MPWide

A communication library for distributed supercomputing

Derek Groen[1,2], Steven Rieder[1,2], Paola Grosso[2], Simon Portegies Zwart[1] and Cees de Laat[2]

[1] Leiden Observatory, Leiden, the Netherlands
[2] University of Amsterdam, Amsterdam, the Netherlands

Application

We originally developed MPWide to manage the long-distance message passing in the CosmoGrid[a] project. This is a large-scale cosmological project whose primary goal is to perform a dark matter simulation using supercomputers on two continents.

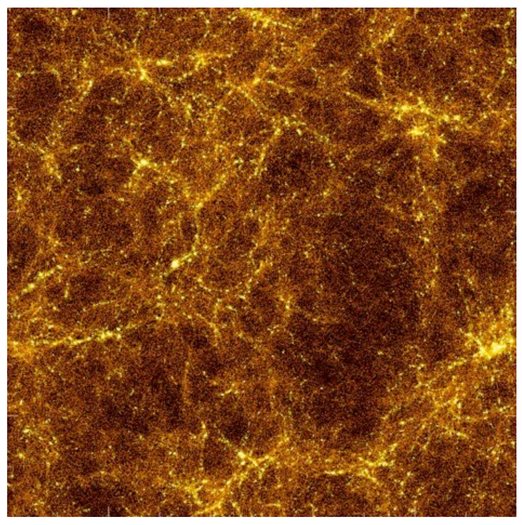
In this simulation, we use the cosmological Λ Cold Dark Matter model[b] to simulate the dark matter particles using a parallel tree/particle-mesh N-body integrator, TreePM[c]. This requires relatively little communication between different sites after each timestep. This integrator calculates the dynamical evolution of 2048^3 (8.5 billion) particles. More information about the parameters used and the scientific rationale can be found in [a]).

The integrator can be run as a single MPI application, or as two separately launched MPI applications on different supercomputers.

[a]) Portegies Zwart et al., 2009; IEEE Computer (submitted)

[b]) Guth, 1981; Physical Review D

[c]) Yoshikawa and Fukushige, 2005; PASJ

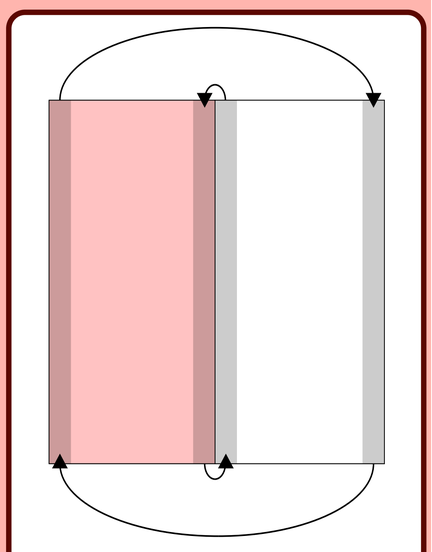


A snapshot of the CosmoGrid simulation. The (bright) dense areas form a cosmic web structure.

Motivation

We use MPWide to manage the wide area communications in the CosmoGrid project, where cosmological N-body simulations run on grids of supercomputers connected by high performance optical networks. To take full advantage of the network light paths in CosmoGrid, we need a message passing library that supports the ability to use customized communication settings (e.g. custom number of streams, window sizes) for individual network links among the sites. The supercomputers we use vary both in hardware architectures and software setup.

Many supercomputers have a recommended MPI implementation which has been optimized for the network architecture of that particular machine. Installing and optimizing a homogeneous MPI implementation on multiple supercomputer platforms is a task that may be politically difficult to initiate, and requires considerable effort and man hours to complete. This has led us to develop MPWide, a light-weight communication library which connects two applications, each of them running with the locally recommended MPI implementation.



After each computation step, the data in grey regions is transferred to the other supercomputer.

Tokyo Cray XT-4

Huygens Power6

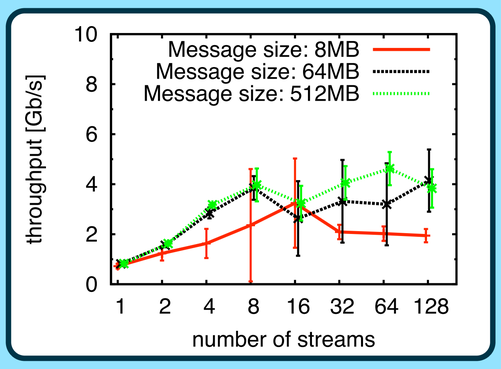
Benchmarks

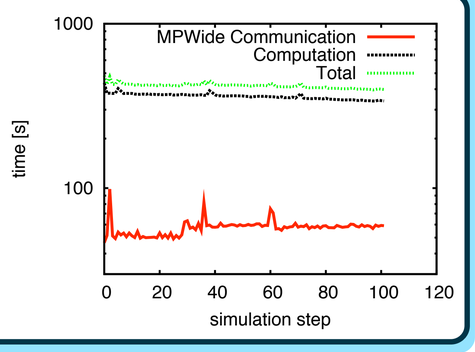
We measured the performance of MPWide between two nodes on different supercomputers, one located in The Netherlands, the other in Finland. These supercomputers are connected with a 10 Gbps interface. The round trip time for this network is 37.6 ms.

Each test consists of 100 two-way message exchanges, where we record the average throughput and the standard error. We performed the tests over a shared network with frequent background traffic.

Our tests show increased performance when using more streams, especially for larger message sizes.

We also tested MPWide in a production environment, during a CosmoGrid run. In this run, we used the Huygens supercomputer in Amsterdam and the Cray supercomputer in Tokyo. In this run, the calculation time dominated the overall performance, with the communication time constituting about one eighth of the total execution time.



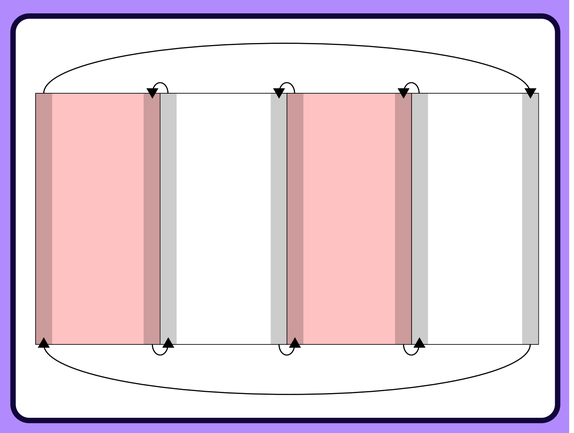


Related work and future

The MPI implementation most closely related to our work is the PACX-MPI[d] implementation. Like MPWide, this implementation connects different machines, while making use of the vendor MPI library on the system. The main difference between PACX-MPI and MPWide lies in the fact that MPWide supports a de-centralized startup, where PACX-MPI does not. For CosmoGrid, support for this is required, as it is not possible to start the simulation on all supercomputers from one site.

Other implementations of MPI, like Open MPI and MPICH-G2, differ further from MPWide, and do not support manual specification of the network topology, required by CosmoGrid.

In the near future, we will expand the CosmoGrid simulation to run on four supercomputer sites, and we will implement support for this in MPWide.



[d]) http://www.hlrs.de/organization/av/amt/research/pacx-mpi/