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 (Cray XC50) was completed in June 2018 as the tenth-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.

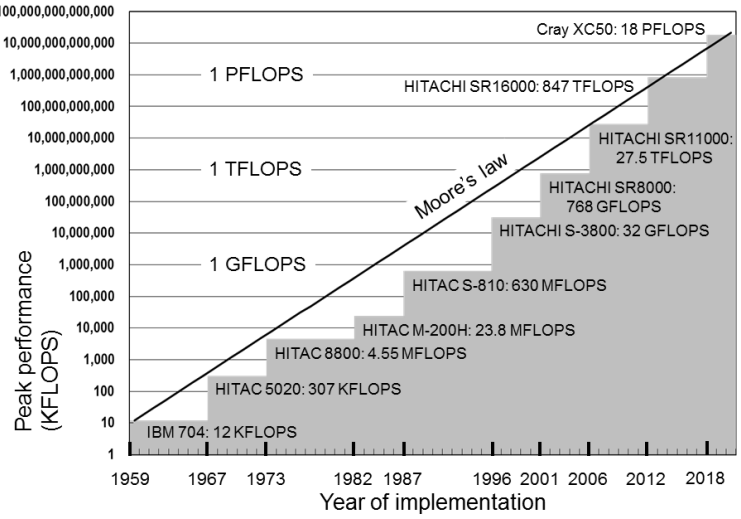

Figure 1.1.1: History of computers used at JMA and their peak performance. The line "Moore's law" represents the projection of peak performance using Moore's law from the first computer (IBM 704).

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

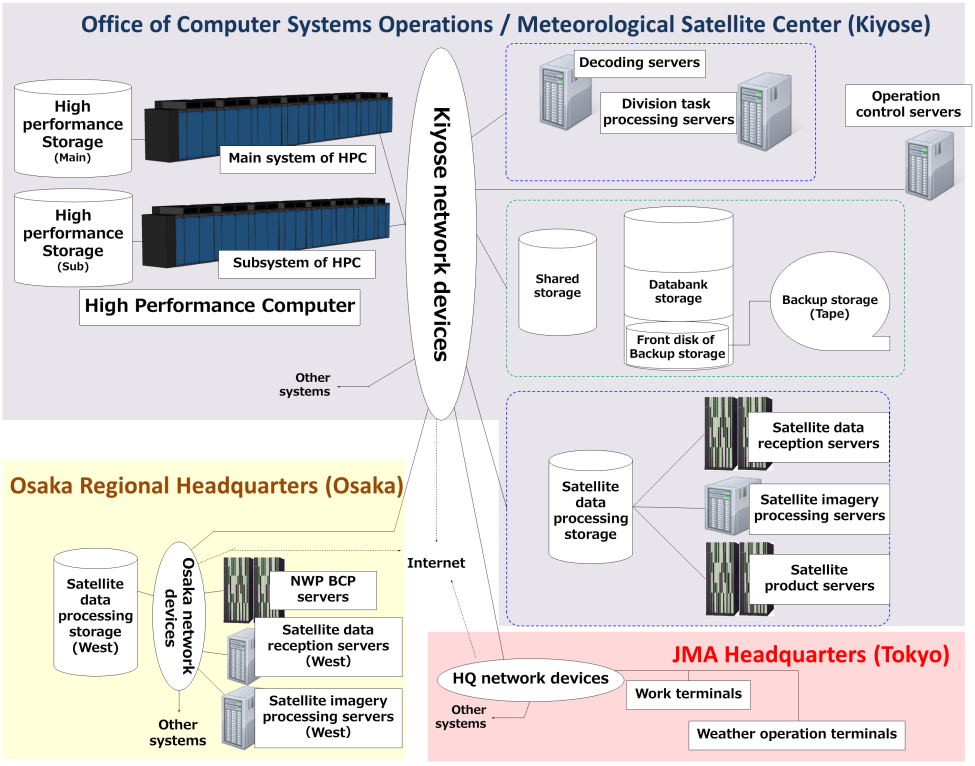

Figure 1.2.1: Schematic illustration of computer system

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 Cray XC50 high performance computers, server computers, storages, terminals, and networks. The system has been in operation since 5 June 2018. Most of the computing facilities are located at the Office of Computer Systems Operations and the Meteorological Satellite Center in Kiyose 24 km west of JMA's central-Tokyo HQ, and some servers are located at the Osaka Regional Headquarters for business continuity planning (BCP). A wide area network (WAN) links the Kiyose, HQ and Osaka sites. The specifications of the high-performance computers and server computers are summarized in Table 1.2.1, Table 1.2.2 and Table 1.2.3, respectively.

1.2.2 High Performance Computer

Two independent systems called a main system and a subsystem of a Cray XC50 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	Cray XC50
Number of systems	2
Computational nodes	
Processor, clock frequency	Intel Xeon Platinum 8160, 2.1 GHz
Cores per processor	24
Cores per logical node	48
Logical nodes per system	2741(ESM), 75(MAMU), 8(Tier2), 16(spare)
Peak performance per logical node	3.2256 TFLOPS
Peak performance per system	9,083 TFLOPS
Memory per logical node	96 GiB
Memory per system	264 TiB
Operating system	Cray Linux Environment 6.0(ESM), SUSE 12.2(MAMU,Tier2)
I/O nodes	
Processor, clock frequency	Intel Xeon E5-2699v4, 2.2 GHz
Cores per processor	22
Cores per logical node	22
Logical nodes per system	2(SDB), 4(PBS-MOM), 2(boot), 2(router), 7(network), 15(LNET), 6(data sync), 2(login gateway)
Memory per logical node	128 GiB
Operating system	SUSE 12.2
Login servers	Dell PowerEdge Server
Processor, clock frequency	Intel Xeon Gold 6148, 2.4 GHz
Cores per processor	20
Cores per logical node	40
Number of servers per system	4
Memory per logical node	768 GiB
Operating system	SUSE 12.2

The Cray XC50 consists of computational nodes, I/O nodes, and login servers.

Each of its computational nodes has two sockets for Intel Xeon Platinum 8160 processors with a clock frequency of 2.1 GHz. One socket of the Xeon processor houses a multi-core chip with 24 separate cores, making $2\times24 = 48$ cores in each logical node. The theoretical performance per logical node is 3.2256 TFLOPS, and the total memory capacity is 96 GiB per logical node[2]. The computational nodes are ESM[3](2741), MAMU[4](75), Tier2[5](8), and spare(16) types. The theoretical performance per system is 9,083 TFLOPS for only ESM and MAMU nodes. Each ESM node runs the CLE(Cray Linux Environment) 6.0 operating system, and each MAMU and Tier2 node runs SUSE Linux Enterprise Server 12.2 independently. The inter-node communication rate between each node and the hub processor is 14 GB/s for one-way communication.

The I/O nodes consist of an Intel Xeon E5-2699v4 (2.2 GHz) processor. The types and numbers of these nodes are SDB[6](2), PBS-MOM[7](4), boot[8](2), router[9](2), network[10](7), LNET[11](15), data sync[12](6), and login gateway[13](2).

The login system involves four Dell PowerEdge servers with two Intel Xeon Gold 6148 (2.4 GHz) processors. The operating system for both I/O nodes and login servers is SUSE Linux Enterprise Server 12.2.

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

[3]Extreme Scalability Mode nodes. Used for high performance Massively Parallel Processing(MPP) runs.

[4]Multiple Applications Multiple User nodes. Used for smaller applications.

[5]Distribution of computational environments to ESM and MAMU.

[6]Service DataBase node with PBS installation.

[7]PBS Mom daemon applied for ESM node.

[8]Used for boot step.

[9]Used for connection with surveillance network.

[10]Used for connection with storage network and servers.

[11]Lustre NETwork node. Used for connection with Lustre high performance storage.

[12]Used for connection with main system and subsystem.

[13]Used for connection with login servers.

Table 1.2.2: Specifications of server computers at Kiyose

	Satellite data reception servers	Satellite imagery processing servers	Satellite product servers
Computer	HPE ProLiant DL360 Gen9	HPE ProLiant DL580 Gen9	HPE ProLiant DL380 Gen9
Processor, clock frequency	Intel Xeon E5-2620v3, 2.4 GHz	Intel Xeon E7-8880v3, 2.3 GHz	Intel Xeon E5-2670v3, 2.3 GHz
Cores per processor	6	18	12
Cores per server	12	72	24
Number of servers	5	8	10
Memory per server	64 GiB	256 GiB	192 GiB
Operating system	RHEL 7.3	RHEL 7.3	RHEL 7.3

	Operation control servers	Division task processing servers	Decoding servers
Computer	HITACHI HA8000 RS210AN1	HPE ProLiant DL580 Gen9	HPE ProLiant DL580 Gen9
Processor, clock frequency	Intel Xeon E5-2640v3, 2.6 GHz	Intel Xeon E7-8880v3, 2.3 GHz	Intel Xeon E7-8860v3, 2.2 GHz
Cores per processor	8	18	16
Cores per server	16	72	64
Number of servers	8	12	2
Memory per server	32 GiB	128 GiB	256 GiB
Operating system	RHEL 7.3	RHEL 7.3	RHEL 7.3

The main system and subsystem have high-performance storage configured with a Lustre file system, and have capacities of 1.6x3 PB each. Every time an operational job running on the main system is completed, its output files are copied to the high-performance storage on the subsystem to ensure that the subsystem is ready to run with further operational jobs if operation is switched to it.

1.2.3 Server and Terminal Computers at Kiyose

A number of server computers are used for various tasks, such as processing and decoding of observational data, weather chart analysis and operational suite management.

The satellite data reception servers, satellite imagery processing servers and satellite product servers are used for automatic processing of various kinds of satellite observation data. The five satellite data reception server are HPE ProLiant DL360 Gen9 units with two Intel Xeon E5-2620v3 (2.4 GHz) processors. The eight satellite imagery processing servers are HPE ProLiant DL580 Gen9 units with four Intel Xeon E7-8880v3 (2.3 GHz) processors. The ten satellite product servers are HPE ProLiant DL380 Gen9 units with two Intel Xeon E5-2670v3 (2.3 GHz) processors.

The eight operation control servers used for control of operational suite job groups are HITACHI HA8000 units with two Intel Xeon E5-2640v3 (2.6 GHz) processors.

The division task processing servers are used for weather chart analysis and small operational jobs that are transaction-intensive rather than compute-intensive. The 12 servers of this type are HPE ProLiant DL580 Gen9 units with four Intel Xeon E7-8880v3 (2.3 GHz) processors.

The two servers used for decoding observational data jobs are HPE ProLiant DL580 Gen9 units with four Intel Xeon E7-8860v3 (2.2 GHz) processors.

Other server computers are also used to that manage the operational suite for numerical weather prediction, satellite data processing and other jobs. Server and terminal computers are additionally used to monitor and manage the computer system.

1.2.4 Mass Storage System

Shared, data bank and backup storage systems are used to share data between high-performance computers and server computers.

The shared and databank storage systems are used for jobs running on high-performance computers or server computers. Configuration involves an IBM Spectrum Scale (ISS) file system with RAID 6 magnetic

disks[14] Shared storage comprises three units with a total capacity of 6PB, and databank storage comprises three units with a total capacity of 25PB, one of which is used as a front disk for backup storage.

The backup storage system is used for long-term archiving. It automatically makes backup copies from the front disk of the data bank storage system, and consists of a tape library and four management servers. Its total capacity is about 80 PB.[15]

1.2.5 Networks

The Kiyose network connects the high-performance computers, server computers and other network/server elements in the computer system described above.

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

Users at HQ remotely log in to computers at the Kiyose site through a WAN consisting of three independent links with transfer speeds of 100 Mbps, 100 Mbps and 1 Gbps (best effort), respectively. The two 100 Mbps links are used for operational jobs, while the 1 Gbps link is used for development jobs. All network equipment is redundantly configured to prevent single failures from causing catastrophic interruption.

The Osaka site is also connected to the Kiyose site through a WAN with two 100 Mbps links.

1.2.6 Server and Terminal Computers at Osaka

Equipment is located in Osaka for NWP BCP operations and redundancy processing of satellite data. There are two HPC ProLiant DL360 Gen9 servers with two Intel Xeon E5-2680v3 (2.5 GHz) processors, which are used for NWP BCP operations.[16] The satellite data reception servers(West) and satellite imagery processing servers(West) are used for processing of satellite observations data in Osaka. The two satellite data reception servers(West) are HPE ProLiant DL360 Gen9 units with two Intel Xeon E5-2620v3 (2.4 GHz) processors. The four satellite imagery processing servers(West) are HPE ProLiant DL360 Gen9 units with two Intel Xeon E5-2698v3 (2.3 GHz) processors.

Table 1.2.3: Specifications of Osaka server computers

	NWP BCP servers	Satellite data reception severs(West)	Satellite imagery processing servers(West)
Computer	HPE ProLiant DL360 Gen9	HPE ProLiant DL360 Gen9	HPE ProLiant DL360 Gen9
Processor, clock frequency	Intel Xeon E5-2680v3, 2.5 GHz	Intel Xeon E5-2620v3, 2.4 GHz	Intel Xeon E5-2698v3, 2.3 GHz
Cores per processor	12	6	16
Cores per server	24	12	32
Number of servers	2	2	4
Memory per server	256 GiB	64 GiB	128 GiB
Operating system	RHEL 7.3	RHEL 7.3	RHEL 7.3

1.3 Operational Aspects

1.3.1 Operational Suite

The JMA operational suite described in later chapters consists of about 80 job groups, including global analysis and global forecasting, with a total of around 20,600 jobs per day. All jobs are submitted from the Routine Operation and Scheduling Environment (ROSE). There are approximately 4,000 and 17,500 constant and variable datasets, respectively.

[14]RAID stands 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 for two drive failures.

[15]The total capacity depends on the volume of the tape cartridge. A capacity of 80 PB is estimated with a 10-TB tape cartridge.

[16]Current NWP BCP operations involve online acquisition of gridded data from overseas NWP center sources and processing to create JMA's product format.

1.3.2 ROSE:Job Scheduler

ROSE is a job flow control computer program that automatically controls execution of all operational jobs. Following on from the start of its development in 2008, it has been used by JMA since 2009 in numerical prediction model development environments and other areas. Based on the current system, the program was adopted for operational NWP control.

ROSE is installed on operational control servers, and is used to control all operational jobs based on submission to PBS[17].

1.3.3 RENS:Operational Job Management System

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 assure that jobs run correctly without human error, JMA developed the comprehensive RENS[18] resource using database management systems (DBMSs).

All job information, input/output datasets, and executables are registered in RENS. Dependencies between these elements can be checked using utility programs.

RENS is comprised of four file types, two DBMSs, and several utility programs to allow registration of information, checking of consistency and other tasks as detailed below.

- 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 PBS 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 PBS.

  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:

- 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 RENS 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.

[17] Portable Batch System (the computer program used to perform job scheduling)

[18] RENS : Routine Environment for Numerical weather prediction System.

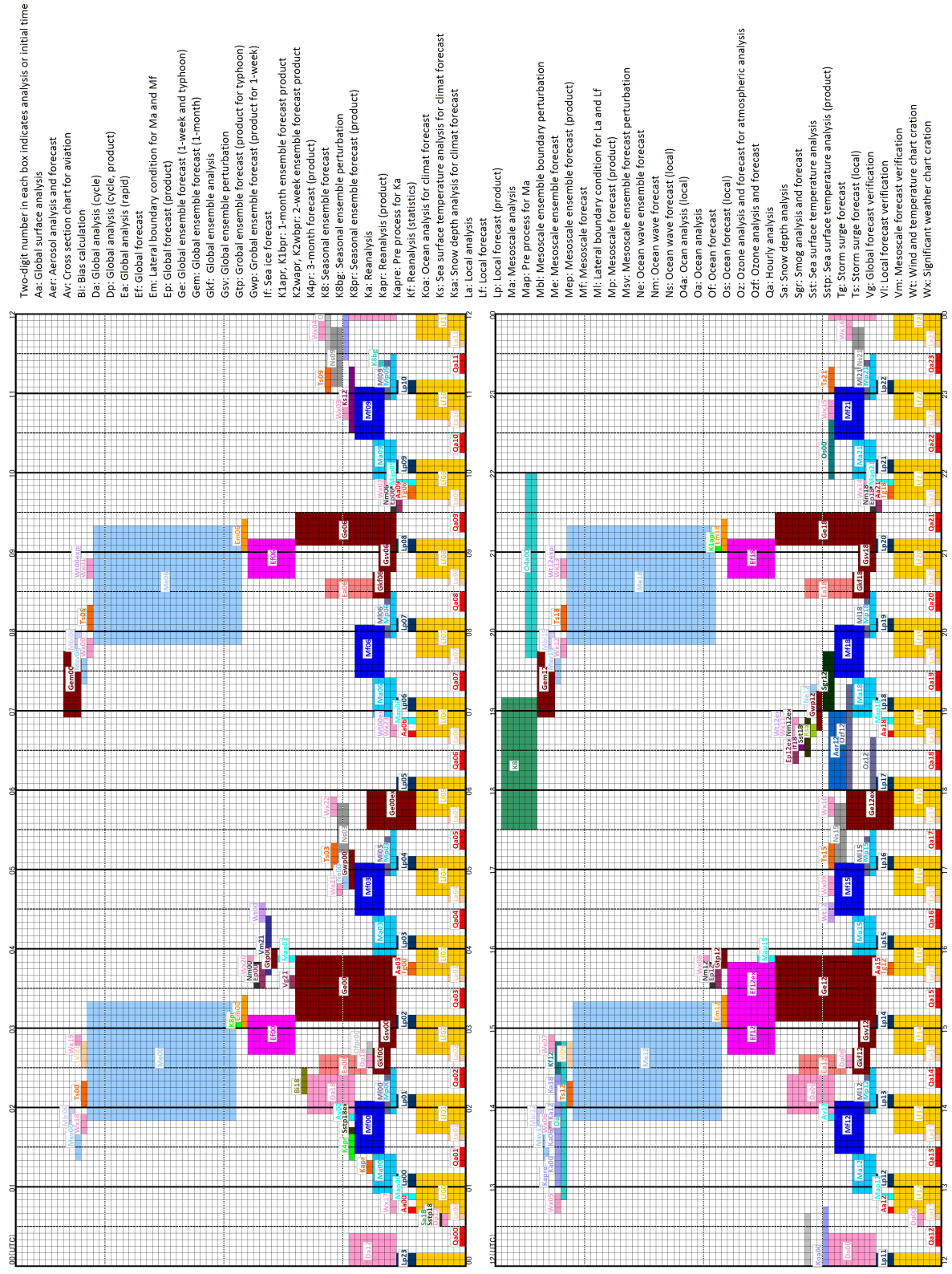

Figure 1.3.1: Daily schedule of the operational suite running on the main system of the high-performance computer as of August 2018. The height and width of each box indicate the approximate number of nodes and the time range, respectively.