Performance analysis of the LA-MPI communications library

Richard L. Graham Mitchel W. Sukalski L. Dean Risinger
David J. Daniel Nehal N. Desai Marydell Nochumsen
Ling-Ling Chen

Los Alamos National Laboratory*
Advanced Computing Laboratory
MS-B287
Los Alamos, NM 87545 USA
lampi-support@lanl.gov

Abstract

This study reports on the performance of the Los Alamos Message Passing Interface library (LA-MPI) on a 128 processor SGI Origin 2000 computer. LA-MPI is an open source, high performance, MPI compliant communications library. Performance is evaluated using a sample scientific application software, CICE [1]. CICE is representative of the type of calculation and communication performed in modern scientific computing. LA-MPI performance is measured against the communications library provided by SGI and Argonne National Labs (ANL). Early results show that LA-MPI library performance is on par with SGI's well-optimized commercial offering.

1 Introduction

Successfully exploiting parallelism in scientific calculations requires a way for the different pieces of the calculation to exchange information. The software implementing this exchange is called the message passing layer. In the past, the primary challenge for designers and implementers of messaging libraries was the implementation of a low latency and high bandwidth library. However, with the introduction of commodity clustering and high performance networks other important criteria have emerged, namely resiliency and reliability. The new challenge is to retain the performance of previous message passing libraries while integrating software components for resiliency and reliability. LA-MPI is an effort to do just this. The present study examines the performance of LA-MPI on a typical scientific code, CICE [1]. The application code is more representative of the type of calculations and communications done by modern scientific applications. The LA-MPI's performance relative to a commercial offering and a widely used MPI compliant messaging library (MPICH) [2] are discussed in the next 3 sections.

*Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the National Nuclear Security Administration of the United States Department of Energy under contract W-7405-ENG-36. LA-UR-02-1004

message size in bytes	*number of messages*
< 100	21786
101-1000	18410
1001-10001	18410
10001-100000	18410
> 100000	26090

Table 1: Number of messages by size range in CICE for point-to-point communication. The distribution of message sizes is fairly even with slightly more very small and very large messages.

2 LA-MPI Overview

The Los Alamos Message Passing Interface (LA-MPI) is a open-source high performance MPI compliant message passing library. LA-MPI was specifically designed to overcome many of the shortcomings of current commercial and non-commercial message passing libraries including:

- lack of scalability
- lack of reliability or a mechanism to insure reliable message transport
- inability to customize or augment the messaging layers
- lack of support for multiple network protocols and interfaces

Currently, LA-MPI implements the full MPI 1.2 standard features reliably and is resilient in the face of network errors and faults. LAMPI can reliably deliver messages in the presence of errors due to the PCI bus, network card, and even wire transmission. LAMPI can survive network card and path failures and guarantees delivery of in-flight messages during failure. In addition the modular LAMPI networking architecture implements striping at both the message and message fragment levels by enabling the use of multiple (possibly heterogeneous) network devices concurrently.

3 CICE Overview

CICE [1] is a production code for efficiently modeling sea ice in a fully coupled atmosphere-ice-ocean-land global climate model. CICE is a widely used community model developed by scientists at LANL, the National Center for Atmospheric Research (NCAR), and other universities.

CICE is a good benchmark program for evaluating MPI implementations. It is written in Fortran 90 using a wide assortment of MPI features including point-to-point communication, broadcasts, reductions, MPI groups, and MPI datatype operations. In addition, CICE uses a fairly even distribution of message sizes with slightly more very small and very large messages (see Table 1). This removes strong biases toward particular message sizes when evaluating performance.

CICE is typically run on eight processors on the SGI Origin 2000 at LANL. Figures 1 through 4 shows the performance of CICE for 64 and 128 timesteps for MPICH, SGI MPT, and LA-MPI. Performance with LA-MPI is within 3% of SGI MPT for the 64 time step case and within less than 1% of SGI MPT for the 128 time step case.

The figures show that LA-MPI compares extremely favorably with the SGI MPI, and is much better than the MPICH.

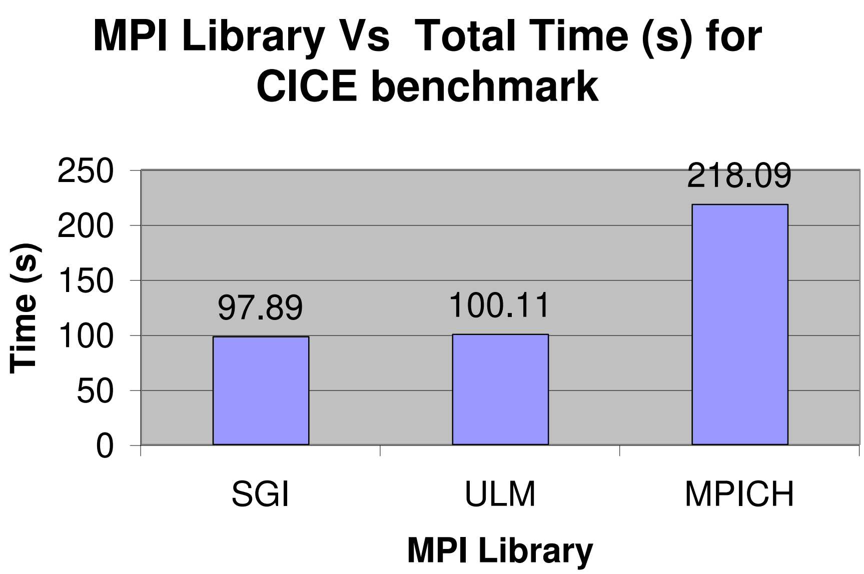

Figure 1: Total Time for CICE benchmark for 64 timesteps

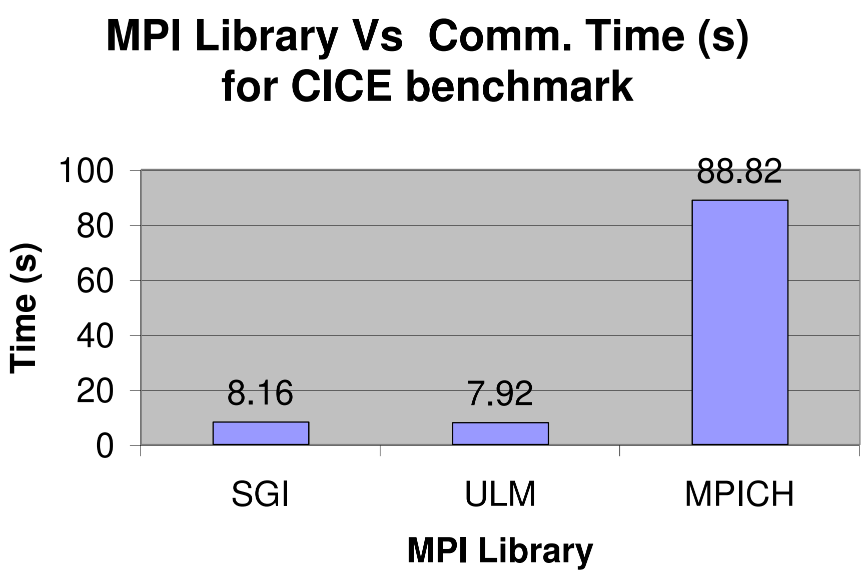

Figure 2: Communication Time for CICE benchmark for 64 timesteps

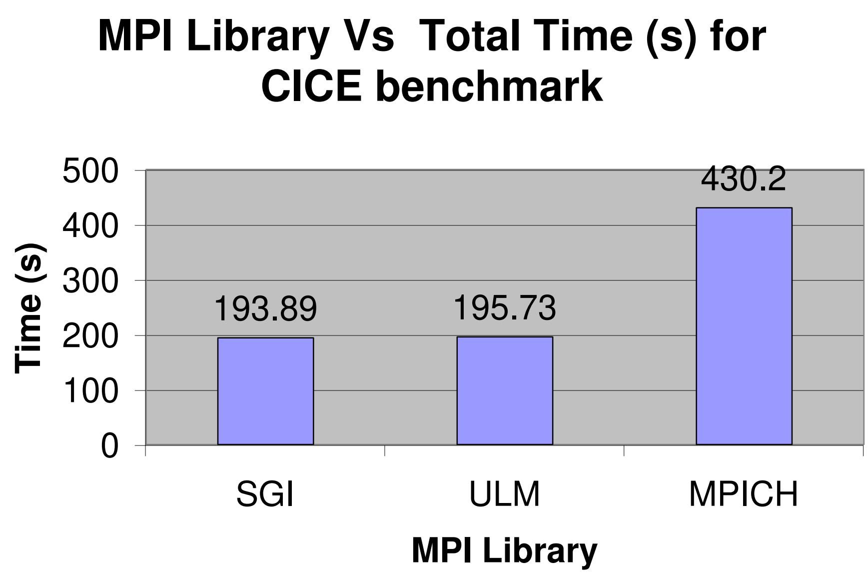

Figure 3: Total Time for CICE benchmark for 128 timesteps

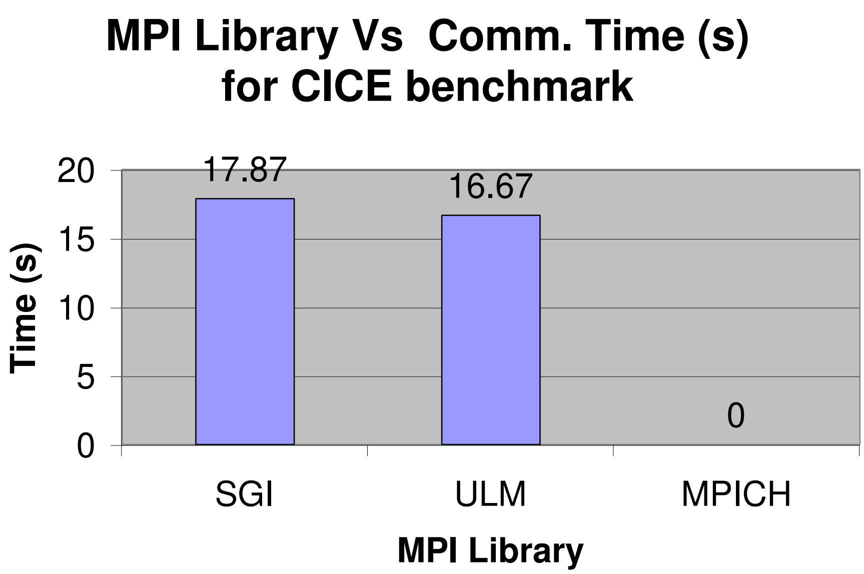

Figure 4: Communication Time for CICE benchmark for 128 timesteps

4 Appendix A. CICE Configuration

```
&ice_nml
    year          = 1997
  , istep0        = 0
  , dt            = 7200.0
  , ndte          = 120
  , npt           = 64
  , diagfreq      = 12
  , histfreq      = 'm'
  , dumpfreq      = 'm'
  , hist_avg      = .true.
  , restart       = .false.
  , print_points  = .true.
  , kitd          = 1
  , kdyn          = 1
  , kstrength     = 1
  , evp_damping   = .false.
  , advection     = 'remap'
  , grid_type     = 'displaced_pole'
  , grid_file     = 'global_192x128.grid'
  , kmt_file      = 'global_192x128.kmt'
  , dump_file     = 'iced'
  , restrt_pnter  = 'ice.restart_file'
  , history_file  = 'iceh'
  , diag_file     = 'ice_diag.d'
/

&icefields_nml
    f_hi        = .true.
  , f_hs        = .true.
  , f_Tsfc      = .true.
  , f_aice      = .true.
  , f_aice1     = .false.
  , f_aice2     = .false.
  , f_aice3     = .false.
  , f_aice4     = .false.
  , f_aice5     = .false.
  , f_uvel      = .true.
  , f_vvel      = .true.
  , f_Fswdn     = .true.
  , f_Flwdn     = .true.
  , f_snow      = .true.
  , f_rain      = .true.
  , f_sst       = .true.
  , f_sss       = .true.
  , f_uocn      = .true.
  , f_vocn      = .true.
```

```
  , f_Focnht    = .true.
  , f_Fswabs    = .true.
  , f_albsni    = .true.
  , f_Flat      = .true.
  , f_Fsens     = .true.
  , f_Flwup     = .true.
  , f_evap      = .true.
  , f_congel    = .true.
  , f_frazil    = .true.
  , f_snoice    = .true.
  , f_meltb     = .true.
  , f_meltt     = .true.
  , f_Ffrshw    = .true.
  , f_Fioht     = .true.
  , f_straix    = .true.
  , f_straiy    = .true.
  , f_strtltx   = .true.
  , f_strtlty   = .true.
  , f_strcorx   = .true.
  , f_strcory   = .true.
  , f_stroix    = .true.
  , f_stroiy    = .true.
  , f_strintx   = .true.
  , f_strinty   = .true.
  , f_strength  = .true.
  , f_divu      = .true.
  , f_shear     = .true.
  , f_sig1      = .true.
  , f_sig2      = .true.
  , f_dvidtt    = .true.
  , f_dvidtd    = .true.
  , f_daidtd    = .true.
  , f_salt      = .true.
/
```

References

[1] E. C. Hunke and W. H. Libscomb. Cice: the los alamos sea ice model, documentation and software. Technical Report LA-CC-98-16, 1999.

[2] W. Gropp, E. Lusk, G. for, m Implementation, o MPI, and M. Computer. Argonne national laboratory, 1996.