

# Communication Overhead for Space Science Applications on the Beowulf Parallel Workstation

Thomas Sterling  
Center of Excellence in Space Data  
and Information Sciences  
Code 930.5 NASA Goddard Space Flight Center  
Greenbelt, MD 20771  
tron@cesdis.gsfc.nasa.gov

Donald J. Becker  
Center of Excellence in Space Data  
and Information Sciences  
Code 930.5 NASA Goddard  
Greenbelt, MD 20771  
becker@cesdis.gsfc.nasa.gov

Daniel Savarese  
Department of Computer Science  
University of Maryland  
College Park, MD 20742  
dfs@cs.umd.edu

Bruce Fryxell    Kevin Olson  
Institute for Computational Science  
and Informatics  
George Mason University  
{fryxell@neutrino, olson@jeans}.gsfc.nasa.gov

## Abstract

*The Beowulf parallel workstation combines 16 PC-compatible processing subsystems and disk drives using dual Ethernet networks to provide a single-user environment with 1 Gops peak performance, half a Gbyte of disk storage, and up to 8 times the disk I/O bandwidth of conventional workstations. The Beowulf architecture establishes a new operating point in price-performance for single-user environments requiring high disk capacity and bandwidth. The Beowulf research project is investigating the feasibility of exploiting mass market commodity computing elements in support of Earth and space science requirements for large data-set browsing and visualization, simulation of natural physical processes, and assimilation of remote sensing data. This paper reports the findings from a series of experiments for characterizing the Beowulf dual channel communication overhead. It is shown that dual networks can sustain 70% greater throughput than a single network alone but that bandwidth achieved is more highly sensitive to message size than to the number of messages at peak demand. While overhead is shown to be high for global synchronization, its overall impact on scalability of real world applications for computational fluid dynamics and N-body gravitational simulation is shown to be modest.*

## 1 Introduction

The Beowulf parallel workstation defines a new operating point in price-performance for single-user computing systems. Beowulf couples the low cost, moderate performance of commodity personal-computing subsystems with the emergence of de facto standards in message passing hardware and software to realize a 1 Gops workstation with exceptional local file storage capacity and bandwidth. This experimental system is motivated by requirements of NASA Earth and space science applications including data assimilation, data set browsing and visualization, and simulation of natural physical systems. It exploits parallelism in processor, disk, and internal communication, all derived from mass market commodity elements. Thus enabling large temporary data sets to be buffered on the workstation in order to reduce demand on shared central file servers and networks while greatly improving user response time. This paper presents results of experiments to characterize the communication overhead of the Beowulf parallel workstation and establish the regime of effective operation.

While most distributed computing systems provide general purpose multiuser environments, the Beowulf distributed computing system is specifically designed for single user workloads typical of high end scientific workstation environments. The Princeton Shrimp [2] project is also targeted to parallel workstation systems comprising multiple personal-computer proces-

sors. This Pentium based distributed computer employs a custom communication unit to support a distributed shared memory model. Beowulf, by contrast, incorporates no special purpose parts, depending instead on parallel ethernet communication channels to achieve adequate sustained interprocessor message transfer rates. This has required some software enhancements at the operating system kernel level but has been achieved with commercial off-the-shelf hardware elements, specifically low cost Ethernet cards.

Much of the workstation operational demand is very coarse grained job stream parallelism. But, as with all workstations, some of the required workload is computationally intensive. Thus, there is a need to exploit parallelism within a single application. Ironically, where the solving of the parallel processing problem would ordinarily prove a challenge, in the computational sciences community, many active applications have already been crafted in the communicating sequential processes parallel programming style in order to run effectively on larger distributed computers such as the Intel Paragon [7], the TMC CM5 [14], or the CRI T3D [3]. Beowulf benefits from this parallel programming investment within the community it is intended to serve and provides an equivalent programming and compilation environment at the parallel workstation level. A number of parallel applications have been successfully, sometimes even easily, ported to Beowulf in this manner.

Beowulf inter-processor communications is provided by standard 10 Mbps Ethernet using dual channels with each channel connecting all 16 processing elements. These channels are equally accessible to all processors and the operating system kernel (based on Linux [8]) has been modified to dynamically distribute message packet traffic to load balance across both networks. Interprocessor communications performance may be characterized in several ways and this paper presents experimental results reflecting these aspects of communication on execution performance. Basic network capacity is characterized in terms of throughput, both as byte transfer rate and number of messages passed per unit time. These measurements are presented as functions of message size, message demand, and number of channels employed (one or two). Finally, the impact of parallel interprocessor communication is explored through two real-world application programs from the Earth and space sciences community. One problem is a computational fluid dynamics application employing a regular static data structure well suited to a system of Beowulf's architecture. The second is an N-body gravitational simulation with ir-

regular dynamic global data that challenges the capabilities of Beowulf's communication. Both the scaling properties of these two applications and the communication overhead encountered will be presented.

## 2 Beowulf architecture

The Beowulf parallel workstation architecture comprises 16 PC processor subsystems, each with a half GByte disk and controller. The Beowulf prototype incorporates the Intel DX4 processor with a 100MHz clock rate and 256 KBytes of secondary cache. The DX4 delivers greater computational power than other members of the 486 family not only from its higher clock speed, but also from its 16 KByte primary cache (twice the size of other 486 primary caches) [6]. Each processing subsystem has installed 16 MBytes of DRAM for a total system main memory of 256 MBytes. The processing elements are interconnected by two parallel Ethernet networks with peak capacity of 10 Mbps per network. For purposes of experimentation, one network is twisted pair using a small, inexpensive hub while the other is multidrop thin-net. Two processor subsystems include separate ethernet interfaces to external LAN's for remote access and data transfer. Two additional processor subsystems include high resolution video controller/drivers, one of these also providing user keyboard