Distributed Computing over Optical Networks

Wei Guo, Yaohui Jin, Weiqiang Sun, Weisheng Hu
State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University, Shanghai, 200240, China, Email: vguo@sjtu.edu.cn
Xinhua Lin, Min-You Wu
Grid Computing Center, Shanghai Jiao Tong University, Shanghai, 200240, China
Hong Liu, San Fu
Institute of Aerospace Science and Technology, Shanghai Jiao Tong University, Shanghai, 200240, China
Jun Yuan
Shanghai Supercomputer Center, Zhangjiang Hi-tech Park, Shanghai, 201203, China

Abstract: This paper overviews the opportunities and challenges of distributed computing over optical network. The Terabits Optical Network Integrated Computing Environment (TONICE) project and the Integrated Resources Management System (IRMS) are presented.

OCIS codes: (060.4250) Networks; (060.4510) Optical communications

1. Introduction

Distributed computing is developed to enable an efficient utilization of geographical distributed and heterogeneous computational resources, such as supercomputer, cluster, storage and display devices, to complete the large-scale scientific computing applications. These applications require multi-gigabits flow with low loss, low latency, and minimal jitter network connection for the large data transfers associated with executions on remote computational resources. The TCP/IP network with “best effort” service can not satisfied these requirements. Consequently, an infrastructure that employs optical network to interconnect computational resources with deterministic, high-bandwidth network connections is being investigated. This infrastructure performs well for large data transfers in distributed computing. However, as a promising and new type of technology, the infrastructure presents new challenges:

- In the infrastructure, the computational resources and optical network resources are time-shared by all of the users. The applications may dynamically request any type of resource - not only high-performance computers, but also network connections. It is necessary to allocate the computational resources and network resources dynamically to optimize the costs.

- In general, a distributed computing application consists of a set of tasks with dependent relationship. Each task requires different types of resources for execution, e.g. CPU for computation and network connection for data communication. We can choose different available computational resources to run the computation and choose different available network connection to transport the data. However, the different task scheduling approach may lead to different execution time, and this is crucial for real-time distributed computing applications. So we need to design an optimal task scheduling algorithm to provide better performances for the applications.

- In distributed computing application, sources and destinations of the network connections are dependent upon where the computational tasks are assigned. They are unknown initially. The different computational resources allocations may require different network connections to be established. In the mean time the different optical network connections also affects the duration of computational resources. Thus, we need to allocation computational and network resources jointly to achieve higher resources utilization ratio.

In the past few years, several active researches and on-going projects are investigated the infrastructure to support distributed computing over optical network, such as Enlightened and G-Lambda, PHOSPHORUS, DRAGON, OptIPuter, UCLP and VIOLA [1]. These researches assumed that the computational tasks are independent. Most of them manage the computational resources and optical network resources in different layer. In these efforts, the computational tasks are scheduled only by the computational resources manager. The computational resources manager invokes the optical network control plane to setup a connection when data communication is needed. We call this method as overlay resources management system (ORMS). A few of them has focused on coordinated reservations of both network and computational resources. However, how to design an integrated resources management system that can simultaneously allocate both computational resources and optical network resources and schedule the tasks to execute within a distributed computing environment remains to be investigated.

In China, the Terabits Optical Network Integrated Computing Environment (TONICE) project was initiated to undertake the challenges. It focuses on developing a framework to allow distributed computing applications over
optical network in a coordinated manner. The objective is to develop an integrated resources management system (IRMS) that can provide better performances for applications and achieve higher resources utilization ratio. In this paper, we will present the framework of TONICE first. Then, we will describe the model and implementation of the IRMS. We will also give a living demonstration to prove the effectiveness of this system.

2. TONICE Framework

TONICE is a large-scale interdisciplinary project to address a framework of distributed computing over optical network. The project focuses on coordinated management of both computational resources and optical network resources. It also focuses on designing and developing a national experimental testbed to demonstrate the effectiveness of this architecture.

TONICE is developed based on the China high-performance broadband information network (3TNET) testbed supported by the national high technology research and development program [2]. An experimental testbed for TONICE is deployed in Yangtze River Delta region centered at Shanghai city, including Nanjing city of Jiangsu Province, and Hangzhou city of Zhejiang Province in China. Fig. 1 shows the framework of TONICE. The core is an optical backbone equipped with Automatic Switching Optical Network (ASON) and Dense Wavelength Division Multiplexing (DWDM) nodes. A DWDM long haul with $4 \times 40$ GB/s bandwidth connects Shanghai and Hangzhou Cities. There are totally 13 ASON nodes. Each node has 360 Gb/s (expandable to 1.2Tb/s) switching capability with VC-4 granularity and support GE interface by Next Generation Synchronous Digital Hierarchy (NG-SDH). An extended Generalized Multiprotocol Label Switching Protocols (GMPLS) control plane is developed to control the network.

Four high performance computers including one supercomputer and three clusters are interconnected by the optical network. Currently, the total computational nodes number is 88. The maximum computational capacity is about 0.5TFlops and the total storage is about 1.025Tbits. The DAWNING 4000A supercomputer, which is the fastest supercomputer open to public in China, will be connected to the testbed in the near future. This supercomputer possesses 10TFlops maximum computational capacity and is operated by Shanghai Supercomputer Center (SSC).

3. System Model and Implementation

A distributed computing application is modelling by means of Directed Acyclic Graph (DAG) in this paper. Each node in the DAG represents an executable task. Each directed edge represents data communication between two tasks. The resources coordinated allocation and task scheduling can be defined as finding an optimal approach to map each node of DAG to an appropriate computational resource and each edge of DAG to an appropriate optical network connection. The objective is to minimize the execution time of the application (we call this time as scheduling length or makespan) and maximize the resources utilization ratio. We have developed several task scheduling algorithms to achieve optimal task scheduling result [3, 4, 5].

![Fig. 1 The framework of TONICE](image-url)
The architecture of IRMS shows in Fig. 2. Three functional modules: Resources Manager, Task Scheduler and Execution Manager are designed to achieve resources discovering, resources monitoring, resources allocating and task execution. Resources Manager acquires and storages both the computational resources and optical network resources information, such as resource’s name, address, type, available computational and storage capability, network topology, available bandwidth, etc. Task Scheduler runs the specified task scheduling algorithm to get an optimized task scheduling result according the application requirements and the available resources information. Execution Manager conducts the computational resources to execute computational tasks and signals optical network to setup or release the connections according to scheduling result on time.

4. Demonstration

We deployed a distributed computing platform in our testbed to support the distributed computing application in aircraft shape design [6]. Generally, aircraft shape design is a complex system. An efficient aircraft shape design involves multiple disciplines and is network centric. To complete the design process efficiently, all specialists in various disciplines at different places must be jointed in a collaborative optimization process and the interaction between all designer and design process must be fast enough. The demonstration results show that the IRMS can make the design process be easily controlled and monitored and can conduct interaction quickly. As a result, the application execution time can be shortened and the resources utilization ratio can be improved efficiently.

5. Conclusion

To support the distributed computing application over optical network, the TONICE project is proposed and the IRMS is developed to manage computational resources and optical network resources jointly. We successfully deploy the testbed and demonstrate the applications on the testbed. The experimental results show that the IRMS is feasible. It can make the application process be easily controlled and monitored and can shorten the applications’ execution time and improve the resources utilization ratio efficiently.

Acknowledgements

This work was supported jointly by NSFC (No. 60672016), 863 program of China and Shanghai Sci&Tech committee.

References