Parallel Computation of a Demagnetizing Field in a Distributed Environment

Y. Nakatani

Department of Computer Science, University of Electro-Communications
1-5-1, Chofugaoka, Chofu-shi, Tokyo 182

A parallel computation system for calculating a demagnetizing field in a distributed environment is presented. The system uses a hierarchically organized communication structure to reduce the time needed for communication between workstations. When 139 workstations were used, the total calculation speed was the same as that of a vector-type supercomputer with a computing speed of 1 GFLOPS.

Key words: micromagnetics, computer simulation, calculation of demagnetizing field, parallel and distributed calculation, workstation cluster

1. Introduction

Computer simulation based on micromagnetics has been popular as a result of the increasing demand for high-density recording systems. The simulations performed to date include the magnetization reversal mechanisms of fine particles, the magnetization distribution in magneto-optical-disk media, and the magnetization structure of vertical Bloch line and Bloch points. Because the calculation of a demagnetizing field requires a long computing time, these simulations have been done on supercomputers with vector processors or highly parallel computers such as the CM2.

On the other hand, there has been a remarkable advance in the computing power of small general-purpose computers, called workstations, and there are workstations whose computing power is equivalent to that of mainframe computers one generation ago. Widespread use of these workstations is progressing rapidly, owing to their small size and low price. They are usually connected by local area networks, and means for communicating between the processes on different workstations are provided for general users. It is expected that if a parallel and distributed calculation system for executing micromagnetic calculations is provided on these workstations, such problems can be solved without relying on a supercomputer.

This paper proposes a distributed calculation system for the most time-consuming portion of the micromagnetic calculation, namely, calculation of the demagnetizing field, and describes an implementation of the system.

2. Parallel Calculation of a Demagnetizing Field

We use the conventional method for calculating a demagnetizing field. In this section, we first briefly review the conventional method, and then present a parallel calculation model for obtaining the demagnetizing field by this method.

2.1 Calculation of a demagnetizing field by the conventional method

The whole computing region is divided into a number of cells in which the magnetization is assumed to be uniform. The demagnetizing field is obtained by superposing the dipole field generated by the magnetization \( \mathbf{M} \) within each source cell at the center \( P(x, y, z) \) of an observing cell:

\[
\mathbf{H}^0(P) = -\frac{1}{\text{cell}} \sum_{Q} \frac{1}{r^3} \left[ \mathbf{M}(Q') \right. \\
-3(\mathbf{M}(Q') \cdot r) \frac{r}{r^2} \left. dx'dy'dz' \right],
\]

(1)

where the integrations are carried out within the volume \( V' \) of a source cell. \( r \) denotes the distance vector drawn from a source point \( Q'(x', y', z') \) within the cell \( Q \) to the observing point \( P \). The magnitude of \( r \) is denoted by \( r \). Since the magnetization \( \mathbf{M}(Q') \) is assumed to be a constant vector within each source cell, eq. (1) is reduced to a matrix equation of the following form:

\[
\mathbf{H}^0(P) = \sum_{Q} \mathbf{x}(P-Q) \cdot \mathbf{M}(Q),
\]

(2)

The elements of matrix \( \mathbf{x} \) (hereafter called demagnetizing coefficients) can be obtained analytically when the shape of the calculation cell is a rectangular prism.

2.2 Strategy

The process for calculating the demagnetizing field according to eq. (2) can be described as follows:

for obs = 1 to \( n \) do
    for so = 1 to \( n \) do
        \( \mathbf{H}(\text{obs}) = \mathbf{x}'(\text{obs, so}) \cdot \mathbf{M}(\text{so}) \),

where \( n \) is the total number of calculation cells and \( \mathbf{x}' \) is the coefficient matrix for the demagnetizing field. While the observing cells are swept in the outer loop, the inner loop sweeps the source cells. This routine shows that the demagnetizing fields at different observing cells can be calculated independently. We divide the outer loop calculation into a certain number of sub-
loop calculations, which are executed concurrently by multiple processes that reside on different workstations connected by a local area network.

2.3 Data communication in a distributed environment

It is worthwhile examining the method for transmitting message data between processes on different workstations, since the time required for the inter-processor communication significantly influences the performance of a parallel computation.\(^\text{9,10}\)

We assume that our distributed environment is composed of UNIX*1 workstations and local area networks using an Ethernet or FDDI. A group of workstations are connected through a common network, hereafter called a subnet. A subnet is connected with another subnet at a workstation that is usually called a gateway. In terms of the connection of workstations, a distributed environment can be regarded as a parallel computer with distributed memory whose processors are connected by hierarchically connected busses.

Two methods for inter-processor communication are available at present: the transmission control protocol (TCP) and the user datagram protocol (UDP).\(^\text{11}\) TCP allows the execution of a single reliable data communication, while UDP can handle multiple but unreliable data communications. Since we need reliable data communication, we use TCP. Thus, from the viewpoint of data communication, a distributed environment can be regarded as a parallel computer whose processors are connected with each other by hierarchically connected star networks with distributed memory.

A communication message includes the header information that is added to the main body of the data to be communicated. The proportion of the header information to the communication data increases with a decrease in the data length. Data collisions will occur as the number of communications on an Ethernet increases, and the speed of communication will decrease as a result.\(^\text{13}\) If we consider the characteristics of the communication on an Ethernet, it is preferable to send messages consisting of long packages occasionally rather than short messages frequently, so that we can maintain the communication rate. In the next subsection, we will construct a parallel calculation model that takes account of these circumstances.

2.4 Parallel calculation model

In our calculation model, we parallelize only the part of the calculation for obtaining the demagnetizing field. All the other parts of the calculation are executed in one process, which we will call the parent. A process for calculating a demagnetizing field is called a child.

---

*1 UNIX is a registered trade mark of AT&T Bell Laboratories in the United States and other countries.
Fig. 3 Enhanced parallel calculation model.

Fig. 4 Communication and calculation scheme (enhanced model).

Fig. 5 Speedup achieved with the proposed model (estimated).

The calculation of the demagnetizing field but only mediate the data communications. While the children are busy calculating the demagnetizing field, those workstations that have only the parent or a bridge would not have jobs to do, and would stand idle. To avoid this situation, when we create bridge processes, we can also create a child on a workstation that has a parent and/or a bridge. Thus this model does not actually need extra workstations.

The total calculation time and the speedup of the model can be written as follows:

$$t_{cal} = t_{send} + t_{dem} + t_{cal}.$$  

$$t_{send} = \frac{(n \cdot 8 \cdot 2 + 4 \cdot 2) \cdot b}{rate'}, \quad \frac{(n \cdot 8 \cdot 2 + 4 \cdot 2) \cdot c}{b \cdot rate},$$

$$t_{dem} = \frac{n \cdot n \cdot dem}{c},$$

$$t_{cal} = \frac{n \cdot 8 \cdot 3}{c \cdot rate'} + \frac{n \cdot 8 \cdot 3}{b \cdot rate'},$$

$$SP = \frac{n \cdot n \cdot dem}{t_{cal}}.$$  

where $c$ is the number of children, $b$ is the number of bridges, $rate'$ is the communication rate between a parent and a bridge, $rate$ is the communication rate between a bridge and a child, and $dem$ is the time necessary for the calculation of the demagnetizing field per unit cell.

$t_{send}$ is the communication time necessary for sending the magnetization distribution and the addresses of the calculation regions from the parent to the bridges and from the bridges to the children. The magnetization is given by two double-precision (8-byte) floating-point numbers that representing the polar and azimuthal angles, because the magnitude of the magnetization does not change throughout the calculation. The addresses of the calculation regions are represented by two integers (4-byte) of the long-word type. $t_{dem}$ is the time taken to calculate the demagnetizing field. $t_{cal}$ is the communication time needed to collect the calculat-
ed demagnetizing fields from the children to the bridges and from the bridges to the parent. (In this model, the bridge sends the collected data to the parent after receiving data from all of the children. If it sends the collected data to the parent each time it receives data from the child, \( t_{\text{data}} \) seems to be reduced. However, this increases the frequency of data transfer, and the communication rate will decrease owing to data collisions.) The first term of \( t_{\text{end}} \) is proportional to \( b \) and the second term is proportional to \( c/b \). The communication time, \( t_{\text{end}} \), decreases in proportion to \( 1/b \) when \( b \) is small, but increases when \( b \) becomes large. Usually a constant term that represents the overhead of the communication is included in \( t_{\text{end}} \) and \( t_{\text{coll}} \). But this term is dropped for the sake of simplicity.

Figure 5 shows the estimated dependency of the speedup on the number of children and the total number of calculation points. The total number of the calculation points is varied from 1,000 to 300,000, the number of the children is varied from 1 to 1,000, and \( b \) is fixed at ten. The experimental values used for \( \text{rate} \) and \( \text{dem} \) were obtained by using an Ethernet and a Sony NWS3470 workstation; \( \text{rate} \) and \( \text{dem} \) are 1 Mbytes/s and \( 1.5736 \times 10^{-4} \) s, respectively, and the ratio of \( \text{rate} \) to \( \text{rate} \) is assumed to be unity.

The calculation speed increases with an increase in \( c \) when \( c \) is less than 100, but it reaches a maximum at a certain value of \( c \), depending on \( n \), and then decreases. The model has the feature that the speedup eventually decreases when \( c \) is increased while \( n \) is kept constant. This situation occurs in cases where there are many workstations in one subnetwork, which is generally an undesirable way of using a distributed environment. Usually there are ten workstations in one subnetwork, and such subnetworks are connected in a certain hierarchical structure. The number of workstations in a subnetwork can of course be reduced by increasing the depth of the hierarchy of communication, that is, by adding second-level bridges, each of which mediates between its children and its first level-bridges, each of which in turn mediate between the second-level bridge and the parent. We planned to implement the communication model so that it can be fitted into the structure of our network, so, we did not extend the level of the communication hierarchy further.

Details of the effect of the number of processors and the data communication on the performance are discussed by Hwang\(^{[10]} \) and Zhu\(^{[10]} \).

3. Results

We used one hundred and thirty-one workstations in the computer room of the author's department, and eight workstations in our laboratory. In the computer room, eight high-performance workstations (NWS5000 VI or NWS5900) are connected by an FDDI network. One of them is the gateway of the room, and the other seven workstations are the gateways of the subnetworks in the room. Six of them have two subnetworks, each composed of ten or nine medium-preformance workstations (NWS3470) connected to one Ethernet, and one workstation has two subnetworks, each composed of five or four medium-performance workstations.

Two experiments were performed: one to obtain the speedup curves for various numbers of workstations, and the other to obtain the peak speed using all the (one hundred and thirty-nine) workstations that were available. This speed was directly compared with the calculation speed of a supercomputer. These experimental data were obtained by simulating the annihilation process of a vertical Bloch line pair,\(^{[13]} \) without other users.

The socket system call is used for data communications, and the rexec system call is used for spawning processes on the bridges and children from the parent or bridges.

Table 1 Calculation times on supercomputers, workstations, and using the proposed parallel calculation method.

<table>
<thead>
<tr>
<th>Machine</th>
<th>FLOPS</th>
<th>CPU Time</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>S800/480</td>
<td>8 G</td>
<td>15,8706</td>
<td>1.00</td>
</tr>
<tr>
<td>S800/60</td>
<td>1.5 G</td>
<td>90,1224</td>
<td>6.10</td>
</tr>
<tr>
<td>S810/20</td>
<td>630 M</td>
<td>167,1816</td>
<td>11.31</td>
</tr>
<tr>
<td>NWS5000VI</td>
<td></td>
<td>3207,6599</td>
<td>217.04</td>
</tr>
<tr>
<td>NWS5000SA</td>
<td></td>
<td>4318,1514</td>
<td>272.08</td>
</tr>
<tr>
<td>RS3330</td>
<td></td>
<td>6719,5830</td>
<td>454.68</td>
</tr>
<tr>
<td>NWS3470</td>
<td>2.3 M</td>
<td>12262,9971</td>
<td>772.69</td>
</tr>
<tr>
<td>Parallel(^{1})</td>
<td></td>
<td>108,0000</td>
<td>6.81</td>
</tr>
</tbody>
</table>

\(^{1}\text{Parallel: NWS5000VI} \times 7, \text{NWS5000SA} \times 1, \text{RS3330} \times 1, \text{NWS3470} \times 130\)

As mentioned before, this model only parallelizes the routine for calculating the demagnetizing field; the rest of the simulation is carried out by the parent. We have only discussed the calculation time used in the demagnetizing field calculation.

3.1 Speedup curve

We used six clusters of workstations consisting of the same number of workstations; that is, we used six bridges. To obtain the speedup curve, the child processes were run on an NWS3470, and the parent and bridges were run on an NWS5000VI.

Figure 6 shows the dependence of the speedup on the number of the children, with the total number of calculation points \(n\) varied from 1,000 to 30,000. Here we made twenty calculation steps and summarized the minimum calculation time. When the number of children is small, the speedup is in all cases proportional to the number of children. As \(c\) increases, the speedup increases and finally becomes saturated, when \(n\) is small. But the speedup does not decrease when \(n\) is large. Of course, it will become saturated and eventually decrease with a further increase of \(c\). The speedup does not increase smoothly when \(n\) is small, where the calculation time is very small and hence the accuracy of the timer routine is not very precise.

Next we will compare the calculation time obtained experimentally with that estimated from eq. (3). Because the workstations that have the processes assigned to bridges and parent are connected with the FDDI, the communication speed \(rate'\) is different from \(rate\). The experimental value 4.5 is used for the ratio of \(rate'\) to \(rate\).

The differences between the calculation time obtained experimentally and the time obtained from the estimation equation are shown in Fig. 7. The error reaches about 100% when \(n\) is small, but becomes smaller as \(n\) is increased, eventually falling to several percent. We dropped the term representing the overhead of communication in the estimation of the communication time; this is why the error becomes large when \(n\) is small.

3.2 Peak speed

We used all of the one hundred and thirty-nine workstations that were available to the author in order to achieve the maximum calculation speed. The bridge processes were run on the seven previously described gateway machines, and the parent process was also run on one of them. The gateway machine in the author's laboratory is connected with the gateway machine in the Department's computer room via the university's local area network. One of the bridges was run on the gateway machine of the author's laboratory. In the experiment described in the preceding subsection, the child processes were run only on an NWS3470. In the present experiment, however, the child processes were made to run on all workstations. Because of the difference in the computing speed among the child processes, it was necessary to use a load-balancing technique so that we could make good use of all the workstations.\(^{2}\)

We carried out twenty steps of calculation, with the shortest calculation time regarded as the calculation time obtained by using parallel calculation.

Table 1 summarizes the computing speeds of supercomputers, workstations, and the present parallel calculation. The calculation time was measured in the simulation of the annihilation process of a vertical Bloch line pair with the calculation points of \(n = 32,000.\) These data were obtained by using programs tuned to the respective target computers. The top three results were obtained on Hitachi supercomputers. The 3800/480 is a parallel supercomputer with four processors, but the result shown in the table was obtained with one processor. In these three cases, the VPU : CPU ratios were larger than 99%. This shows that we can obtain excellent performance by using supercomputers. The parallel calculation was performed by using 139 workstations: seven NWS5000VI, one NWS5000 SA, one RS3330, and 130 NWS3470. The calculation time for each workstation is also shown in the table. The calculation time for the parallel calculation is about 6.81 times longer than that of the supercomputer S3800/480 with a peak computing speed of 8 GFLOPS. This shows that the calculation speed of the present parallel calculation with 139 workstations is the same as that of a supercomputer with a peak computing speed higher than 1 GFLOPS.

4. Conclusion

A parallel calculation system for calculating a demagnetizing field in a distributed environment was presented. The system was based on inter-workstation communication organized hierarchically to reduce the communication time. When 139 workstations were
used, the total calculation speed was the same as that of a vector-type supercomputer with a computing speed of 1 GFLOPS.

TCP was used in this paper as a means of interprocess communication, because of the limitations of the other available protocols. The overall calculation speed was thus limited by the inter-process communications. Recently, reliable methods of multicasting have been investigated. If a standard reliable multicasting method becomes available, it should diminish the amount of communication drastically, and we can expect to obtain a further speedup by using parallel computation in a distributed environment.

The system we described was constructed in a distributed environment in which workstations were connected by local area networks. If we extend the system in an environment where workstations and supercomputers are connected by the Internet, the calculation system will have a computing speed in excess of 1 TFLOPS.

Acknowledgment I would like to thank Prof. Nobuo Hayashi of the University of Electro-Communications for several enlightening discussions.

References


Received July 14, 1995; Accepted Jan. 12, 1996