2. COMPUTER SYSTEM

2.1 Introduction

The computer system at the Japan Meteorological Agency (JMA) has been repeatedly upgraded since IBM 704 was firstly installed in 1959. The current system has been completed in March 2006 as the eighth generation since then.

2.2 System configuration and specifications

Figure 2.2.1 illustrates major components of the computer system including HITACHI SR11000 supercomputer, servers, storages, terminals and networks. Most of the computing facilities are installed at the Office of Computer Systems Operations (OCSO) and the Meteorological Satellite Center in Kiyose City, 24 km west of central Tokyo. A wide area network (WAN) connects the Kiyose site and the headquarters (HQ). Specifications of the computers are listed in Table 2.2.1.

2.2.1 Supercomputer (HITACHI SR11000)

Three clusters of SR11000 supercomputer are installed at the Kiyose site. One 50-node cluster of SR11000/J1 (Power5 1.9GHz; hereafter referred to as Cluster 1) and two 80-node clusters of SR11000/K1 (Power5+ 2.1GHz; hereafter referred to as Cluster 2 and 3) have been in operation since March 2005 and March 2006, respectively. The operating system on the supercomputer is AIX 5.2. Each node consists of 16 processors and 64 GiB[1] of memory. Cluster 1 has 6.08 teraflops (TF) of total peak performance and 3.1 TiB of total memory. Cluster 2 or 3 has 10.8 TF of total peak performance and 5.0 TiB of total memory. Each cluster has 6.9 TB magnetic disks connected via storage area network (SAN). As a faster media for data transfer between successive jobs, a part of the memory is configured as extended storage.

SR11000 is a clustered symmetric multiprocessing computer. Nodes in a cluster are interconnected by a full-duplex crossbar switch network to process an MPI (message passing interface) job efficiently. Data transfer speed between two nodes is 8 GiB/s in a single direction. An MPI task uses processors in a node by symmetric multiprocessing (SMP). A processor-intensive job is programmed with combination of MPI and SMP parallelisms.

Cluster 1 and 3 are used for operational runs of the NWP models, and Cluster 2 is used for development and preoperational runs of the NWP models. Cluster 2 is used for operational runs when Cluster 1 or 3 is in maintenance or failure. SAN disks share a fibre channel network, so that the disks for Cluster 1 or 3 can be mounted on Cluster 2 in that case.

[1] Industrial conventions use powers of 1024 instead of that of 1000 for prefixes of units. Symbols such as GiB or TiB refer to the former sense in contrast gigabytes (GB) or terabytes (TB) meaning the latter..

2.2.2 Server Computers

A number of server computers are installed for various kinds of tasks related to NWP, filling wide spectrum of required computing power. High-availability (HA) cluster is used for mission-critical tasks; a cluster has a standby node that takes over a virtual IP address of another node in case of maintenance or failure.

Very Short-Range Forecast model (VSRF; Section 5.4) runs on three-node cluster of EP8000/570 every 30 minutes. Operational run uses two nodes, and another one is a standby node. Another standalone node of the EP8000/570 is installed at HQ for visualization and statistical processing at the forecast operation office.

A notable change from the previous system is the introduction of PC-based servers. Nine two-node clusters of HA8000/130W with Intel Xeon processors are installed at Kiyose site for small-scale batch jobs, such as some part of verification (Section 4.9) or data monitoring.

The JMA NWP Operating System (JNOS; Section 2.3) runs on two-node cluster of EP8000/520 continuously to control batch jobs on supercomputers and servers described above. Another cluster of EP8000/520 called databank server provides an interface to the mass storage system.

2.2.3 Mass storage system

The mass storage system archives a large volume of NWP products used for long-term activities such as verification, model diagnosis, or statistical studies. The system consists of two StorageTek PowderHorn 9310 tape libraries, front disks, archive management server (Three-node HA cluster of SGI Origin350), and the databank server. Each tape library can store up to 5,000 volumes of cartridge tapes each of which has 200 GB, and the maximum capacity of the mass storage system is 2.00 petabytes (PB). All equipment is connected to the SAN network, and 18 TB front disk is attached for each tape library.

The archive management server provides hierarchical storage management for computers on the SAN including supercomputers. Computer output is written first to the front disk, and later migrated to the tape media when its amount exceeds an arranged value. Wherever it is, the data can be accessed as a UNIX file through CXFS filesystem (Shepard and Eppe, 2003).

Computers out of the SAN can retrieve the archived data from the databank server using hypertext transfer protocol (HTTP). Grid point value (GPV) data in NuSDaS format (Toyoda, 2005) can be accessed by HTTP-based Pandora Protocol (Toyoda, 2002), and data in other formats are served by conventional file-based HTTP.

2.2.4 Terminal computers

Many IA-32 based terminal computers are installed in both Kiyose site and HQ. Terminals at OCSO are used to

monitor and operate the computer system. Terminals at the forecast operation office of HQ are used to display the analysis and forecast result of NWP and perform the forecast operation. Peripheral devices such as printers and I/O media are installed in NWP-related offices in HQ.

2.2.5 Networks

Gigabit local area networks (LANs) connecting all equipment are implemented both in Kiyose site and HQ. Each LAN has network-attached storage (NAS) disks to store small, short-living or frequently updated data such as UNIX home directory or program source codes.

The WAN between Kiyose and HQ LANs consists of two independent links of 100 Mbps. Operational traffic uses one link, and the other is standby that is normally used for less-critical tasks such as development.

Table 2.2.1 Specifications of computers.

	Supercomputer		Server	
	Cluster 1	Cluster 2/3	VSRF	Fcst Office
Operational since	March 2005	March 2006	March 2006	March 2006
Model	SR11000/J1	SR11000/K1	EP8000/570	EP8000/570
Processor	Power 5	Power 5+	Power5	Power5
	1.9 GHz	2.1 GHz	1.65 GHz	1.65 GHz
Processors per node	16	16	8	8
Nodes per cluster	50	80	3	1
Clusters	1	2	1	1
Node Peak Performance	121.6 GF	134.4 GF		
Cluster peak performance	6.08 TF	10.8 TF		
Node SPECint_rate2000			85	85
Memory per node (GiB)	64	64	16	16
Memory per cluster (TiB)	3.1	5.0		
Cluster-shared disk	6.878 TB	6.878 TB	694 GB	—
Operating system	AIX 5.2	AIX 5.2	AIX 5.2	AIX 5.2

	Server (*continued*)		
	Databank	JNOS	IA-32
Operational since	March 2006	March 2006	March 2006
Model	EP8000/520	EP8000/520	HA8000/130W
Processor	Power5	Power5	Xeon
	1.65 GHz	1.65 GHz	2.80 GHz
Processors per node	2	2	2
Nodes per cluster	2	2	2
Clusters	1	1	9
Node SPECint_rate2000	23	23	18.2
Memory per node (GiB)	4	4	2
Cluster-shared disk	—	202 GB	—
Operating system	AIX 5.2	AIX 5.2	Linux

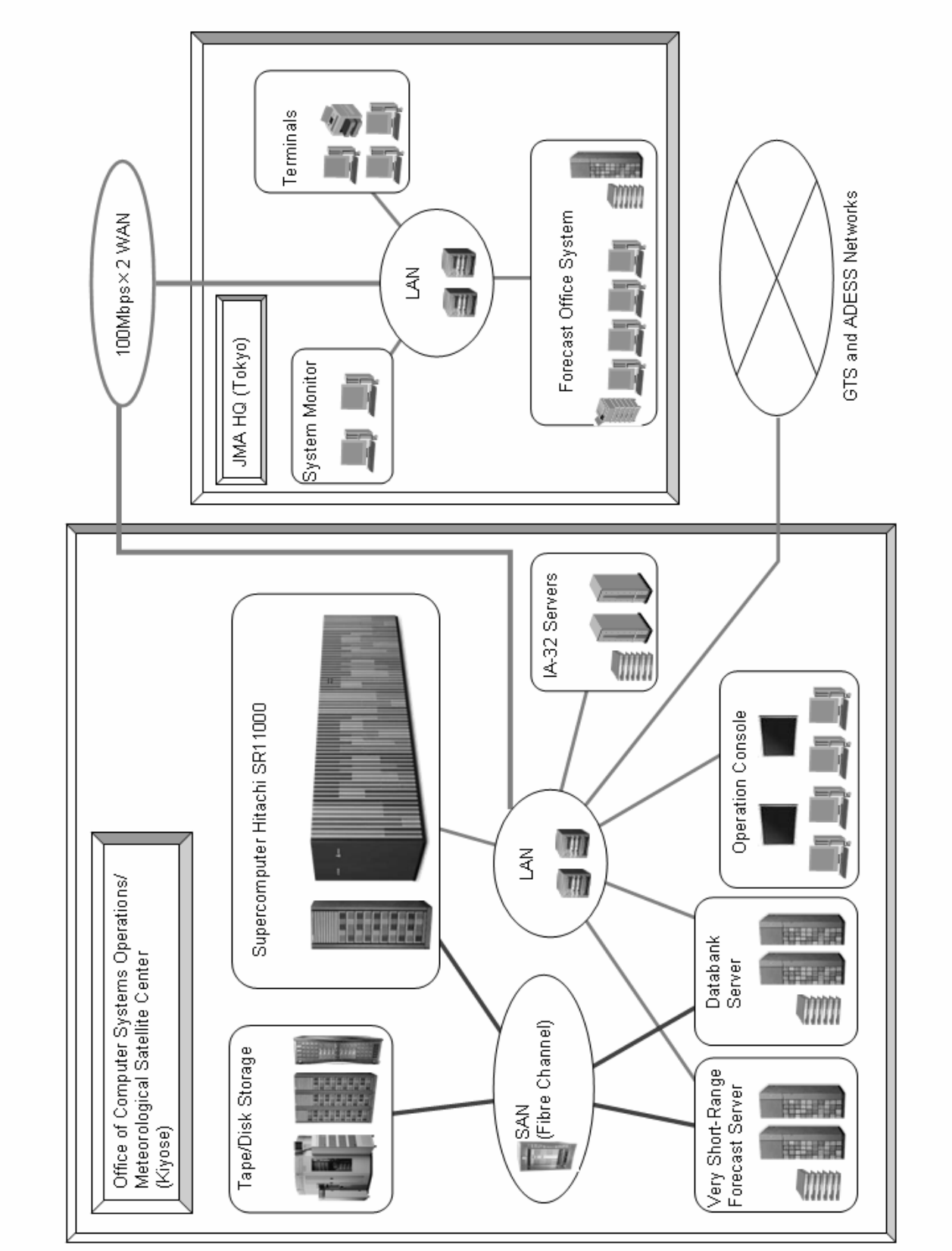

Figure 2.2.1 Schematic illustration of the computer system.