CHAPTER 7

Computing, Network, and Data Services

7.1 Computing services

CDC operates a centralized computing facility, based on high-end Unix workstations, that emphasizes shared resources and is designed for the benefit of all CDC science projects. Resource allocation and policy issues are dealt with by an internal CDC review group, the Computer Users Advisory Committee, which makes recommendations to CDC's systems management. The goal of CDC's systems services is to provide near-state-of-the-art computational and storage facilities. The purpose is to enable CDC to efficiently fulfill its mission and research obligations and to allow our scientists to remain competitive with their peers at other institutions.

At the high end, CDC is unable to internally support traditional supercomputing services. CDC researchers who require access to supercomputing or massively parallel processor (MPP) computing must seek those resources on their own. Currently, various scientists at CDC are making use of free supercomputing time on the NCEP Cray, the NCAR Cray, and the Alaska CIFAR Cray. CDC also has access to FSL's Intel Paragon MPP, but so far there has been no actual usage of this system due to computer programming impediments.

The bulk of CDC's computer facility investment is in mid-range computing, utilizing a tightly integrated network of Unix workstations and servers (detailed in **Fig. 7.1**). Total capacity of this system is approximately 1.75-Gflops of aggregate throughput, with 250-Mflops peak symmetric multi-processor (SMP) throughput and 125-Mflops peak single processor throughput. Individual machines are often dedicated to specific functions to minimize system downtime and segregate competing demands. For instance, four of our smaller servers are dedicated to serving the users' home file systems to the other cpu's on the network. Another is dedicated to Web and email functions. Our two largest systems are also segregated for specific uses - one as a compute server and one as a data file server. About one-third of our total system-wide computing capacity (500-Mflops) is readily accessible to large batch jobs, such as diagnostic general circulation model runs.

Total on-line storage capacity includes 600-Gbytes of traditional disks and 400-Gbytes of optical disks on a single jukebox. Many of the traditional disks are "striped" to increase throughput or are "RAIDed" to increase reliability. Approximately one-third of our traditional disks (200-Gbytes), plus the opti-

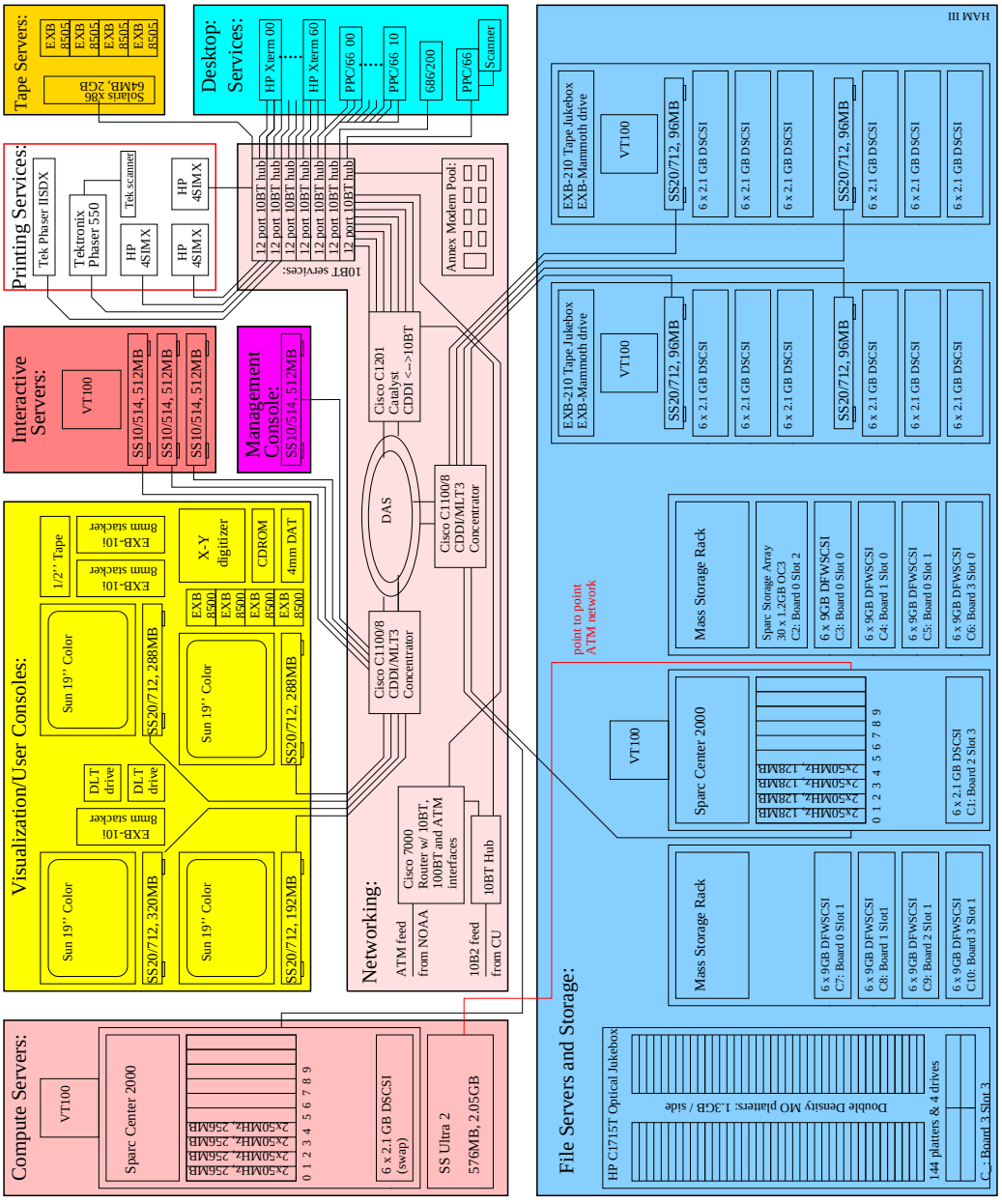

Fig. 7.1. Diagram of the primary mid-range computing facilities at CDC.

cal jukebox, are utilized for on-line storage of our high-demand, shared data sets, such as COADS Release 1 and portions of the NCEP Reanalysis. These data sets are generally available to both internal and external users.

For large personal data sets and for archival purposes, users have access to a variety of tape devices: twelve 8-mm Exabyte drives (including three stackers), two DLT 4000 drives, an IBM 3480 drive (primarily used to ingest Reanalysis tapes), a 4-mm DAT drive, a 1/2-inch open reel drive, and a 1/4-inch cartridge drive. In addition, systems staff has two 8-mm Mammoth tape jukeboxes dedicated to systems backups.

For hard copy output, CDC provides a variety of networked printing options. Three high-capacity 600-dpi duplex laser printers are located in the computer user rooms for the bulk of printing needs. In addition, there is a large-platen laser printer available for output up to 11”x17”, a 1200-dpi color laser printer, and a 300-dpi dye sublimation printer for high-end color prints. A high-resolution color scanner allows hard copy materials to be input to the system.

For low-end desktop computing needs, such as word processing, presentation graphics, and spreadsheets, CDC has moved away from the traditional PC or Macintosh in every office. Instead, the majority of CDC staff have a high-quality X-terminal on their desk. The X-terminals are used to access a pair of high-end Pentium servers which run a modified version of Windows NT and provide a variety of popular software packages. This arrangement provides a much better way for systems staff to monitor software utilization and significantly reduces the per capita cost of desktop maintenance and support. The X-terminals also serve as the primary gateway to our mid-range facilities and offer much the same functionality as if we provided a workstation console on every user’s desk.

7.2 Network services

CDC’s network facilities can be divided into the internal Local Area Network (LAN) and our connectivity to the Wide Area Network (WAN). Heavy reliance upon X-terminals and the distributed nature of CDC’s mid-range computer facilities makes dependable, high-bandwidth networking a crucial factor for maintaining user productivity. Access to remote supercomputing facilities and the growing importance of CDC on-line data sets and graphical Web products to external users dictate that we also have adequate connectivity to the Internet.

CDC is in the fortunate position of having considerable control over its internal LAN wiring. As a result, CDC moved to high-bandwidth, low-cost, twisted-pair media many years ago. Our current LAN topology is based on a star configuration using network concentrators. At the highest level, we operate a twisted-pair copper cable version of FDDI (known as CDDI) which interconnects all of our Unix workstations and servers. CDDI uses token ring technology to provide a full 100-Mbps of bandwidth without the collisions of traditional Ethernet. In addition, our fastest compute server is connected directly to the large data file server over a point-to-point ATM link at 155-Mbps. The

CDDI network ties into an eight-port Switched Ethernet hub, which, in turn, fans out to eight traditional 12-port Ethernet hubs that can connect up to 88 X-terminals, Macs, PCs, and printers into the network. Thus, every end user effectively shares a 10-Mbps line with ten other users or uses.

For connections to the outside world, CDC is fortunate to have two institutions, NOAA-Boulder and the University of Colorado (CU), to choose from to provide WAN services. CDC is connected into the NOAA-Boulder backbone with an ATM link (155-Mbps) and to the CU backbone at regular Ethernet speeds (10-Mbps). As one or the other of these institutions implement higher bandwidth connections, CDC can direct its network routing to take advantage of the fastest path to the external Internet. Currently, CU has an OC3 (155-Mbps) connection from its campus to an Internet Service Provider (ISP). NOAA-Boulder has a dual-T1 (3-Mbps) linkage that will shortly be upgraded to a fractional-T3 (6-Mbps). Although this may appear to heavily favor the CU route, many factors can affect usable bandwidth, such as the number of simultaneous users and the bandwidth by which the ISP itself connects to the Internet. See **Fig. 7.2** for more details.

7.3 Data services

While many CDC scientists adequately manage their own data needs, there are many instances where shared data sets and shared data expertise make cooperation on data management issues highly desirable. To that end, CDC provides data management services to acquire,

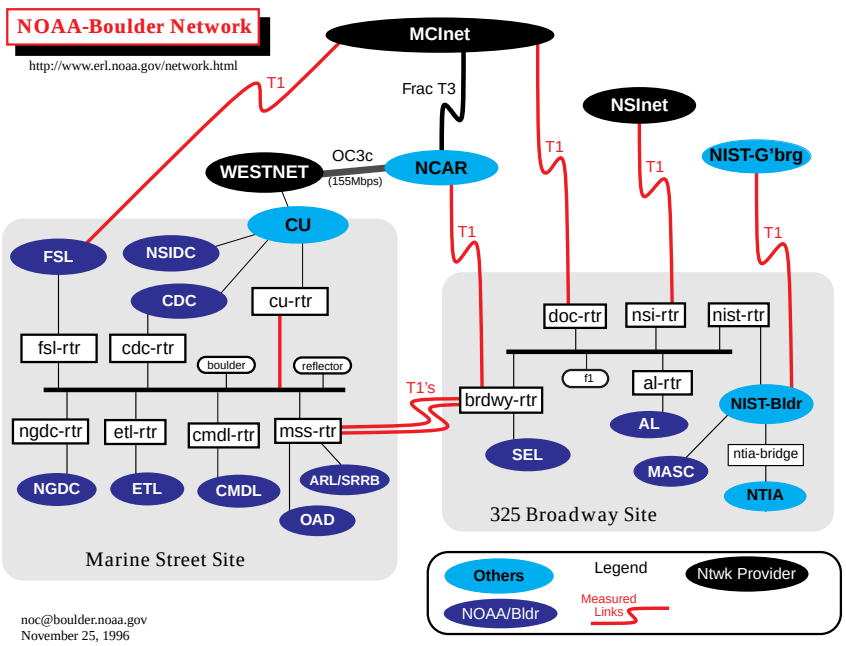

Fig. 7.2. Layout of the NOAA-Boulder Network, including alternative routes from CDC to the Internet through the University of Colorado (CU).

ingest, store, and maintain a wide variety of climate-related data sets at CDC. Most data sets, so maintained, are made available to our internal users as directly accessible files and to outside collaborators through anonymous FTP or tape copies. The use of the Web to facilitate data access is of growing importance at CDC. Advice on which data sets to support is provided to CDC data management staff by the CDC's Computer User Advisory Committee. Currently supported data sets are detailed in **Table 7.1**.

CDC data management has standardized much of its data work on the netCDF format. NetCDF was chosen because of its widespread use in the atmospheric sciences, especially in academia, and because its files are self-describing and machine-independent. Beyond that, however, CDC has cooperated with data managers at PMEL and NCDC to further refine netCDF for use with gridded climate data sets. The result has been the Cooperative Ocean-Atmosphere Research Data Service (COARDS) convention that defines, in more detail than plain netCDF, the ordering and formatting of key variables, a data-packing algorithm, and the minimum required metadata.

Use of the COARDS netCDF convention has allowed CDC and cooperating institutions to develop data sets and data access routines which are more easily exchanged. For example, CDC data management has developed COARDS-compliant access routines for GrADS and IDL, while PMEL, in Seattle, has developed similar routines for MATLAB and their FERRET software package. CDC also provides Fortran-callable access routines so our internal users can interface their own programs with our on-line data archive.

CDC also has made a major commitment to make climate information and products available through the Web. The public portions of our on-line archive can be searched, previewed, and downloaded via either the official NOAAServer or our own search interface. Requests that would constitute an unacceptable imposition on our network bandwidth are instead queued as a tape job. We only request that the external user provide us with replacement tapes. Other Web services, more graphical in nature, are also popular with both internal and external users. The US Climate Page, developed by CDC, has drawn considerable attention from students, travelers, gardeners, recreationalists, and real estate agents. The CDC Map Room pages, which illustrate both traditional and experimental climate diagnostic techniques, have been praised by scientific users everywhere from NCEP to CSIRO. The on-line Reanalysis Atlas is a new Web product designed to simplify user access to any desired combination of variable, time, latitude, longitude, and level within this cumbersome data set.

Total usage of these services are tracked to anticipate adequate allocation of resources to pace user demand. Currently, user "hits" on CDC's various Web pages exceed the third-of-a-million mark per month, corresponding to an average of one hit every ten seconds, day and night, throughout the month. Data set transfers into and out of CDC using anonymous FTP (the default for Web-downloads) are in the range of 40-

Gbytes per month. Some 405 users from around the world have retrieved over 7,000 files from our Reanalysis data archive alone. **Fig. 7.3** and **Fig. 7.4** illustrate some of these trends.

Table 7.1. Summary of Cooperatively Managed Data Sets at CDC.

Data Set Title	Size (GBytes)	Media
Climate Diagnostics Data Base	1.70	Magnetic
CDC Outgoing Longwave Radiation	0.23	Magnetic
Comprehensive Ocean-Atmosphere Data Set (COADS)		
Release 1 Products (COADS1).	0.12	Magnetic
Interim Products (COADS Interim).	0.12	Magnetic
Release 1a Standard Products (COADS1a Standard)	0.62	Magnetic
Release 1a Enhanced Products (COADS1a Enhanced)	0.62	Magnetic
Release 1b Standard Products (COADS1b Standard)	0.67	Magnetic
Release 1b Enhanced Products (COADS1b Enhanced)	0.67	Magnetic
DOE Gridded Surface Precipitation and Temperature Anomalies	0.05	Magnetic
ECMWF (non-public)	1.50	Magnetic
GFDL Consortium (non-public)	1.89	Magnetic
Leetmaa Pacific Ocean Analysis Data.	0.79	Magnetic
Monterey Marine Real-time Marine Data.	0.33	Magnetic
NMC Daily Global Analyses.	11.00	Magnetic
NMC Real-time Marine Data	0.09	Magnetic
NCEP/NCAR Reanalysis Products (all 4x daily)		
Other Flux	15.29	Magnetic
Pressure Level	71.21	Optical
Surface	8.16	Magnetic
Surface Flux	53.50	Optical
Tropopause	1.48	Magnetic
Spectral Coefficients	32.02	Magnetic
CDC Derived NCEP/NCAR Reanalysis Products		
Pressure Level (daily average)	17.75	Magnetic
Surface Flux (daily average)	13.41	Magnetic
Other Derived Products (monthly and long-term means)	2.45	Magnetic
Microwave Sounding Unit (MSU) Data	0.60	Magnetic
Reconstructed Reynolds SST.	0.02	Magnetic
Reynolds Sea Surface Temperature	0.12	Magnetic
Total	**234.71**	

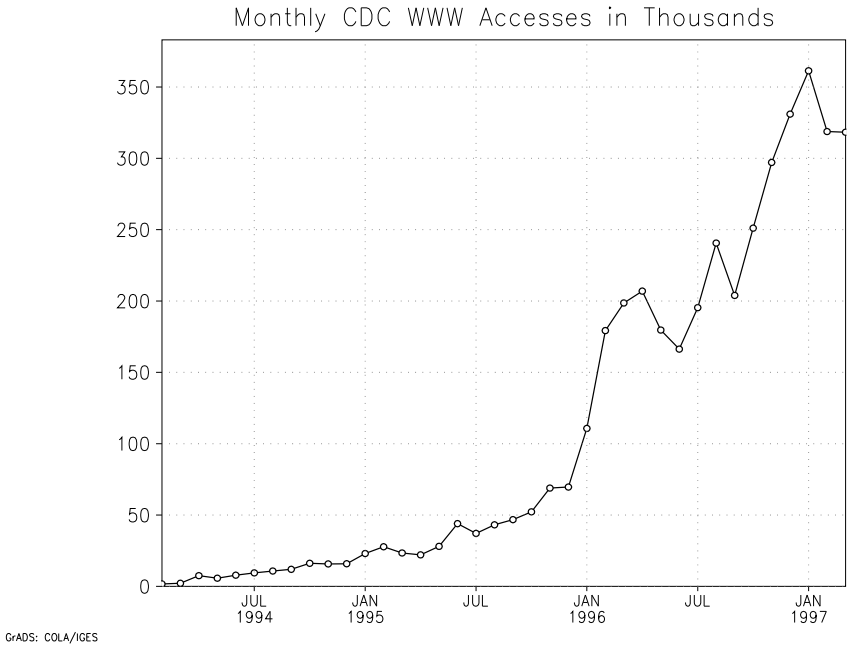

Fig. 7.3. Total number of "hits" per month for all of CDC's Web pages.

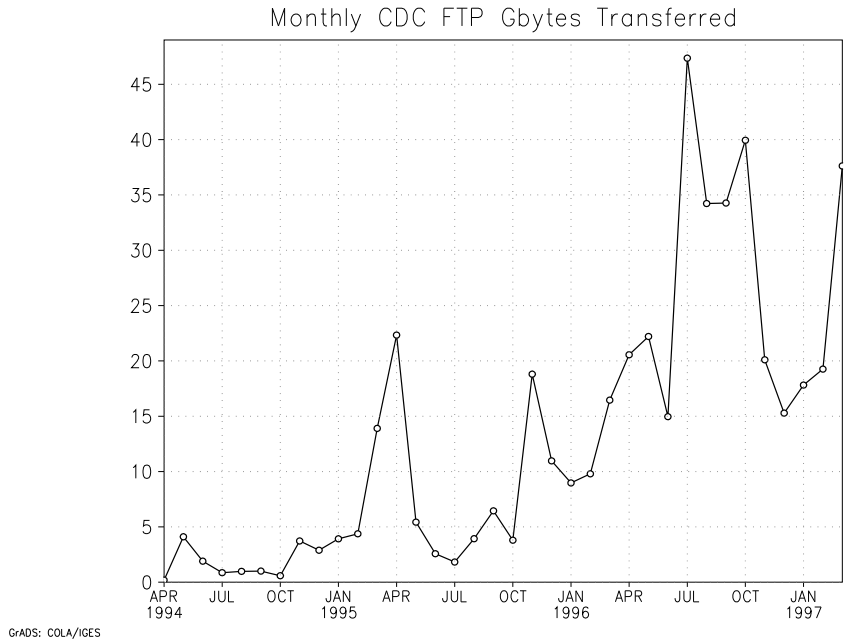

Fig. 7.4. Total gigabytes per month of files transferred using FTP at CDC.

This page intentionally left blank.