Chapter 1

Computer System

1.1 Introduction

The Japan Meteorological Agency (JMA) installed its first-generation computer (IBM 704) to run an operational numerical weather prediction model in March 1959. Since then, the computer system at JMA has been repeatedly upgraded, and the current system (HITACHI SR16000) was completed in June 2012 as the ninth-generation computer. Figure 1.1.1 shows the history of computers at JMA, their peak performance, and a change in peak performance calculated using Moore's law[1] from the first computer (IBM 704). The peak performance of the second (HITAC 5020), the third (HITAC 8800), and the eighth (HITACHI SR11000) computers at the beginning of their implementation was almost the same as that projected using Moore's law, while it was lower during the 1980s, 1990s, and the first half of the 2000s. Recent growth is faster and the peak performance of the current computer is higher than the projection.

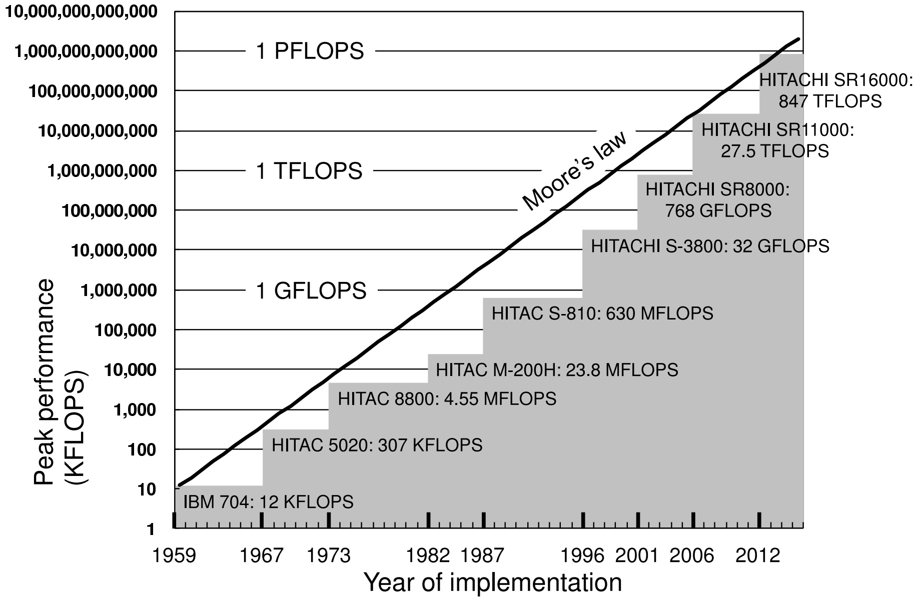

Figure 1.1.1: History of computers used at JMA and their peak performance. The line "Moore's law" represents the projection of peak performance using Moore's law from the first computer (IBM 704).

[1] The term "Moore's law" has many formulations. Here we refer to exponential growth of peak performance which doubles every 18 months.

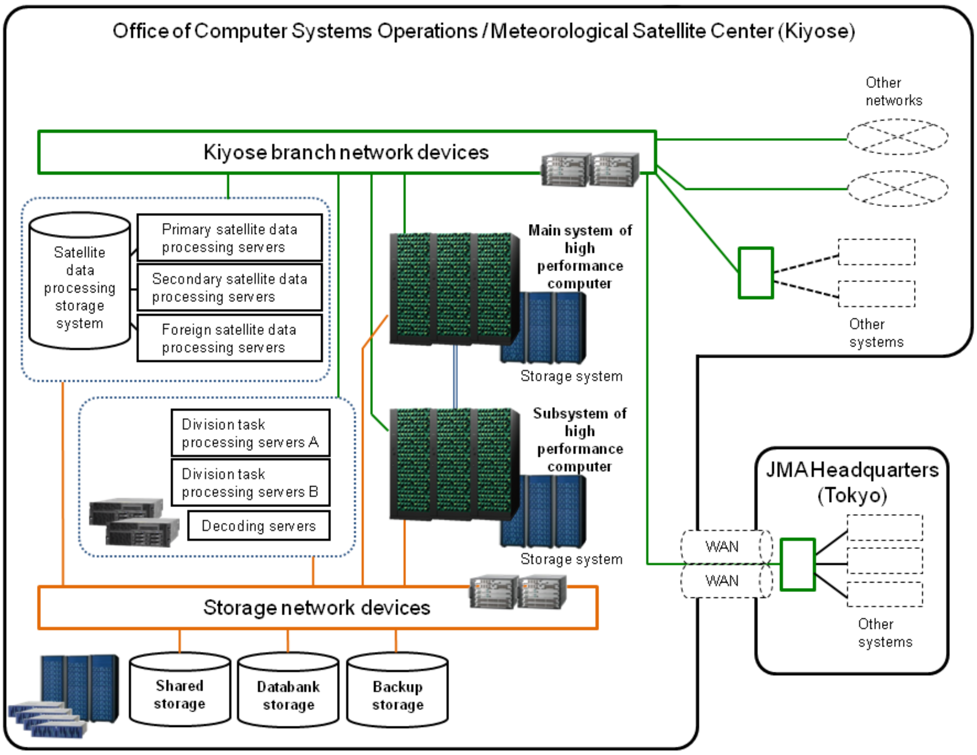
Office of Computer Systems Operations / Meteorological Satellite Center (Kiyose)
Other networks
Kiyose branch network devices
Satellite data processing storage system
Primary satellite data processing servers
Secondary satellite data processing servers
Foreign satellite data processing servers
Main system of high performance computer
Storage system
Other systems
Division task processing servers A
Division task processing servers B
Decoding servers
Subsystem of high performance computer
Storage system
JMA Headquarters (Tokyo)
WAN
WAN
Other systems
Storage network devices
Shared storage
Databank storage
Backup storage

Figure 1.2.1: Schematic illustration of computer system

In this chapter, Section 1.2 briefly describes the configurations and specifications of the current computer system at JMA. Section 1.3 outlines the operational suite and the operational job management system on the current computer system.

1.2 System Configurations and Specifications

1.2.1 Overview

Figure 1.2.1 illustrates major components of the computer system at JMA including HITACHI SR16000 model M1 high performance computers, server computers, storages, terminals, and networks. This system has been in operation since 5 June 2012. Most of the computing facilities are installed at the site of the Office of Computer Systems Operations and the Meteorological Satellite Center in Kiyose City located 24 km west of the JMA headquarters (HQ) in central Tokyo. A wide area network (WAN) connects the Kiyose site and the HQ site. The specifications of the high performance computers and server computers are summarized in Table 1.2.1 and Table 1.2.2, respectively.

1.2.2 High Performance Computer

Two independent systems called a main system and a subsystem of an SR16000 model M1 high performance computer with the same specifications are installed at the Kiyose site. The main system usually runs operational numerical weather prediction jobs, while the subsystem usually runs development jobs. However, in case the main system is under maintenance or out of order, the subsystem runs operational jobs to make the system stable for operational use.

Table 1.2.1: Specifications of high performance computers

Computer	SR16000 model M1
Processor, clock frequency	IBM POWER7, 3.83 GHz
Cores per processor	8
Cores per logical node	32
Logical nodes per system	412 (computation), 10 (I/O), 4 (system), 6 (spare)
Number of systems	2
Peak performance per logical node	0.98 TFLOPS
Peak performance per system	423.5 TFLOPS (total), 403.9 TFLOPS (computation)
Memory per logical node	128 GiB
Memory per system	54.0 TiB (total), 51.5 TiB (computation)
Operating system	AIX 7.1

Table 1.2.2: Specifications of server computers

	Primary satellite data processing servers	Secondary satellite data processing servers	Foreign satellite data processing servers
Computer	EP8000/750	EP8000/750	HA8000/RS220AK1
Processor, clock frequency	IBM POWER7, 3.0 GHz	IBM POWER7, 3.0 GHz	Intel Xeon X5670, 2.93 GHz
Cores per processor	8	8	6
Cores per server	16	16	12
Number of servers	3	6	2
Memory per server	128 GiB	128 GiB	32 GiB
Operating system	AIX 6.1	AIX 6.1	Linux

	Division processing servers A	Division processing servers B	Decoding servers
Computer	BS2000	EP8000/520	EP8000/750
Processor, clock frequency	Intel Xeon E5640, 2.66 GHz	IBM POWER6+, 4.7 GHz	IBM POWER7, 3.0 GHz
Cores per processor	4	2	8
Cores per server	8	2	16
Number of servers	16	2	2
Memory per server	48 GiB	32 GiB	64 GiB
Operating system	Linux	AIX 6.1	AIX 6.1

The computational basis of SR16000 model M1 is an IBM POWER7 processor with a clock frequency of 3.83 GHz. One socket of a POWER7 processor is a multi-core chip which has eight separate cores.

One logical node comprises one multi-chip module with four sockets of POWER7 processors and dual inline memory modules. Therefore, the number of cores in one logical node is $4 \times 8 = 32$. The theoretical performance per logical node is 980.48 GFLOPS and the total memory capacity is 128 GiB per logical node[2]. The inter-node communication rate between each POWER7 processor and a hub processor is 96 GiB/s for one-way communication.

One physical node consists of eight logical nodes. Therefore, the number of cores in one physical node is $8 \times 4 \times 8 = 256$. The theoretical performance per physical node is 7,843.84 GFLOPS. Each logical node within a physical node is connected to another one with an inter-node communication rate of 24 GiB/s for one-way communication.

One super node consists of four physical nodes. Each physical node within a super node is connected to another one with a communication rate of 5 GiB/s for a one-way path. Eight paths are available for each communication between physical nodes, and therefore the total rate of communication becomes 40 GiB/s for one-way communication.

One system is composed of fourteen super nodes. Since one of fourteen super nodes in a system has only two physical nodes, the total numbers of logical nodes and physical nodes within a system are 432 and 54,

[2] The International Electrotechnical Commission approved names and symbols for the power of $2^{10} = 1,024$ instead that of 1,000 for prefixes of units. Symbols such as GiB or TiB refer to the former sense. In contrast, symbols such as GB or TB mean the latter.

respectively. In a system, 432 logical nodes are assigned to 412 computational nodes only for computation, 10 I/O nodes for data transfer between the system and external storages, 4 service nodes for system management, and 6 spare nodes as reserve stocks. Each logical node runs the AIX 7.1 operating system independently. Therefore, one system can be regarded as an aggregation of 432 separate computers. The theoretical performance per system is 423.5 TFLOPS for total 432 logical nodes and 403.9 TFLOPS for only 412 computational nodes. The total memory capacity per system is 54.0 TiB for total 432 logical nodes and 51.5 TiB for only 412 computational nodes.

The main system and subsystem have high-speed magnetic disks with capacities of 135 TB and 210 TB, respectively. Every time an operational job running on the main system is completed, the output files from the job are copied to the disk on the subsystem to keep the subsystem ready to run succeeding operational jobs if the operation is switched to it.

1.2.3 Server and Terminal Computers

A number of server computers are installed for various kinds of tasks such as processing and decoding of observational data, analyses of weather charts, management of the operational suite, and other small jobs.

The primary, secondary, and foreign satellite data processing servers are used for automatic data processing of various kinds of satellite observations. The primary satellite data processing servers consist of three servers of EP8000/750 with two IBM POWER7 (3.0 GHz) processors. The secondary satellite data processing servers consist of six servers of EP8000/750 with two IBM POWER7 (3.0 GHz) processors. The foreign satellite data processing servers consist of two servers of HA8000/RS220AK1 with two Intel Xeon X5670 (2.93 GHz) processors.

The division task processing servers A and B are used for weather chart analyses and small operational jobs that are transaction-intensive rather than compute-intensive. The division task processing servers A consist of sixteen servers of BS2000 with two Intel Xeon E5640 (2.66 GHz) processors. The division task processing servers B consist of two servers of EP8000/520 with one IBM POWER6+ (4.7 GHz) processor.

The decoding servers are used for decoding jobs of observational data and consist of two servers of EP8000/750 with two IBM POWER7 (3.0 GHz) processors.

There are other server computers that manage the operational suite of numerical weather predictions, satellite data processing, and other jobs. In addition, there are server and terminal computers to monitor and manage the computer system.

1.2.4 Mass Storage System

Three kinds of storage systems are available to share data between the high performance computers and server computers. They are shared storage, data bank storage, and backup storage systems.

The shared storage system is used from jobs running on the high performance computers or server computers. This system comprises four network-attached storage (NAS) units which consist of RAID 6 magnetic disks[3] with a total capacity of 754 TB.

The data bank storage system is used for long-time archiving. This system comprises six NAS units which consist of RAID 6 magnetic disks with a total capacity of 2,932 TB.

The backup storage system automatically makes backup copies under specified directories of the data bank storage system. This system consists of a tape library, two management servers, and a shared storage between the servers. The total capacity of the backup storage system is 1,520 TB.

1.2.5 Networks

The Kiyose branch network connects the high performance computers, server computers, and other networks and servers out of the computer system mentioned above.

The storage network connects the high performance computers, server computers, shared storage system, databank storage system, and backup storage system.

[3] The term RAID is short for redundant array of independent disks or redundant array of inexpensive disks. In particular, RAID 6 utilizes block-level striping with double distributed parity and provides fault tolerance of two drive failures.

Users at the HQ site remotely log in to computers at the Kiyose site through a WAN (Figure 1.2.1). This WAN consists of two independent links with a transfer speed of 100 Mbps each. One link is used for operational jobs, while the other is used for development jobs. All the network equipment is configured redundantly to avoid a single equipment failure causing a total interruption.

1.3 Operational Aspects

1.3.1 Operational Suite

The operational suite of JMA that will be described in later chapters consists of about 70 job groups including the global analysis, global forecast, and so on. The number of total jobs composing all the job groups is about 10,600 per day. All the jobs are submitted using a parallel job scheduling system, LoadLeveler. The numbers of kinds of constant and variable datasets are about 3,000 and 8,800, respectively.

Figure 1.3.1 illustrates the daily schedule of the operational suite job groups[4] running on the main system of the high performance computer as of February 2013.

1.3.2 Management System of Operational Jobs

There are complicated dependencies between jobs in a job group and between input and output datasets. To manage a vast number of operational jobs and datasets systematically and to assure the jobs run correctly by eliminating man-made errors, JMA developed a comprehensive system using database management systems (DBMSs). All the information about jobs, input and output datasets, and executables is registered in the DBMSs. The dependencies between these elements can be checked using utility programs.

The management system of operational jobs is comprised of four kinds of files, two DBMSs, and several utility programs to register information, check the consistency and so on:

- Files

  Registration form: Information about job groups, jobs, datasets, executables, and so on. A registration form is submitted when jobs are added or deleted, datasets or executables are updated, or the configurations of job groups or jobs are modified.

  Job definition file: Information about a job group and jobs within the job group such as the job group name, the job name, the schedule (time to run), the order of job groups and jobs (preceding job groups and jobs), and computational resources required (the LoadLeveler job class, the number of nodes, the computational time).

  Job control language: Information about executables such as a shell script, a ruby script, an awk script and a load module, and input and output datasets used in each job. A job control language file is converted into a shell script using a utility program to be submitted to LoadLeveler.

  Program build file-format: Information about source files, object modules, libraries, options for compilation, and so on. A program build file-format is converted into a makefile using a utility program to compile load modules.

- DBMSs

  DBMS for registration: Information from the above four files is registered using utility programs.

  DBMS for job management: Information from the DMBS for registration is stored and this information is used by job schedulers.

When a job control language is converted into a shell script, the following procedures are made:

[4] Semi-operational job groups running on the main system of the high performance computer and operational job groups running on the decoding servers are not included here.

Two-digit number in each box indicates analysis or initial time
Aa: Global surface analysis
Ar: Aerosol analysis and forecast
Av: Cross section
Bi: Bias calculation
Da: Global analysis (cycle)
Dp: Global analysis (cycle, product)
Ea: Global analysis (rapid)
Ef: Global forecast
Em: Lateral boundary condition for mesoscale analysis and forecast
Ep: Global analysis (product)
If: Sea ice
K1bg: 1-month ensemble analysis
K1t: 1-month ensemble forecast
K1tpr: 1-month ensemble forecast (product)
K1w: 1-month ensemble forecast
K1wpr: 1-month ensemble forecast (product)
K2wm: 2-week ensemble forecast
K2wmpr: 2-week ensemble forecast (product)
K2ws: 2-week ensemble forecast
K2wspr: 2-week ensemble forecast (product)
K4pr: 3-month forecast (product)
K8e: Seasonal forecast
K8ebg: Seasonal analysis
K8pr: Seasonal forecast (product)
Kapr: Seasonal analysis (product)
Kc: Reanalysis
Kcst: Reanalysis (statistics)
Ko: Ocean analysis
Ks: Sea surface temperature analysis
Ma: Mesoscale analysis
Mf: Mesoscale forecast
Mp: Mesoscale forecast (product)
Mr: Lateral boundary condition for local analysis and forecast
Nm: Ocean wave forecast
Ns: Ocean wave forecast
Oa: Ocean analysis
Of: Ocean forecast
Oz: Ozone analysis and forecast
Qa: Hourly analysis
Ra: Local analysis
Rf: Local forecast
Rp: Local forecast (product)
Sa: Snow depth analysis
Sg: Ozone analysis and forecast
Sst: Sea surface temperature analysis
Te: Typhoon ensemble forecast
Tg: Storm surge forecast
Ti: Typhoon ensemble analysis
Tp: Typhoon ensemble forecast (product)
Ts: Storm surge forecast
Vg: Global verification
Vm: Mesoscale verification
We: Weekly ensemble forecast
Wi: Wind and temperature
Wp: Weekly ensemble forecast (product)
Ws: Significant weather
Wsv: Weekly ensemble forecast perturbation

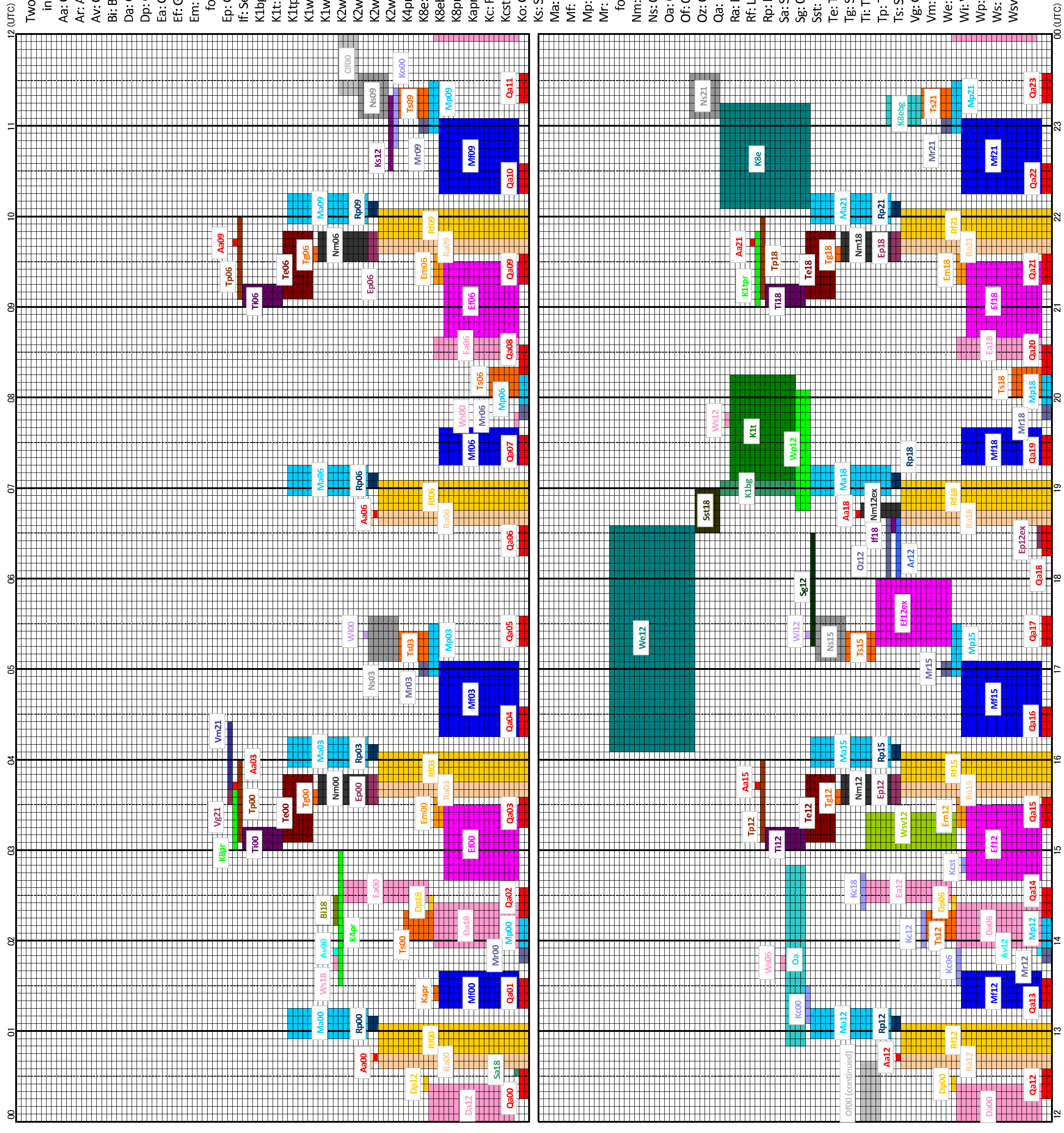

Figure 1.3.1: Daily schedule of operational suite running on main system of high performance computer as of February 2013. Height and width of each box indicate the approximate number of nodes and time range, respectively.

- Existence test: A shell script tests the existence of all non-optional input datasets at the beginning in order to avoid wasting time if the preceding job failed.
- Quasi-atomic output: Every step of a job calling an executable creates output files with temporary names at first and renames them to final names when the step successfully terminates.

The development of the management system of operational jobs was started in 2004 on the seventh computer system and installed in the operational system in 2006 when the eighth computer system was implemented. The number of man-made errors after the inclusion of this management system was reduced to about one sixth of that before the adoption.