CSE 557 Presentation
4/12/2000, 9:25 am

Beowulf: A Parallel Workstation for Scientific Computations

by

Thomas Sterling, Donald J. Becker, John E. Dorband, Daniel Savarese, Udaya A. Ranawake and Charles V. Packer

(early 1995)

Presented by

Anirudh Modi

Overview

- Results from experiments measuring the scaling characteristics of Beowulf clusters:
  - communication bandwidth
  - file transfer rates
  - processing performance
- Evaluation uses:
  - a computational fluid dynamics (CFD) code
  - an N-body gravitational simulation program
- Aims to show that Beowulf architecture provides a new operating point in price/performance for high performance computing

Beowulf Architecture

- 16 single processor machines with Intel 486 DX4 100MHz chip (16 KB primary cache, 256 KB secondary cache on the motherboard) [iComp index: 435, for P-III 1 GHz = 3280]
- 16MB RAM on every node [256 MB total]
- 500 MB hard disk on every node [8 GB total]
- Two 10baseT ethernet adapters on every node
- Operating system: Linux (kernel v 1.1), with PVM for parallel programming
- *Note*: This work was done in *1994*, when this was top of the line PC architecture!!!

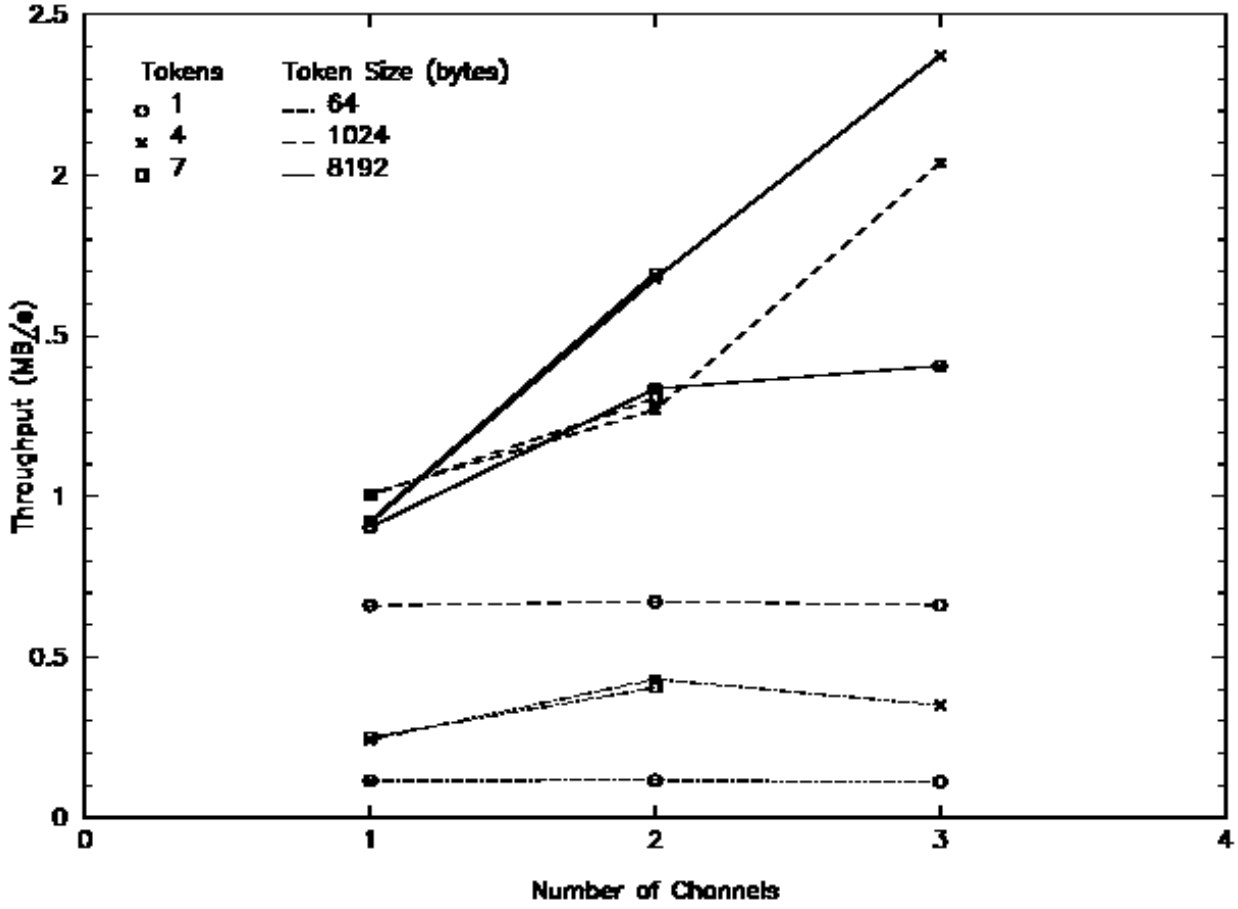
Internode Communications

- Several ethernets on each node are tied by *channel bonding* at device driver level.
- Routing of packets is thus transparent to the calling program, since the ethernet device driver takes care of this.
- The device driver was written by one of the authors, Donald Becker.
- User Datagram Protocol (UDP) was used to perform token exchanges used to measure network throughput.
- Processors communicate in pairs for the benchmark.

Internode Communications

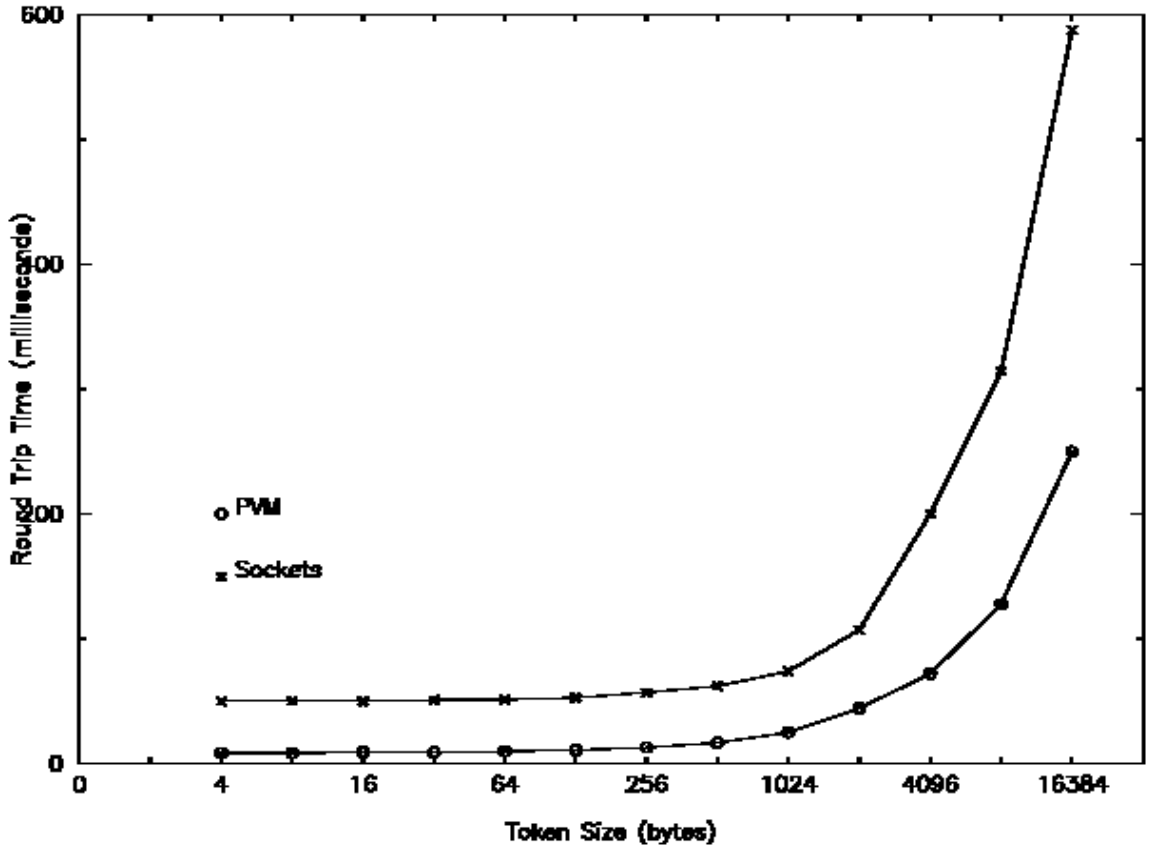
- Theoretical peak speed of 10baseT ethernet: 10 Mbits/sec = 1.25 MB/sec
- # of channels was varied from 1-3 for the tests (only 4 pairs of processors were used for the 3 channel test owing to the lack of more ethernet cards)
- Token size was varied from 64 bytes to 8192 bytes
- *Rountrip time* was also used as a performance characteristic: defined as time taken by the token to return to the original sender after visiting all the intermediate nodes once (15 visits in this case)

Internode Communications



(Note: Higher throughput is better)

Internode Communications



(Note: Lower Round Trip time is better)

Internode Communications

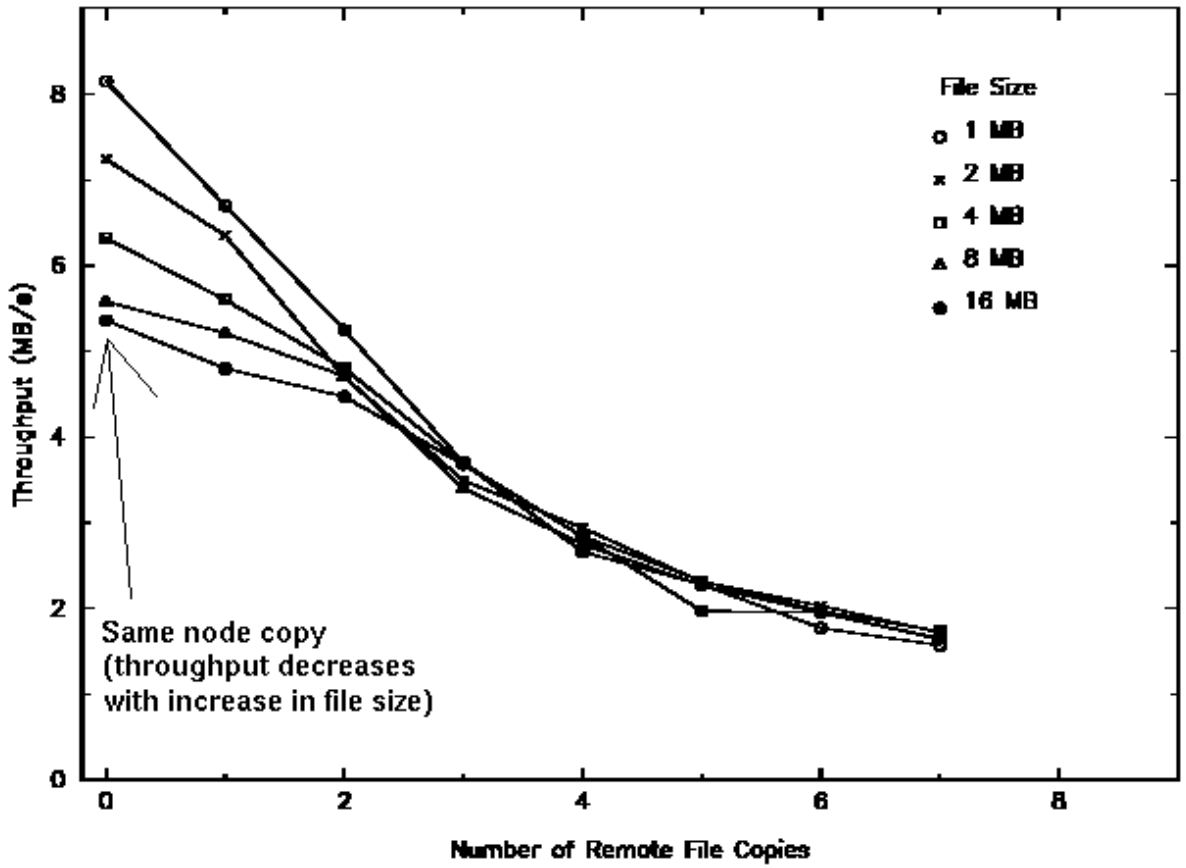
Performance summary:

- Peak throughput obtained was:
  - 1 channel: 1.0 MB/s (80% of theoretical peak)
  - 2 channels: 1.7 MB/s (68% of theoretical peak)
  - 3 channels: 2.4 MB/s (64% of theoretical peak)
- 4-pair 8192 byte token gives best network throughput followed by 4-pair 1024 byte case
- Rountrip time is constant (10ms for socket connection, 60ms using PVM) till a token size of 1024 bytes, after which it increases exponentially!
- A token size of 1024 bytes is *ideal* from the above results

Parallel Disk I/O

- Simultaneous file transfers involving varying file sizes were carried on across a mix of interprocessor copies and intraprocessor copies.
- File copies were performed using Unix *read()* and *write()* system calls.
- Remote file transfers were carried on using BSD sockets and UDP.
- File sizes were varied from 1 MB to 16 MB.
- Remote file transfers were done in pairs, hence upto 8 simultaneous copies were possible (only 7 could be executed since only 15 nodes were available at the time of the experiment).
- No processor was involved in more than one file transfer, therefore no disk or processor contention, only bandwidth contention.

Parallel Disk I/O

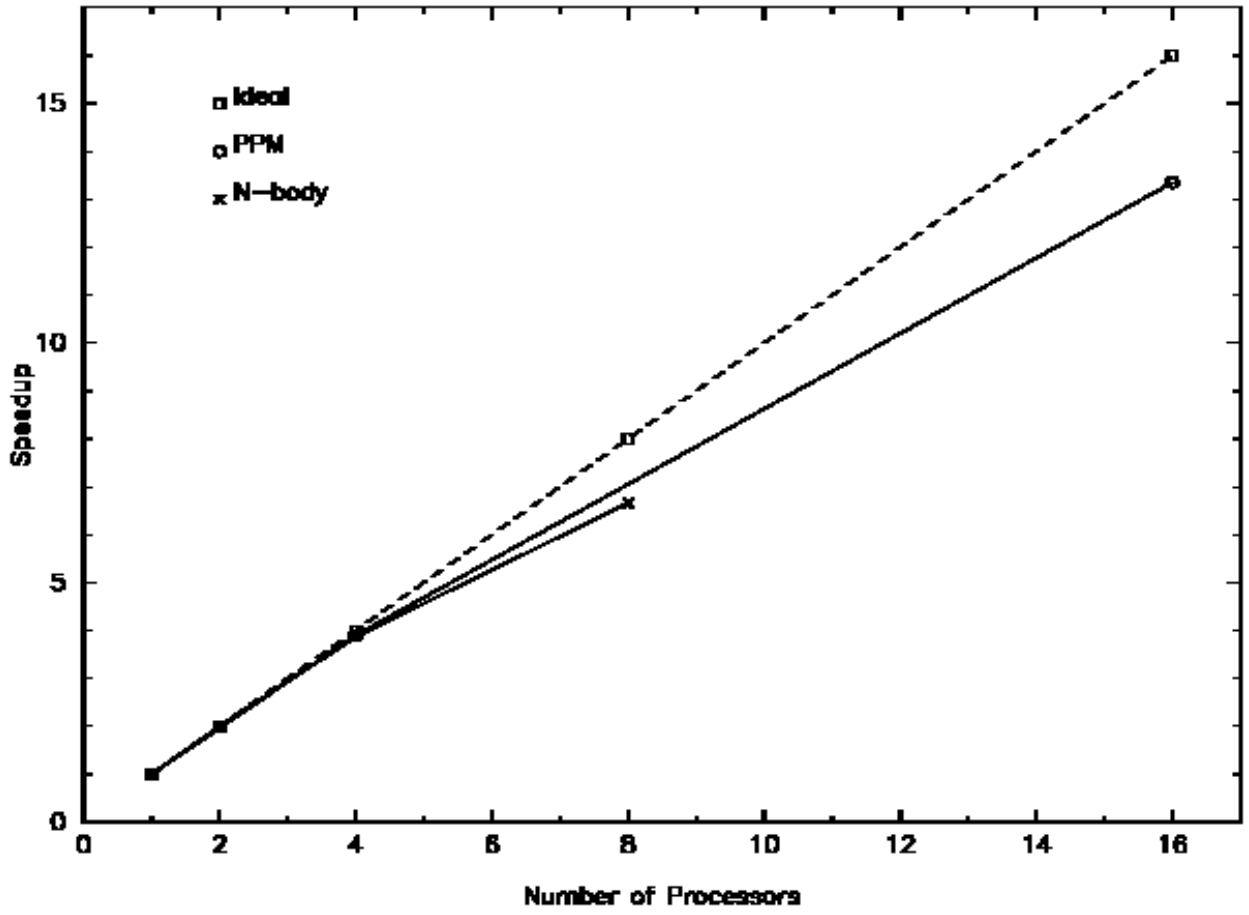


(Note: Higher throughput is better)

Benchmark of Scientific Codes

- A 2-D compressible fluid dynamics (CFD) code called Prometheus was used.
  - The code solves Euler's equation for gas dynamics
  - Uses structured rectangular mesh with Piecewise Parabolic Method (PPM)
  - Ported to PVM from its previous version running on IBM SP-1 and Intel Paragon
  - Domain decomposition was used for parallelization with 128x128 nodes for each processor.
- N-body simulation code was also used for the benchmark
  - Code used to study the structure of gravitating, star forming, interstellar clouds.
  - It has runtime of O(NlogN).
  - A range of particles from 32K to 256K were used.

Benchmark of Scientific Codes



(Note: Higher Speedup is better)

Benchmark of Scientific Codes

Summary of results:

- CFD simulation:
  - Good scaling characteristics (almost linear); performance for 16 processor case was just 16% below ideal.
  - Single processor performance was 4.5 Mflops, whereas performance with 16 processors was 60 Mflops (Speedup of 13.3).
  - Cray T3D was 2.5 times slower for 16 processor case!!
- N-body simulation:
  - Scaling characteristics poorer, but not bad; performance for 8 processor case was 19% below ideal.
  - Single processor performance was 1.9 Mflops, whereas performance with 8 processors was 12.4 Mflops (Speedup of 6.5).

Conclusions

- Interprocessor communication proved to be the most interesting aspect of the Beowulf cluster.
- Channel bonding showed excellent scaling factors.
- Parallel file transfers seemed to be limited by the network bandwidth.
- Scientific codes showed good performance and scaling characteristics.
- Excellent price/performance; about 10-20 times cheaper than supercomputers with similar performance.
- Architecture should benefit from future technological advances in networking, processor speeds, disk speeds, etc. and with improvements in software/OS.