Chapter 1

Computer System

1.1 Introduction

The Japan Meteorological Agency (JMA) installed its first-generation computer (IBM 704) to run a numerical weather prediction (NWP) model in March 1959. Since then, the computer systems at JMA have been repeatedly upgraded, and the current systems were completed in March 2024 as the eleventh-generation computer. Table 1.1.1 shows the history of computers at JMA.

Table 1.1.1: History of computers used at JMA

Generation	Date of implementation	Main computer
I	1959/3	IBM 704
II	1967/4	HITAC 5020/5020F
III	1973/8	HITAC 8700/8800
IV	1982/3	HITAC M-200H (2 units)
V	1987/9	HITAC M-680
	1987/12	HITAC S-810/20K
VI	1996/3	HITAC S-3800/480
VII	2001/3	HITACHI SR8000E1
VIII	2006/3	HITACHI SR11000K1 (2 units)
IX	2012/6	HITACHI SR16000M1 (2 units)
X	2018/6	Cray XC50 (2 units)
XI	2023/3	Fujitsu PRIMEHPC FX1000 (2 units)
	2024/3	Fujitsu PRIMERGY CX2550 M7 (2 units)

JMA currently operates a main system built around two on-premise high performance computers (HPCs) at Kiyose (Fujitsu PRIMERGY CX2550 M7 units equipped with Intel Xeon CPUs, simply referred to as the Supercomputer System) and an additional system for prediction of linear precipitation zones built around two HPCs located at Fujitsu's cloud data center (Fujitsu PRIMEHPC FX1000 equipped with Fujitsu A64FX CPUs, referred to as Linear Precipitation Zone Prediction Supercomputer).

Section 1.2 here briefly describes the configurations and specifications of the current computer systems at JMA. Section 1.3 outlines the suite in production and the job management system on the current computer systems.

1.2 System Configurations and Specifications

This section describes the Supercomputer System and the Linear Precipitation Zone Prediction Supercomputer. Both are independent, with loose coupling via closed internal networking and integration.

1.2.1 Supercomputer System

1.2.1.1 Overview

Figure 1.2.1 illustrates major components of the Supercomputer System, including two HPCs, server computers, storages, terminal computers and networks. This system has been in operation since 5 March 2024. Most related facilities are located at the Office of Computer Systems Operations and the Meteorological Satellite Center in Kiyose, 23 km west of JMA's central-Tokyo Headquarters (HQ), with some servers at Fujitsu's cloud data center and others at the Osaka Regional Headquarters for Business Continuity Planning (BCP). Wide area networks (WANs) link Kiyose, HQ, Osaka and the cloud data center. The specifications of the HPCs and server computers are summarized in Table 1.2.1, Table 1.2.2, Table 1.2.3 and Table 1.2.4.

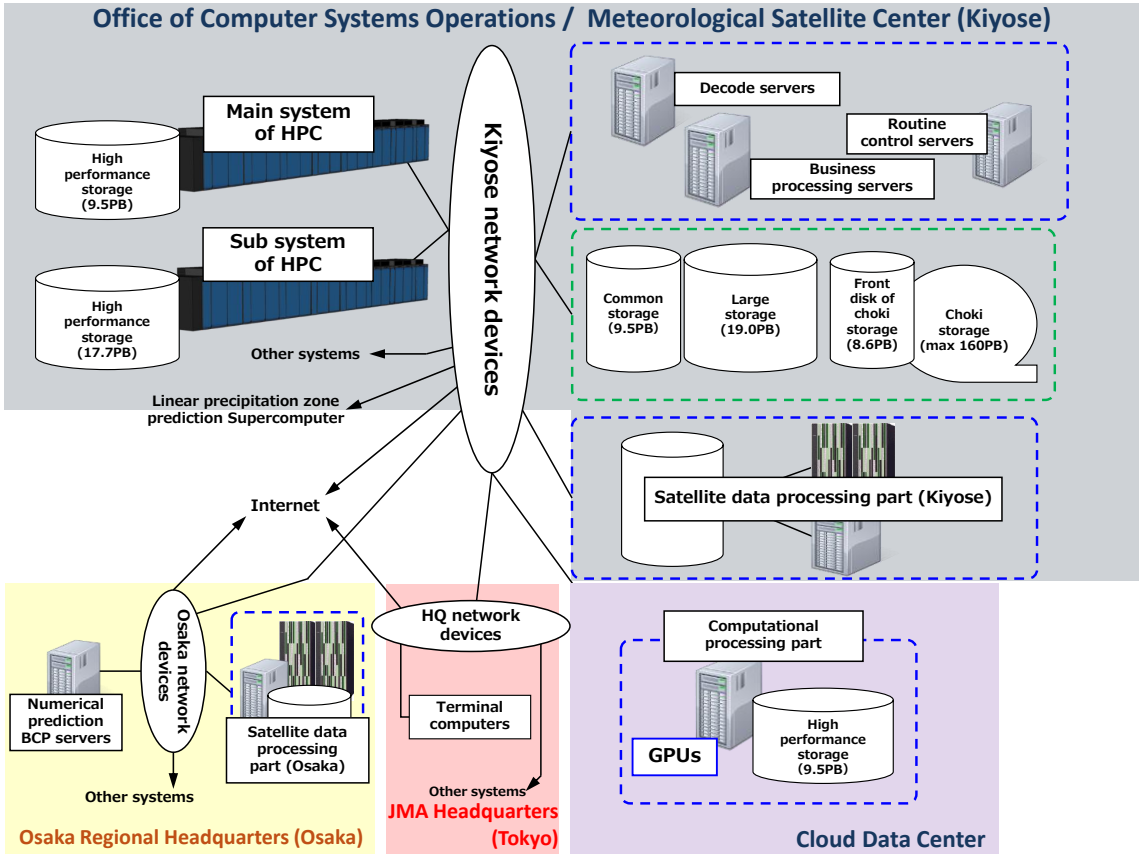

Figure 1.2.1: Schematic illustration of the Supercomputer System

1.2.1.2 High-Performance Computers

The Kiyose site is home to independent facilities referred to as the main system and the sub-system of HPC, with almost identical specifications. The main system hundles operational NWP and the sub-system development. If the main system is under maintenance or out of order, the sub-system takes over.

Table 1.2.1: Supercomputer System HPC specifications. Peak performance, memory and bandwidth per HPC do not include spare nodes.

Number of systems	2(main system and sub-system)
Compute node	Fujitsu PRIMERGY CX2550 M7
Processor, clock frequency	Intel Xeon Max 9480, 1.90 GHz
Cores per processor	56
Cores per logical node	112
Logical nodes per HPC	484(available), 12(spare)
Peak performance per logical node	6.810 TFLOPS
Peak performance per HPC	3.295 PFLOPS
Memory type	HBM2e
Memory per logical node	128 GiB
Memory per HPC	60.5 TiB
Bandwidth per logical node	3280 GB/s
Bandwidth per HPC	1587 TB/s
Operating system	Red Hat Enterprise Linux (RHEL) 8.6
Login node (LN) and job execution login node (JL)	Fujitsu PRIMERGY RX2540 M7
Processor, clock frequency	Intel Xeon Platinum 8458P, 2.70 GHz
Cores per processor	44
Cores per logical node	88
Logical nodes per HPC	4(LN), 6(JL) for main system 7(LN), 3(JL) for sub system
Memory per logical node	1024 GiB
Operating system	RHEL 8.6
Network connection node (NC) and supercomputer connection node (SC)	Fujitsu PRIMERGY RX2540 M6
Processor, clock frequency	Intel Xeon Silver 4314, 2.40 GHz
Cores per processor	16
Cores per logical node	32
Logical nodes per HPC	4(NC), 4(SC)
Memory per logical node	512 GiB
Operating system	RHEL 8.6

Each HPC consists of compute nodes, login nodes, job execution login nodes, network connection nodes, supercomputer connection nodes and other nodes for management.

- Compute node
  Each HPC includes 496 compute nodes, of which 484 are operational and 12 are spares. Each has two sockets for Intel Xeon Max 9480 processors with a clock frequency of 1.90 GHz. One socket in the Xeon processor houses a multi-core chip with 56 separate cores, making $2 \times 56 = 112$ cores in each logical node. The theoretical performance per logical node is 6.810 TFLOPS, and that per HPC (not including spare nodes; the same applies hereafter) is 3.295 PFLOPS. An Intel Xeon Max 9480 processor uses HBM2e memory with high bandwidth. The total memory capacity per logical node and that per HPC are 128 GiB[1] and 60.5 TiB, respectively. The total bandwidth per logical node and that per HPC are 3280 GB/s and 1587 TB/s, respectively. Inter-node communication is implemented with InfiniBand NDR200, and allows each node to communicate at 200 Gbps.

- Login nodes (LNs)
  HPC users log in via an LN consisting of a Fujitsu PRIMERGY RX2540 M7 with two Intel Xeon Platinum 8458P (2.70 GHz) processors to perform various interactive processes and submit batch jobs. The main system has four LNs, while the sub-system has seven.

- Job execution login nodes (JLs)
  JLs have the same specifications as LNs, but are batch-job submission destinations for various general-

[1]The International Electrotechnical Commission approves names and symbols for the power of $2^{10} = 1,024$ rather than 1,000 for unit prefixes. Symbols such as GiB and TiB refer to the former, while those such as GB and TB refer to the latter.

purpose processes (other than calculation) rather than login. The main system has six JLs, while the sub-system has three.

- Network connection nodes (NCs)
  Each HPC includes four NCs using Fujitsu PRIMERGY RX2540 M6 with two Intel Xeon Silver 4314 (2.40 GHz) processors. These are used for connection with storage networks and servers.

- Supercomputer connection nodes (SCs)
  Similarly, each HPC includes four SCs, each using Fujitsu PRIMERGY RX2540 M6 with two Intel Xeon Silver 4314 (2.40 GHz) processors. The NC and an SC specifications are similar, but connections differ. SCs are used for connection with the main system and the sub-system.

Table 1.2.1 summarizes the nodes used.

The main system and the sub-system feature high-performance storage configured with a Lustre file system, with capacities of 9.5 and 17.7 PB, respectively. Files output from the main system are stored on the sub-system via SCs to ensure readiness for further operation on the sub-system as required.

1.2.1.3 Server and Terminal Computers at Kiyose and HQ

The various server computers at Kiyose are outlined below.

- Routine control servers
  Six routine control servers are used to control job groups using Fujitsu PRIMERGY RX2530 M6 units with two Intel Xeon Silver 4310 (2.10 GHz) processors.

- Business processing servers
  These are used for weather chart analysis and small jobs that are transaction-intensive rather than computer-intensive. The 10 servers of this type have Fujitsu PRIMERGY RX2530 M6 units with two Intel Xeon Gold 5318Y (2.10 GHz) processors.

- Decode servers
  The two servers used for decoding observational data jobs are Fujitsu PRIMERGY RX2540 M6 units with two Intel Xeon Gold 5318Y (2.10 GHz) processors.

A summary of these servers is shown in Table 1.2.2.

Additional servers are used at Kiyose for automatic processing of satellite observation data. Other servers are also used to manage NWP, satellite data and other content, and to monitor and manage the computer system. Terminal computers at Kiyose and HQ are additionally used for weather chart analysis and for computer monitoring and management.

Table 1.2.2: Kiyose server computers specifications

	Routine control servers	Business processing servers	Decode servers
Computer	Fujitsu PRIMERGY RX2530 M6	Fujitsu PRIMERGY RX2530 M6	Fujitsu PRIMERGY RX2540 M6
Processor, clock frequency	Intel Xeon Silver 4310, 2.10 GHz	Intel Xeon Gold 5318Y, 2.10 GHz	Intel Xeon Gold 5318Y, 2.10 GHz
Cores per processor	12	24	24
Cores per server	24	48	48
Number of servers	6	10	2
Memory per server	16 GiB	256 GiB	256 GiB
Operating system	RHEL 8.6	RHEL 8.6	RHEL 8.6

1.2.1.4 Mass Storage at Kiyose

Common storage, large storage and *choki*[2] storage are used to share data among HPCs and server computers at Kiyose.

Common and large storage are used for HPCs and server computers, with configuration involving a Lustre file system and total capacities of approximately 9.5 and 19.0 PB, respectively.

Choki storage is used for long-term archiving, with automatic backup from front disk. Total capacity depends on the number and type of tape cartridges, with current and maximum capacities of approximately 120 and 160 PB, respectively. The total capacity of the front disk for *choki* storage is around 8.6 PB.

1.2.1.5 Cloud Data Center Computers

Some Supercomputer System equipment, including the computational processing part, is hosted at Fujitsu's cloud data center.

Computational processing part involves six high-performance AI computers, two login nodes (cloud) and nodes for management.

- High-performance AI computers
  High-performance AI computers with GPUs are used for research on GPU porting of NWP models and development of weather models using AI. Each consists of Fujitsu PRIMERGY GX2570 M6 with eight NVIDIA A100 80 GB SXM GPUs and eight InfiniBand HDR (200 Gbps) ports for interaction with other high-performance AI computers. Five of these are active, and the other is a spare.

- Login nodes (cloud)
  Each login node (cloud) consists of Fujitsu PRIMERGY RX2530 M6 with two Intel Xeon Gold 5318Y (2.10 GHz) processors. A login node (cloud) is used for submission of batch jobs to high-performance AI computers.

A summary of these specifications is shown in Table 1.2.3.

The computational processing part has high-performance storage with a Lustre file system and a capacity of approximately 9.5 PB.

Table 1.2.3: Specifications of the computational processing part. Total peak GPU performance and total memory do not include spare computer.

High-performance AI computers	Fujitsu PRIMERGY GX2570 M6
GPU	NVIDIA A100 80 GB SXM
The number of GPUs per computer	8
Processor, clock frequency	Intel Xeon Gold 6338, 2.00 GHz
Cores per processor	32
Cores per computer	64
The number of computers	5(available), 1(spare)
Peak GPU performance per computer	77.6 TFLOPS (64-bit vector)
Total peak GPU performance	388 TFLOPS (64-bit vector)
GPU Memory per computer	640 GiB
CPU Memory per computer	2048 GiB
Total Memory	13.12 TiB
Operating system	RHEL 8.6
Login node (cloud)	Fujitsu PRIMERGY RX2530 M6
Processor, clock frequency	Intel Xeon Gold 5318Y, 2.10 GHz
Cores per processor	24
Cores per logical node	48
The number of logical nodes	2
Memory per logical node	256 GiB
Operating system	RHEL 8.6

[2] *Choki* means long-term in Japanese.

1.2.1.6 Networks

The Kiyose network in 10-Gigabit Ethernet connects HPCs, servers and other network/server elements.

The storage network in InfiniBand HDR connects LNs, JLs and NCs of HPCs, servers, common storage, large storage and *choki* storage.

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 Kiyose site is connected to the cloud data center via WAN with two 10 Gbps links and to the Osaka site via WAN with a 1 Gbps link and a 100 Mbps link.

The Supercomputer System and Linear Precipitation Zone Prediction Supercomputer are connected via two 10 Gbps links at the Kiyose site.

1.2.1.7 Server and Terminal Computers at Osaka

The Osaka site has two Fujitsu PRIMERGY RX2530 M6 servers with two Intel Xeon Silver 4310 (2.10 GHz) processors for NWP BCP operation[3] as per the specifications shown in Table 1.2.4. There are also servers for satellite data processing, and servers and terminals in management.

Table 1.2.4: Specifications of Osaka server computers

	Numerical prediction BCP servers
Computer	Fujitsu PRIMERGY RX2530 M6
Processor, clock frequency	Intel Xeon Silver 4310, 2.10 GHz
Cores per processor	12
Cores per server	24
Number of servers	2
Memory per server	128 GiB
Operating system	RHEL 8.6

1.2.2 Linear Precipitation Zone Prediction Supercomputer

1.2.2.1 Overview

Figure 1.2.2 illustrates major components of the Linear Precipitation Zone Prediction Supercomputer, including two HPCs, servers, storage, terminals and networks. The system has been in operation since 1 March 2023. Most of the computing facilities are located at Fujitsu's cloud data center, with some terminals and network devices at the Office of Computer Systems Operations and the Meteorological Satellite Center in Kiyose. WAN links the cloud data center and Kiyose. HPC specifications are summarized in Table 1.2.5.

1.2.2.2 High-Performance Computers

Along with the Supercomputer System, the Linear Precipitation Zone Prediction Supercomputer has independent HPC main and sub-systems with almost identical specifications at the cloud data center. The main system handles operational NWP and sub-system development. If the main system is under maintenance or down, the sub-system takes over.

Each HPC consists of compute nodes, login nodes, job execution login nodes, network connection nodes and other nodes for management.

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

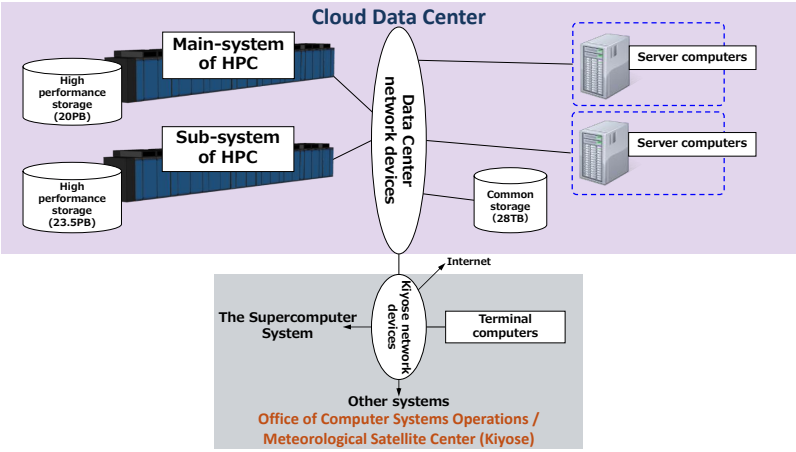

Figure 1.2.2: Linear Precipitation Zone Prediction Supercomputer

- Compute nodes
  Each HPC includes 4,608 compute nodes, of which 4,224 are operational and 384 are spares. Each has a socket for Fujitsu A64FX processors with multi-core chips and 48 separate cores for a clock frequency of 2.20 GHz. Theoretical performance per logical node is 3.3792 TFLOPS, and that per HPC (not including spare nodes; the same applies hereafter) is 14.27 PFLOPS. A64FX processors have HBM2 memory with high bandwidth. The total memory capacity per logical node and that per HPC are 32 GiB and 132 TiB, respectively, with corresponding total bandwidths of 1,024 GB/s and 4,325 TB/s. Inter-node communication is implemented with Torus fusion (TOFU) interconnect D allowing up to 40.8 GB/s of bidirectional interaction at each node.

- Login nodes (LNs)
  LNs consist of Fujitsu PRIMERGY RX2530 M6 with two Intel Xeon Platinum 8380 (2.30 GHz) processors for login to HPCs to allow various interactive processes and submission of batch jobs. Each HPC has 16 LNs.

- Job execution login nodes (JLs)
  JLs and LNs have the same specifications, but JLs are not disclosed as a login destination. Rather, they are batch job submission destinations for various general-purpose processes other than calculation. Each HPC has 12 JLs.

- Network connection nodes (NCs)
  Each HPC includes four NCs, each consisting of a Fujitsu PRIMERGY RX2530 M6 with two Intel Xeon Gold 5320 (2.20 GHz) processors. The NCs are used to exchange data between the main system and sub-system, and with the Supercomputer System.

These nodes are summarized in Table 1.2.5.

The main system and sub-system have high-performance storage configured with a Fujitsu Exabyte File System (FEFS) based on the Lustre file system, with respective high-performance storages of 20 and 23.5 PB. Main system output files are copied to high-performance storage on the sub-system via NCs to ensure ongoing functionality if operation is switched to the sub-system.

1.2.2.3 Networks

The cloud data center network implemented in 10 Gigabit Ethernet connects HPCs and other network/server elements in the computer system described above.

Table 1.2.5: Specifications of Linear Precipitation Zone Prediction Supercomputer HPCs. Peak performance, memory and bandwidth values per HPC do not include spare nodes.

Number of systems	2(main-system and sub-system)
Compute node	Fujitsu PRIMEHPC FX1000
Processor, clock frequency	Fujitsu A64FX, 2.20 GHz
Cores per processor	48
Cores per logical node	48
Logical nodes per HPC	4224(available), 384(spare)
Peak performance per logical node	3.3792 TFLOPS
Peak performance per HPC	14.27 PFLOPS
Type of Memory	HBM2
Memory per logical node	32 GiB
Memory per HPC	132 TiB
Bandwidth per logical node	1024 GB/s
Bandwidth per HPC	4325 TB/s
Operating system	RHEL 8.3
Login node (LN) and job execition login node (JL)	Fujitsu PRIMERGY RX2530 M6
Processor, clock frequency	Intel Xeon Platinum 8380, 2.30 GHz
Cores per processor	40
Cores per logical node	80
Logical nodes per HPC	16(LN), 12(JL)
Memory per logical node	1024 GiB
Operating system	RHEL 8.4
Network connection node (NC)	Fujitsu PRIMERGY RX2530 M6
Processor, clock frequency	Intel Xeon Gold 5320, 2.20 GHz
Cores per processor	26
Cores per logical node	52
Logical nodes per HPC	4
Memory per logical node	256 GiB
Operating system	RHEL 8.4

A WAN consisting of two independent links with a transfer speed of 10 Gbps connects the cloud data center and the Kiyose site.

As described in 1.2.1.6, the Supercomputer System and Linear Precipitation Zone Prediction Supercomputer are connected by two 10 Gbps links within the Kiyose site. Users at HQ remotely log in to HPCs at the cloud data center via this facility.

1.2.2.4 Other Implementation

As per the common storage of the Supercomputer System described in 1.2.1.4, the Linear Precipitation Zone Prediction Supercomputer also implements common storage with an NFS file system. However, as the 28 TB capacity is much lower than that of the entire system, common storage is used for version control of development files and similar.

The cloud data center hosts servers to manage the entire system, with terminals at Kiyose additionally used for monitoring and management.

1.2.2.5 Inter-system Data Transfer

As per 1.2, the Supercomputer System and the Linear Precipitation Zone Prediction Supercomputer are integrated.

To this end, results from the Supercomputer System are required for Linear Precipitation Zone Prediction Supercomputer operation, while the latter should be aggregated to the former. Accordingly, results from both are interchanged using the NCs of each system and copied from the main HPC system to an HPC sub-system as described in 1.2.1.2 and 1.2.2.2 and exemplified 1.3.2.

As data transmission in production between both systems must be prioritized over that for development, Quality of Service (QoS) provision is incorporated in the WAN between the Kiyose site and the cloud data center.

1.3 Management Aspects

1.3.1 Operational Suite

The JMA operational suite described here involves approximately 70 job groups[4], including global analysis and global forecasting, with more than 31,000 jobs per day. Figures 1.3.1 and 1.3.2 show the daily schedule of job groups in production for the Supercomputer System and the Linear Precipitation Zone Prediction Supercomputer, respectively. All jobs are submitted via the Routine Operation and Scheduling Environment (ROSE) software, with approximately 5,000 and 16,000 ancillary and variable datasets, respectively.

1.3.2 ROSE Job Scheduler

JMA's Routine Operation and Scheduling Environment (ROSE) job flow control application automatically controls execution of all operational jobs (as opposed to some development jobs) in the Supercomputer System and the Linear Precipitation Zone Prediction Supercomputer. Developed in 2008, the application has been used by JMA since 2009 in NWP model development environments and other areas. Since the tenth-generation computerization introduced in 2018, ROSE has been officially adopted for NWP control.

ROSE is used on routine control servers in the Supercomputer System based on the job operation software for the Technical Computing Suite (TCS) and the Portable Batch System (PBS). The software for TCS is used for job submission to HPCs and high-performance AI computers of the Supercomputer System and HPCs of the Linear Precipitation Zone Prediction Supercomputer. The PBS is used for submission to other servers of the Supercomputer System.

1.3.3 RENS: Job Management System

There are complicated dependencies between jobs in job groups and between input and output datasets. To manage a vast number of jobs and datasets systematically and assure that jobs run correctly without human error, JMA developed the comprehensive Routine Environment for Numerical weather prediction System (RENS) 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 files: Information about job groups, jobs, datasets, executables, and so on. Registration files are submitted when jobs are added or deleted, datasets or executables are updated, or configurations of job groups or jobs are modified.

  Job definition files: 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 (e.g., numbers of nodes, computational times).

  Job control language: Information about executables such as a shell script, a ruby script, a python script, an awk script and a load module, and input and output datasets used in each job. Job control language files are converted to shell scripts using a utility program for submission as batch jobs.

[4]Including job groups executed on HPCs of the Supercomputer System and the Linear Precipitation Zone Prediction Supercomputer as well as those executed on Supercomputer System decode servers.

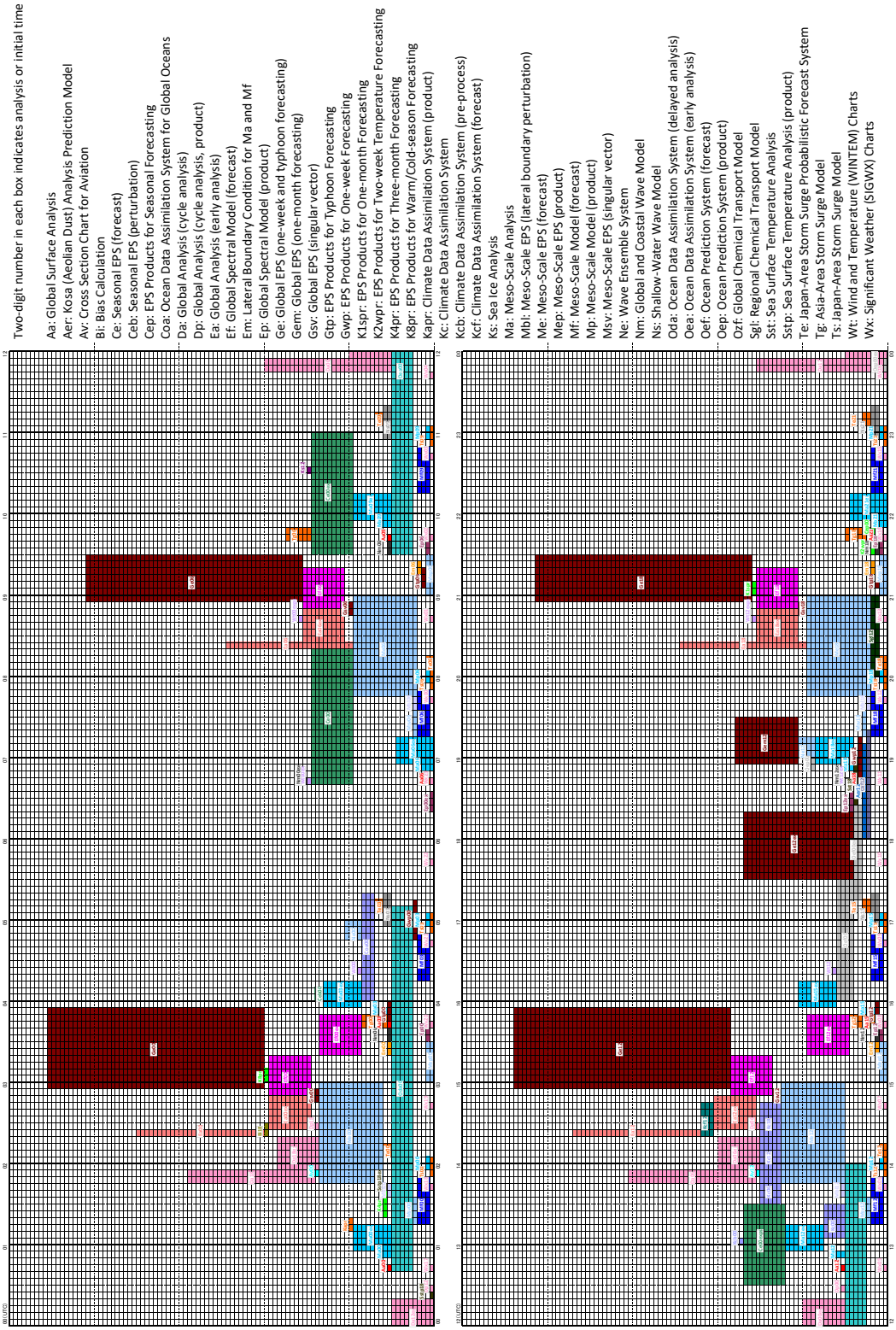

Figure 1.3.1: Daily schedule for the suite of job groups in production on the main HPC facility on the Supercomputer System as of March 2024. The height and width of each box indicate the approximate number of nodes (five per grid) and the time range (five minutes per grid), respectively.

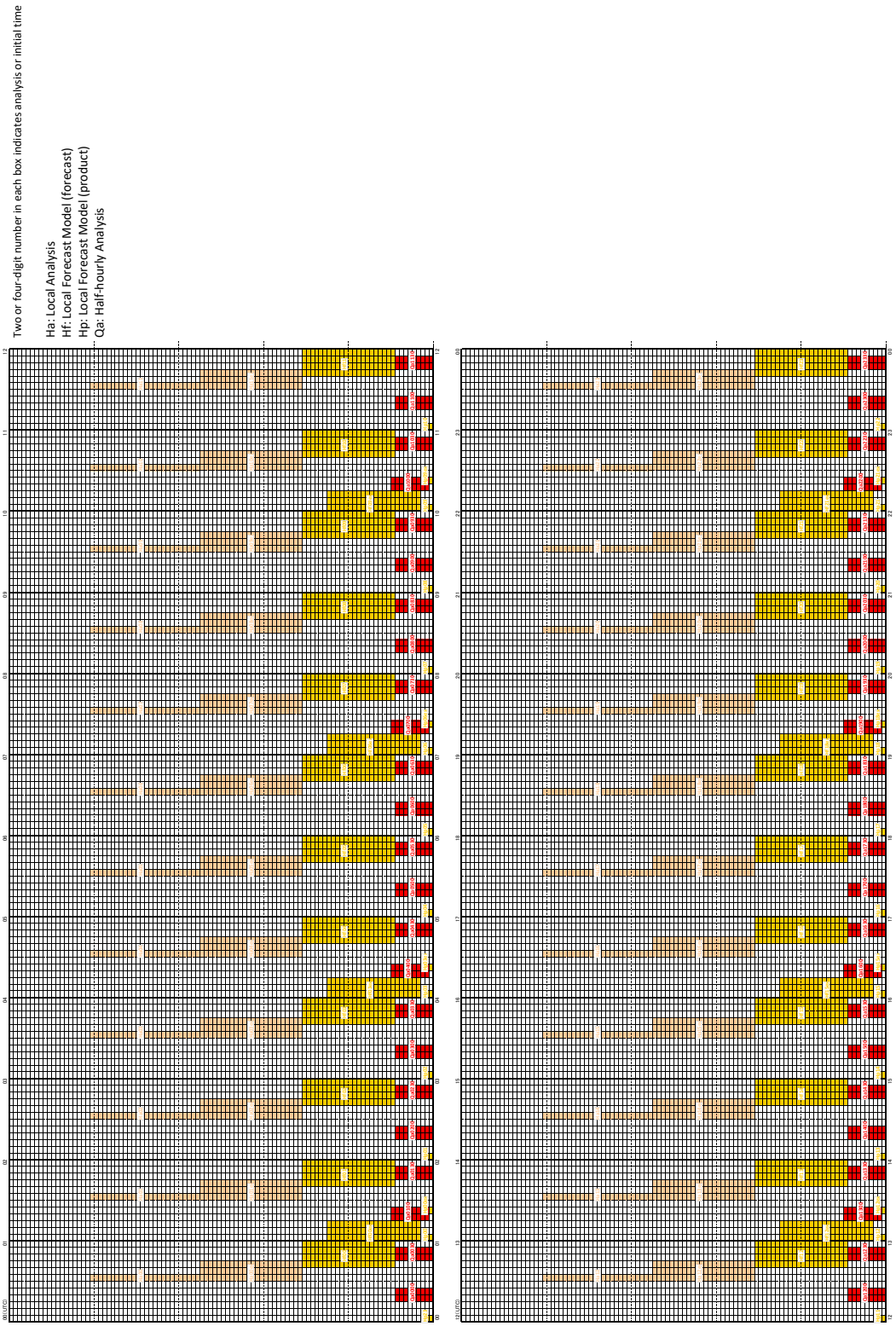

Figure 1.3.2: Daily schedule for the suite of job groups in production on the main HPC facility on the Linear Precipitation Zone Prediction Supercomputer as of March 2024. The height and width of each box indicate the approximate number of nodes (ten per grid) and the time range (five minutes per grid), respectively.

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 file 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: All steps in jobs calling executables create output files with initial temporary names that are finalized once processing terminates.

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