Implementation and Evaluation of Shared Memory System for Establishing λ Computing Environment

Eiji Taniguchi[1], Ken-ichi Baba[2], Masayuki Murata[1]

Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan, Tel: +81-6-6879-4542,

Fax: +81-6-6879-4544, Email: [1]{e-tanigu, murata}@ist.osaka-u.ac.jp, [2]baba@cmc.osaka-u.ac.jp

Abstract

In λ computing environment, wavelength paths are utilized as shared memory for high-speed and high-quality data exchanging and sharing for distributed computing. Our implementation and evaluation of such a shared memory access system are presented.

1 Introduction

Since in standard Grid computing, each node executes parallel applications by exchanging data through TCP/IP protocol stack, it is difficult to achieve good performance in sharing and exchanging a large amount of data because of its overhead such as packet processing delays and packet losses.

We are proposing a new computing architecture, called the λ computing environment [1], that provides virtual channels utilizing optical wavelength paths directly connecting computing nodes. See Fig. 1.

By directly connecting nodes through wavelengths, we can expect to achieve high-speed and high-quality communication for distributed computing environment. The virtual channels may build an environment with a mesh topology, or a virtual ring topology to make it easier to share data among computing nodes. In [1], we have considered the shared memory architecture by utilizing multiplexed wavelengths where the wavelength channel itself is considered as a shared memory. On the other hand, in this paper, we realize a more realistic shared memory system on the photonic ring network. For this purpose, we utilize AWG–STAR system [2] developed by NTT Photonics Laboratory (see Section 2) in order to implement a shared memory system for distributed computing. In this paper, we investigate and discuss how to improve the performance.

2 Brief overview of AWG–STAR system and our contribution

An AWG–STAR system is information sharing network platform realized by the WDM (Wavelength Division Multiplexing) technology and wavelength routing using AWG (Array Waveguide Grating) routers. The AWG router processes optical signal without transforming into electrical signal. Computing nodes connected to the AWG router configure a star topology in physical, but does a ring in logical (Fig. 2). The optical bandwidth of wavelength channel is 2.152Gbps. Each node is equipped with a shared memory board (SMB). It provides shared memory that can contain the identical data at the same address over all nodes by AWG–STAR [2]. However, AWG-STAR does not provide co-operating functions among processes running in parallel, while those are mandatory for realizing distributed computing. One such example is a barrier function that makes some process wait until all processes call this function. We realize those functions and make it possible to execute the distributed program on AWG-STAR.

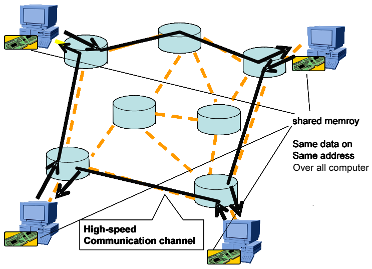

Fig. 1: A conceptual sketch of λ computing environment.

Another problem of the AWG–STAR is that SMB is equipped with computing nodes via a PCI bus. Therefore, an access delay to the shared memory is slower than that of local memory and it may largely affect the entire performance (currently read/write access speed is around 60MB/s). Also, it takes 500ns to delete and/or append transmission frames and to reflect update information to the shared memory. By considering it, we realized a function which allocates the memory for shared variables effectively.

In the later section, we will address how much those factors affect the execution performance of distributed programs and investigate how those can be eliminated.

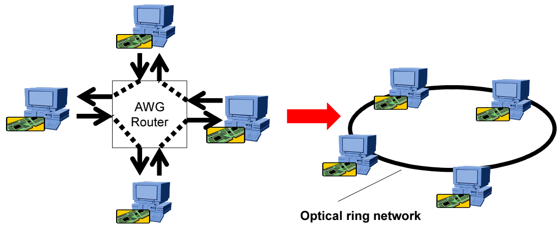

Fig.2: Topology of AWG–STAR system.

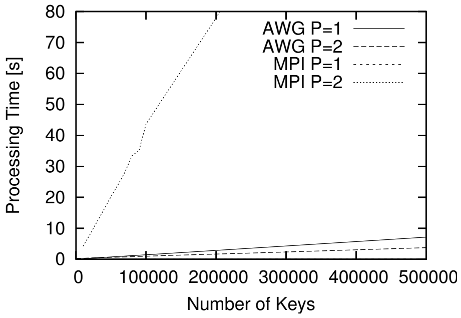

Fig.3: Execution times of radix sort program.

3 Implementation and evaluation

We implemented our shared memory system on the AWG–STAR and evaluated it by porting and executing several typical parallel programs in SPLASH2 [3].

Figure 3 shows execution times for a radix sort program. P is the number of nodes on the ring. We run one process on each node in the experiment. For comparison purpose, we also show the cases for MPI over TCP/IP running on 100Mbps Ethernet. In this case, the performance of our system can get good performance as expected.

However, as shown in Fig. 4, execution time for the LU decomposition program becomes much larger even than that of the MPI case on 100Mbps Ethernet. The reason is due to the delay time of the shared memory because the LU decomposition program calls many write accesses to the shared memory in our first implementation. Note that in the case of the radix sort program, the number of accesses is relatively small.

One solution against the problem is hardware improvement of SMB, which is now undergoing, but meanwhile we have tuned the program by decreasing the number of access times to the shared memory as follows. (1) The program accesses to the shared memory in a block unit not in a matrix element unit, while the latter is a common way. (2) It utilizes the local memory as cache for shared memory and writes back to the shared memory when necessary. Figure 5 shows the effect of tuning in the LU decomposition program for the first and second improvements described above. As shown in the figure, we have confirmed that program tuning is quite important to get performance improvement.

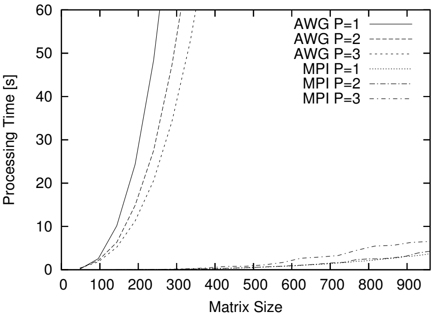

Fig.4: Execution times of LU decomposition program.

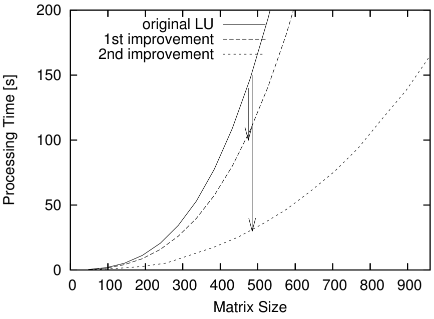

Fig. 5: Improvements of LU decomposition program ($P = 3$).

4 Conclusion

In this paper, we have implemented and evaluated the shared memory system on the photonic ring (the AWG–STAR system) for establishing distributed λ computing environment. We have confirmed that the WDM technology enables the high-speed computing environment, but also found that the memory system at local hosts easily becomes a bottleneck. Program tuning is an important approach to improving the performance as we have demonstrated in this paper, but also we need to investigate a end-host architecture suitable to the photonic-based computing environment, which is our future research topic.

Acknowledement

Last, we thank Dr. Okada of NTT for his helpful suggestions and supports.

References

[1] H. Nakamoto, K. Baba and M. Murata, “Shared memory access method for a λ computing environment,” in *Proc. of IFIP OpNeTech*, Oct. 2004.

[2] A. Okada, H. Tanobe and M. Matsuoka, “Dynamically reconfigurable real-time information-sharing network system based on a cyclic-frequency AWG and tunable-wavelength lasers,” in *Proc. of ECOC2003*, Sep. 2003.

[3] `http://www-flash.stanford.edu/apps/SPLASH/`.