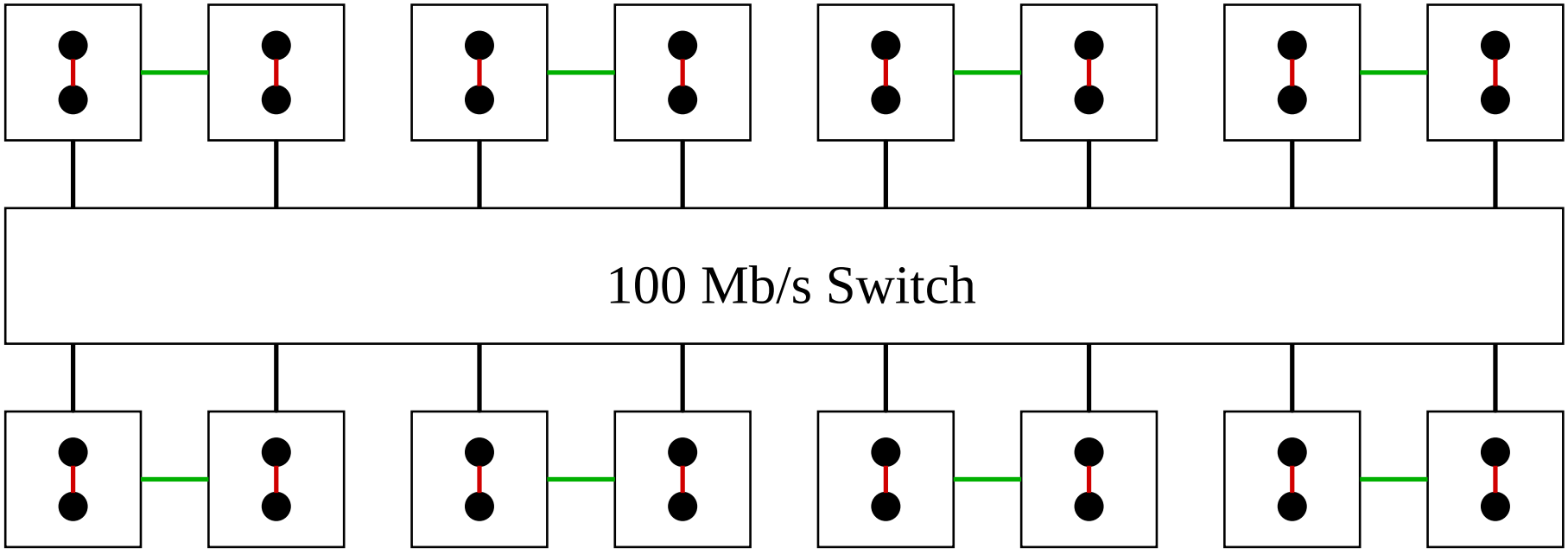
1st dimension: system bus.

2nd dimension: point-to-point gigabit Ethernet.

3rd–5th dimensions: fast Ethernet switch.

(Alternate view: increasing performance of switched PCs with point-to-point links.)

CROW5 topology



1st dim.: Bus (>2660 Mb/s)

2nd dim.: Gigabit Ethernet (1000 Mb/s)

3rd, 4th, 5th dims.: Fast Ethernet (100 Mb/s)

Comparison to Other Network Types

Benchmark: 100 step dynamics simulation of protein HIV Integrase, 1 fs step size.

Table of speedups (as compared to 1 processor):

Topology	Number of Processors				
	2	4	8	16	32
Hier. Hypercube	1.8	3.0	4.4	7.1	9.2
1 Gb/s switch	1.8	3.1	4.6	8.0	10.7
100 Mb/s switch	1.8	2.9	4.0	5.4	6.3

Comparable to a gigabit switch.

Faster than just an ethernet switch.

Redundancy Features

Computers connected with both point-to-point links and hub/switch.

- Redundant links if one fails.
- Possible redundancy between any computers in the cluster.

Conclusions

- Clusters are very suitable for numerically intensive calculations.
- An appropriate topology can provide a noticeable performance improvement.
- We developed a hierarchical hypercube topology based on communication patterns in software to speed up computation.
- The hierarchical hypercube offers a performance improvement over standard switching technology for only a small additional cost.

Acknowledgement

Ministry of Education, Science, and Sport of Slovenia