Supercomputer System for Numerical Weather Prediction by Taiwan Central Weather Bureau

● Fumihiro Takehara ● Hidenori Hayashi ● Junichi Fujita

The Taiwan Central Weather Bureau (CWB) is responsible for forecasting and monitoring weather across Taiwan and for detecting earthquakes and tsunamis. Its numerical weather prediction system is being upgraded in three phases to use Fujitsu's new supercomputer system centered about FUJITSU Supercomputer PRIMEHPC FX10. The sale of the FX10 supercomputer in the first phase to the CWB was the first overseas shipment of this system, which was developed on the basis of the technology used in the K computer.

1. Introduction

The Central Weather Bureau (CWB) of Taiwan is a government agency that forecasts and monitors weather throughout the country and detects earthquakes and tsunamis. In 1983, it began supercomputer operations with the CDC Cyber 205 as a numerical weather prediction (NWP) system based on an overall computerization plan for meteorological services. The NWP system was subsequently upgraded to higher speed supercomputers with the second-generation CRAY YMP-81 and YMP-2E in 1990, the third-generation FUJITSU VPP300E and VPP5000 in 1999, and the fourth-generation IBM P5-575 Cluster 1600 in 2006.

The impact of the recent dramatic changes in worldwide climate and of weather-related disasters on socioeconomic activities has been increasing, and weather-related disasters associated with heavy rainfall caused by yearly typhoons have become a frequent occurrence in Taiwan due to its subtropical climate. There is therefore a need in Taiwan to improve the accuracy of weather observations and forecasts and to strengthen the observation and analysis of long-term climate change.

To deal with the resulting increase in computational demand on the NWP system, the CWB introduced a supercomputer system centered about Fujitsu's PRIMEHPC FX10 supercomputer (four racks) in December 2012 as an upgrade to the IBM P5-575 Cluster 1600 supercomputer. Deployment of this new system is being done in three stages. The first stage in 2012 commenced operations, and the second stage in 2013 expanded the system (four more racks). The third stage in 2014 is slated to further enhance the system. This system was developed on the basis of the technology used in the K computer.

In this paper, we overview the NWP work and the NWP system of the CWB and describe the basic configuration and features of the first-stage PRIMEHPC FX10 system, the computational platform for the NWP system.

2. Overview of NWP work

In principle, the NWP work process of the CWB is divided into four work periods corresponding to worldwide weather observation periods starting at 00:00, 06:00, 12:00, and 18:00 (GMT). In each work period, observation data must be obtained from the upstream system before beginning forecast work. Furthermore, to enable an adequate amount of observation results to be collected, each "major run" of forecast work generally begins three hours after observation time commences, which means GMT + three hours (03:00, 09:00, 15:00, and 21:00), or, in Taiwan time, those start times + eight hours (11:00, 17:00, 23:00, and 05:00). In addition, the latest forecast start time in each period requires that a complete set of information be provided to the forecast center no later than a certain time, and those times in Taiwan time are 14:00, 20:00, 02:00,

and 08:00.

All forecast work begins with analysis and processing by the Global Forecast System (GFS) followed by forecasting with regional forecast models once the data required for those models have been obtained. At the start time of a major run, the observation data may not be fully collected, so many models execute a post run six hours later before executing the major run in the next work period. The forecast results obtained by model calculations performed in the post run using complete observation data can then be compared with those obtained in the major run, and the forecast with the best results can be provided as weather information for the next forecast round.

The NWP system has a close relationship with many other systems. A diagram of the main upstream and downstream flows and peripheral systems for the NWP system is shown in **Figure 1**. The NWP system obtains a variety of observation data and overseas model data from a global data information-collection system, an automatic weather data processing system, and other related upstream systems and services. Then, after outputting numerical predictions, it provides data to a mapping system, grid-data transmission system, real-time forecasting system, and other downstream systems and services for subsequent work. Finally, after completing forecast work, the NWP system stores the output results of each model and work-related data in a large-scale data storage system.

3. NWP system model

Numerical weather predictions are made on the basis of calculations performed on an active system and a backup system consisting of two PRIMEHPC FX10 clusters. The active system provides information to related departments both inside and outside the CWB while the backup system provides an environment in which CWB NWP-system researchers and developers

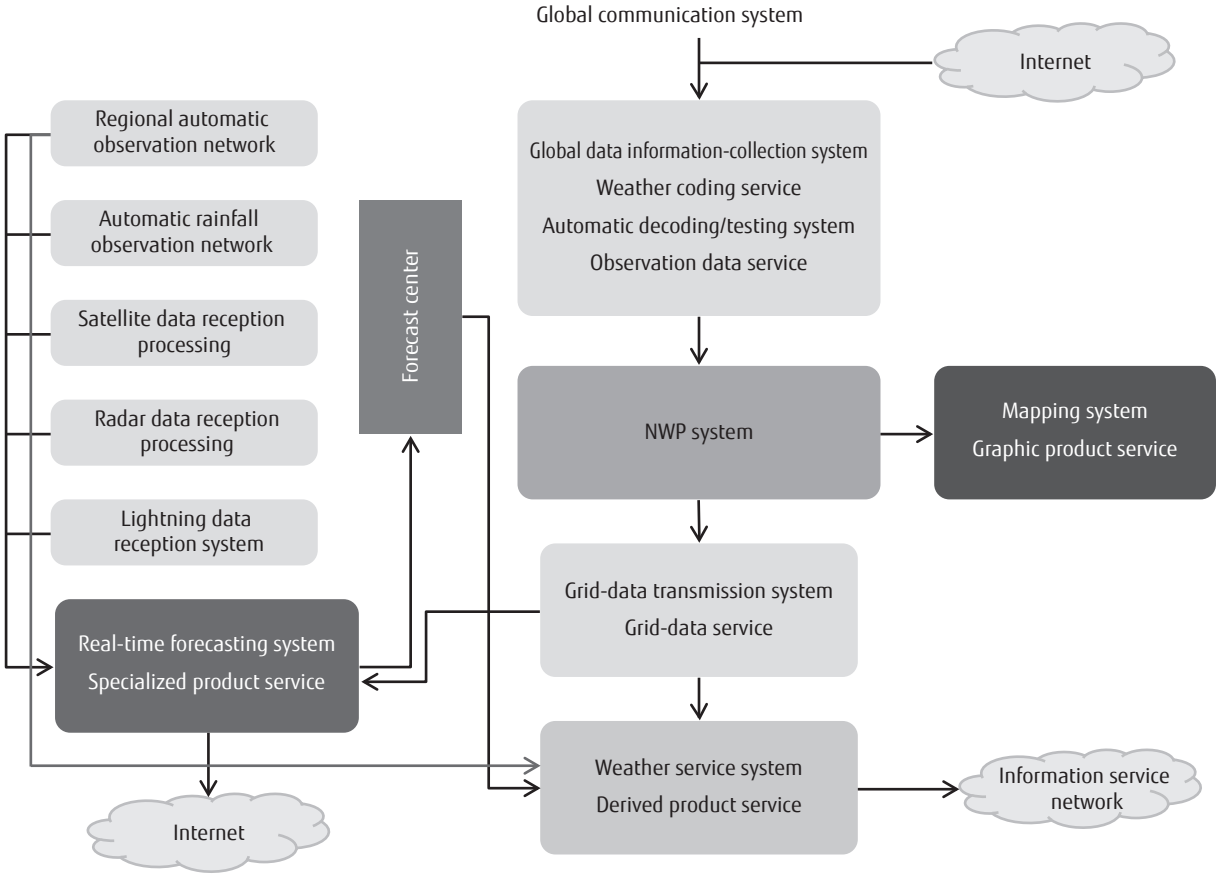

Figure 1
Main upstream and downstream flows and peripheral systems for NWP system.

can directly access and use that system's data.

The system controlling the execution of the NWP work provides output data in various formats to a mapping system for preparing graphics and to a real-time forecasting system for displaying weather data. Weather forecasters and related departments inside and outside the CWB make use of this data.

In addition, the grid-data transmission system provides the results of NWP-model analysis and forecasting to related departments inside and outside the CWB. These services must be provided in a reliable manner within a pre-established amount of time.

The NWP system model consists of two sets of global forecast models, seven sets of regional forecast models (Weather Research and Forecasting [WRF] model and Non-hydrostatic Forecast System [NFS]), and four sets of ensemble forecast models (Ensemble Forecast System [EFS]). All of these models, plus post-forecast processing for each one and evaluation calculations within various meteorological fields, are executed in accordance with a daily schedule.

In this NWP work, it is extremely important that large-scale calculations be performed accurately and that forecast results be provided to related systems within a pre-established timeframe to ensure the safety of the Taiwanese people.

4. NWP system

The NWP system is centered about a work process that inputs various types of observation data, executes the NWP system model, and outputs forecast results. In terms of hardware, the NWP system consists of the PRIMEHPC FX10 supercomputer (four racks), peripheral servers (FUJITSU Server PRIMERGY), and disk storage systems (FUJITSU Storage ETERNUS).

The configuration and features of the NWP system are described below.

4.1 Configuration

The configuration of the NWP system is shown in **Figure 2**. This system consists of two compute clusters forming an active system ("inside" compute cluster 1) and a backup system ("outside" compute cluster 2) and a Fujitsu Exabyte File System (FEFS) cluster forming a high-performance, scalable common file system.

1) Active system (inside compute cluster 1)

The inside cluster executes jobs in accordance with the NWP system model and makes actual weather forecasts. Jobs for making numerical weather predictions are written in shell script (ksh), and the linking and running of many shell scripts generates a product (forecast).

The inside cluster includes a supercomputer (PRIMEHPC FX10: 2 racks, 192 compute nodes, and 10 IO nodes), a login node (4 PRIMERGY RX200 S7 servers), and a datamover node (4 PRIMERGY RX200 S7 servers). Jobs are submitted by the cron job-scheduler utility from the login node and executed on compute nodes. The datamover node is used to obtain input data for job execution from upstream systems and to provide prepared products (forecasts) to downstream systems.

2) Backup system (outside compute cluster 2)

As the name implies, the main role of the backup system (outside cluster) is to act as a backup to the inside cluster. To this end, it is capable of continuing job execution if operations on the inside cluster should become disabled due to a system failure or another problem. This outside cluster also serves as a platform for executing jobs for research and development purposes and for creating and executing jobs by researchers and application developers both inside and outside the CWB.

The outside cluster includes a supercomputer (PRIMEHPC FX10: 2 racks, 192 compute nodes, and 10 IO nodes), a login node (2 PRIMERGY RX200 S7 servers), and a datamover node (2 PRIMERGY RX200 S7 servers). The login and datamover nodes are used in the same way as on the inside cluster.

3) FEFS cluster

The FEFS cluster constitutes a storage system shared by the inside and outside systems. The data storage area consists of an Object Storage Server (OSS) system for access processing (two sets of two PRIMERGY RX300 S7 servers = four servers) and an Object Storage Target (OST) system as a storage disk array for saving FEFS data (two sets of four ETERNUS DX410 S2 disk storage systems = eight disk storage systems). This equates to two OSS groups, each consisting of two OSS servers and four OST systems. Two OSS servers form a redundant configuration, enabling processing to be continued on one server if a failure occurs on the other. An 8-Gb/s fiber channel (FC) connects the OSS and OST systems.

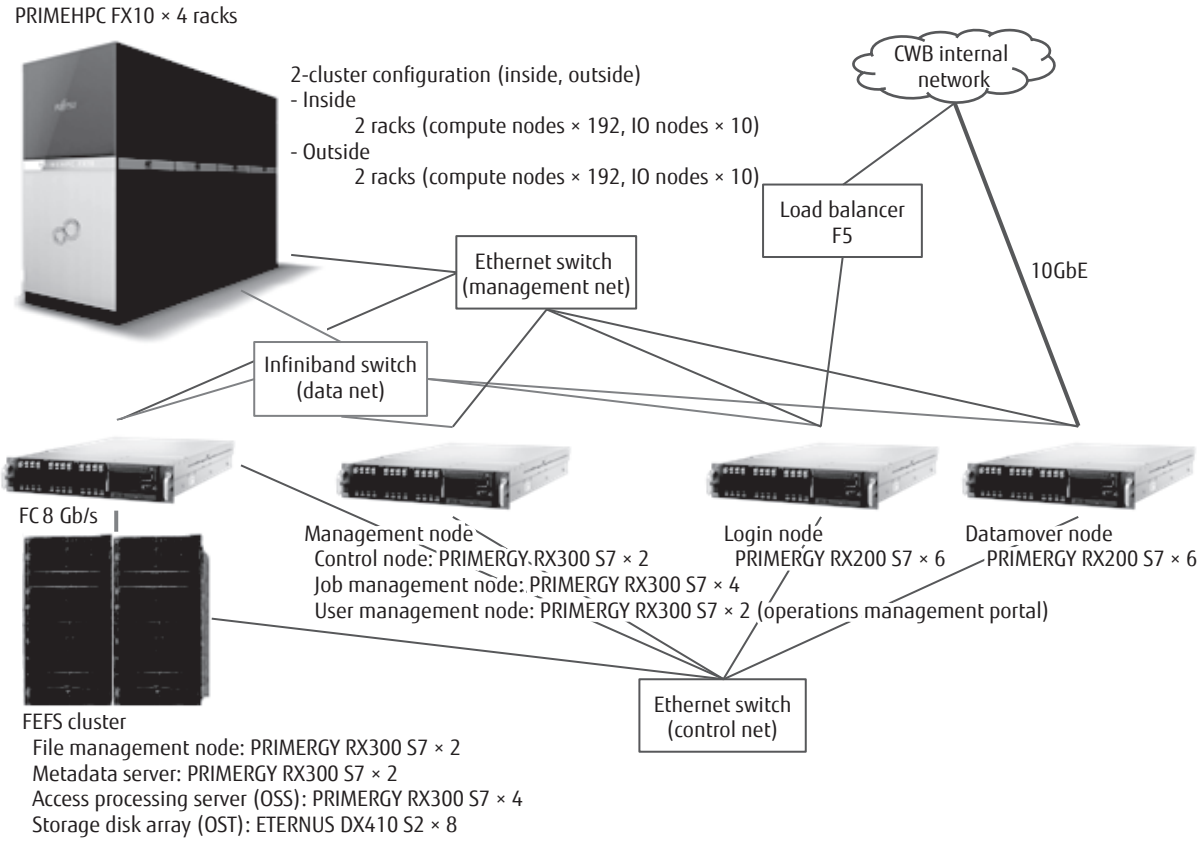

Figure 2
Configuration of NWP system.

Each of the 8 OST systems mounts 62 900-GB, 2.5-inch serial attached SCSI (SAS) disks divided into 9 redundant arrays of inexpensive disks (RAID) groups. Specifically, the 62 disks are broken down into 1 group of RAID 1 + 0 (2 + 2 disks), 8 groups of RAID 5 (6 + 1 disks), and 2 disks as hot spares. The RAID 1 + 0 group is an area for storing the program group for running NWP jobs and features mirroring to ensure high reliability. In short, each of the eight OST systems has nine RAID groups. Combining eight RAID groups configures a single file system, so there are a total of nine file systems using all eight OST systems.

4.2 Features

The NWP system features a redundant configuration for all equipment including network devices so that work can proceed even in the event of a system failure. It also features two compute clusters so that availability is increased. Since the outside cluster serves as a backup system, it can be used to perform work such as software patching that could affect system operations. If no problems are encountered, the same work can then be performed on the inside cluster.

5. Stability test

Given the important role played by the NWP system, it was essential that the work of forecasting continue uninterrupted when switching processing from the old system to the new system. We thus performed a one-month stability test after porting each NWP model in the NWP system to the new system. Specifically, we performed the same NWP work on the old and new systems in parallel for a period of one month. We checked the forecast results of each NWP model and evaluated the accuracy and stability of the new NWP system.

We made the conditions for passing this stability test severe: an availability of 99% was required for the 30-day (720-hour) test period, which means that allowed work stoppage time was only 7.2 hours.

As described above, the NWP system features a redundant configuration for all devices so that work will not come to a stop in the event of a device failure. However, the possibility of a total system halt due to a double failure or software problem must be considered. To facilitate a quick response and system restoration if this were to happen, we prepared monitoring tools and established a 24-hour monitoring system for both hardware and software components.

Although problems were uncovered in the file system and work model during the stability test, speedy, round-the-clock response and restoration measures by engineers monitoring the system kept the work stoppage time to within the allowed 7.2 hours, and an availability of more than 99% was achieved.

6. Conclusion

This paper described the work of numerical weather prediction and the system used for performing this work at the Central Weather Bureau of Taiwan. It also introduced Fujitsu's PRIMEHPC FX10 supercomputer and its features.

This system is being implemented in three stages. The first stage in 2012 brought the system online, and the second stage in 2013 expanded the system. The third stage in 2014 is expanding the system even further.

Looking to the future, the authors and Fujitsu Taiwan will endeavor to make effective upgrades to the NWP system centered about the PRIMEHPC FX10 supercomputer to improve its operability and quality. In this way, we expect to contribute to further improvements in weather observations and forecast accuracy at the Central Weather Bureau of Taiwan and in observations and analysis of long-term climate change in Taiwan. It would give us great satisfaction if our work here helps to mitigate the effects of weather-related disasters in Taiwan and enhance disaster prevention measures.

Fumihiro Takehara
Fujitsu Ltd.
Mr. Takehara is engaged in the introduction of numerical weather prediction systems and support of system operations work at the Central Weather Bureau of Taiwan.

Hidenori Hayashi
Fujitsu Ltd.
Mr. Hayashi is engaged in the introduction of numerical weather prediction systems and support of system operations work at the Central Weather Bureau of Taiwan.

Junichi Fujita
Fujitsu Taiwan Ltd.
Mr. Fujita is engaged in the introduction of numerical weather prediction systems and support of system operations work at the Central Weather Bureau of Taiwan.