CHAPTER 7

Computing, Network, Web, 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. The goal of CDC's systems services is to provide near state-of-the-art computational and storage facilities and to accomplish this within a target budget that is 15% of CDC's total budget. The purpose of CDC's systems services is to allow CDC to efficiently fulfill its mission and research obligations and to enable CDC scientists to compete effectively with their peers at other institutions. Resource allocation and policy issues are considered by an internal CDC review group, the Computer Users Advisory Committee (CUAC), that makes recommendations to CDC's systems and upper management.

The bulk of CDC's computer facility investment is in mid-range computing, utilizing a tightly integrated network of Unix workstations and servers (**Figs. 7.1 and 7.2**). Total capacity of this system is approximately 18-Gflops of aggregate throughput, with 2.4-Gflops peak symmetric multi-processor (SMP) through-

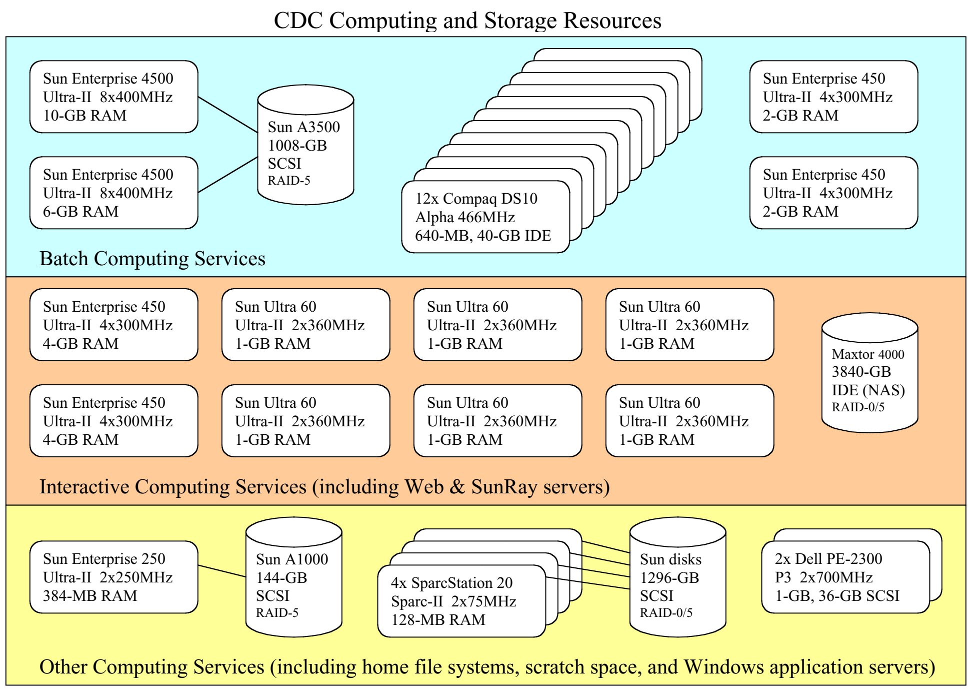

Fig. 7.1. Schematic overview of CDC's computing services and resources.

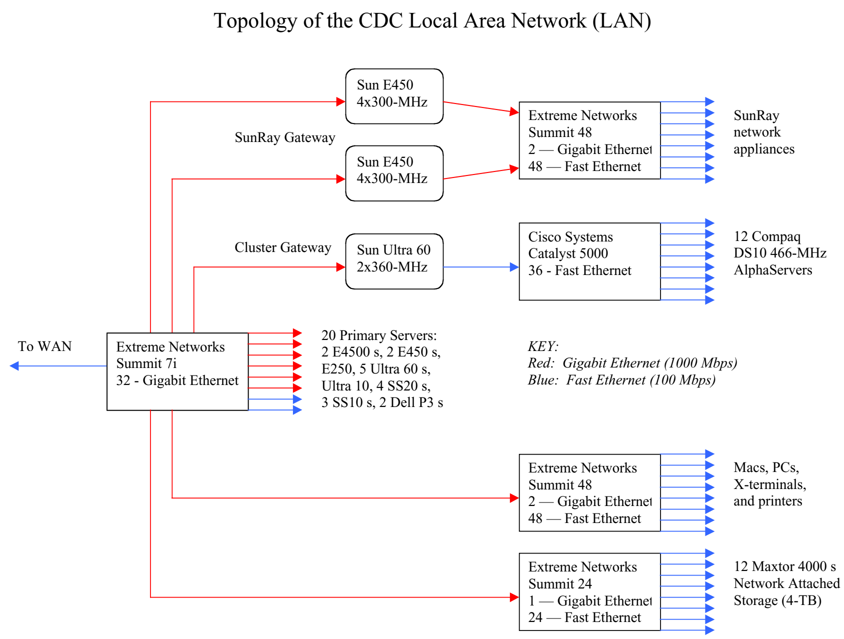

Fig. 7.2. Schematic depiction of CDC's local area network.

put on each Sun Enterprise 4500 and 525-Mflops peak single-processor throughput on each node of our Compaq DS10 Alpha-cluster. Comparable figures for CDC from four years ago are 1.75-Gflops aggregate, 250-Mflops SMP peak, and 125-Mflops single processor peak, respectively. Total aggregate and SMP throughput have increased by an order of magnitude. **Table 7.1** shows the breakout of CDC computing power (in units of aggregate DP LINPACK) by type of processor. Note that one Sun Enterprise 4500 is configured as a large memory machine with 10-GB of RAM.

Industry trends are allowing CDC to increasingly provide traditional supercomputing-class services in-house using commodity workstation technologies. In 1999, when a computer-room fire disabled the National Weather Service's CRAY C-90 supercomputer used for operational forecasts, CDC was able to implement a replacement for the week-two ensemble model runs and allowed NWS to continue issuing short-term climate forecasts until their own systems were repaired. CDC accomplished this using six dual-processor Sun Ultra 60 computers dedicated to the task. Lately, several CDC researchers have brought their compute-intensive modeling activities in-house, as external supercomputer facilities have become less cost effective, outmoded, or have switched to massively parallel processor (MPP) technologies that may require significant recoding (e.g., the National Center for Environmental Prediction's IBM RISC-cluster and the Forecast Systems Laboratory's HPTI Alpha-cluster). Since MPP facilities often employ large numbers of workstation-class processors, it is fairly easy to create small clusters of these same processors that come close to the per-processor performance of their larger brethren, at a fraction of the cost. This

Table 7.1: CDC Primary Computing Resources

CONFIGURATION	CPUs	LINPACK
Compaq Alpha DS10 (12 nodes)	12 x 466-MHz	6.30-Gflops
Sun Enterprise 4500 (2 nodes)	16 x 400-MHz	4.80-Gflops
Sun Ultra 60 (6 nodes)	12 x 360-MHz	3.25-Gflops
Sun Enterprise 450 (4 nodes)	16 x 300-MHz	3.60-Gflops
TOTAL		**17.95-Gflops**

economy is especially true if the computing can be done as a "loose cluster" which uses traditional Ethernet LAN technology for the inter-processor communications. Ensemble model runs are well suited to loose clusters, since each invocation of the model can be run independently on its own processor. Thus, in the past year CDC has concentrated its modest financial resources on providing this type of high-end computing. For example, for typical ensemble model runs, total throughput on our 12-node Compaq DS10 Alpha-cluster is close to 6 Gflops, with a total acquisition cost of only $60,000. For high-end computing needs that cannot easily be met within CDC, some users continue to make some use of outside facilities, primarily at NCEP, FSL, and NCAR.

Several smaller machines, not included in the general computing category, are dedicated to specific functions to minimize system downtime, segregate competing demands, and to maximize system security. For instance, various servers specialize in electronic mail, anonymous-FTP (file transfer protocol), NIS (network information services), DNS (domain name service), tape backups, and NFS (network file system) mounting of the users' home file systems. Two of the general-purpose Ultra 60 servers also act as fully redundant hosts for the CDC web site.

Total on-line (raw) disk storage capacity is approximately 12.5 TB, of which nearly 6 TB is a Dell storage area network (SAN) device that was acquired as excess property from the 2000 Census, but is still in the process of being brought on-line at CDC. The Dell SAN will be used to host climate model output for ease of model intercomparisons and for efficient generation of model diagnostic statistics. CDC's next largest category of disk storage, at 3.8 TB, is composed of newly acquired Maxtor network attached storage (NAS) devices. Each device is basically a stripped down PC with four large disks in a 1U (1.75" high) rack-mount form factor. These NAS devices can be procured for $3200 for 320 GB or approximately a penny per megabyte.

Although they do not provide fast throughput, they are a convenient alternative to magnetic tape. In 1997, total CDC disk capacity was 600 GB of traditional magnetic disks and 400 GB of fairly slow magneto-optical disks. A breakdown of CDC's current magnetic disk storage is shown in **Table 7.2**.

For large data sets that are of interest to only one or two PI's or for archival purposes, users have access to a variety of tape devices: eight DLT-7000 drives, including two auto-loading stackers, six 8-mm Exabyte drives with one stacker, and a 4-mm DAT drive. An IBM-style 4380 drive and stacker is also available to support the occasional ingest of 1/2" square-cartridge tapes. In addition, systems staff uses two Exabyte Mammoth-2 20-slot tape jukeboxes, dedicated to systems backups, with a total single-pass backup capacity of 2.4 TB. All of the users' home file systems, containing their programming code and executables, are incrementally backed up nightly, while the users' data files are backed up approximately once a month. Since 1997, CDC has discontinued support for 1/2" open-reel tapes and 1/4" cartridge QIC tapes.

Systems users have a variety of printing options through the network, including standard monochrome laser printers, both solid-ink and laser color printers, a large-format color ink-jet poster plotter, and a 65-ppm Xerox digital copier that can also staple the output. Color flatbed scanners with optical character recognition software, a very large digitizing tablet, and a bar-code scanner are available as alternative input devices.

There are three solutions provided for desktop computing, depending on the user's needs. CDC's administrative officer and secretaries use PCs that have specialized software for budgetary, personnel, and procurement record keeping. If scientists or their support staff have a strong predisposition to the

Table 7.2: CDC Primary Disk Storage

TYPE	DISKS	CAPACITY
Dell SAN (RAID5)	160 x 36 GB	5760 GB
Maxtor NAS (IDE)	48 x 80 GB	3840 GB
Sun A3500 (RAID5)	56 x 18 GB	1008 GB
Sun A1000 (RAID5)	8 x 18 GB	144 GB
Sun 6-PACKS	36 x 18 GB	648 GB
Sun 6-Packs	72 x 9 GB	648 GB
Alpha-cluster (IDE)	12 x 40 GB	480 GB
TOTAL		**12528 GB**

MacOS look-and-feel, they are provided with Apple Macintosh computers and Citrix client software to access the other CDC systems. Most users, however, are provided with SunRay network appliances. The SunRay has the look-and-feel of a traditional Sun workstation console and can be set up to support a variety of window environments, such as CDE or FVWM, much like an X-terminal. In reality, the SunRay is a further abstraction of the X-terminal concept. Instead of complex X11 calls being transmitted from a server to the user's desktop client, in the SunRay paradigm, only the frame buffer calls are transmitted from the server to the desktop. The hardware components in a SunRay box are those necessary to interface with user: video out, video in, audio out, audio in, keyboard and mouse. These components have been engineered to the limits of human perception, so, in theory, a SunRay will never need to be upgraded. The SunRay costs $350 and takes only minutes to replace if one should fail. A "smart-card" capability allows users to transfer their active session from one SunRay device to another for the purpose of collaboration or presentation. For all of our users, whether on a PC, Mac, or SunRay, CDC provides high-resolution LCD flat-panel displays to reduce eye-strain, glare, energy consumption, and hazardous emissions.

To supply typical office productivity software, CDC operates two dual-processor 700-MHz Pentium-III PCs, running the Windows 2000 Terminal Server Edition (TSE) operating system. Via Citrix, these servers provide access to Microsoft Word, Excel, PowerPoint, WordPerfect, Quatro Pro, Corel Presentations, Systat, Netscape, and Internet Explorer to our SunRay and Macintosh users. All of the Sun systems are running the Solaris 7 or Solaris 8 Unix variants. These Sun operating systems are scrupulously updated with the latest security patches to provide a high level of system integrity. All unnecessary network services are disabled and an automated security check is routinely performed to detect unauthorized attempts to gain access to the CDC systems. Software provided on the Sun systems includes FORTRAN (77, 90, and 95), C, and C++ compilers, IMSL, NAG, IDL, Matlab, Splus, Maple, NCAR Graphics, and GrADS. Other, publicly licensed software (freeware and shareware) available to users on the Sun systems totals 15 GB. The Compaq Alpha cluster runs the Linux operating system, which is considered a potential security weakness. To minimize the risk, CDC has placed the Alpha cluster behind a Sun Ultra 60 running Solaris, which serves as a gateway to separate the cluster from the rest of the CDC local area network. A user must successfully log into the Solaris box to "see" the Linux boxes. The Alpha cluster provides minimal software for the user - only a single FORTRAN compiler. The Alpha cluster and the high-end Sun systems are limited to batch-only jobs. To access the batch machines the user submits their job to the PBS scheduler, which routes the job to the most appropriate computing resource. Once a job starts running on a given processor, it generally will run to completion without swapping. This technique maximizes total system throughput by reducing the system overhead for each job run.

System downtime at CDC has been fairly minimal in the last few years. Redundancy has been designed into much of the systems infrastructure (e.g., two SunRay servers, two web servers, two Enterprise 4500's, two Pentium servers, and disk arrays using RAID-5). If something breaks, users can usually still get work done. Preventative maintenance and hardware/software upgrades are scheduled and performed by CDC systems staff one evening per month. All critical hardware systems are on maintenance contracts with the manufacturers. CDC also keeps a supply of "hot spares" for many commodity hardware components.

7.2 Network Services

CDC's network facilities can be divided into the internal Local Area Network (LAN) and the connectivity to the Wide Area Network (WAN). Heavy reliance upon centralized computing resources, with minimal computing power on the individual's desktop, requires dependable, high-bandwidth networking to maintain user productivity. Access to remote computer systems at NCEP, NCAR, and GFDL and the importance of CDC on-line data sets and graphical web products to external users dictate that CDC also requires adequate connectivity to the Internet.

CDC is in the fortunate position of having considerable control over its internal LAN wiring. Although LAN equipment was provided to CDC as part of the move to the new David Skaggs Research Center (DSRC) in March 1999, it quickly became apparent that a mixture of Asynchronous Transfer Mode (ATM) and Ethernet protocols within the LAN and the use of edge switches in the role of core switches were not providing CDC the level of service needed. Consequently, with the advent of inexpensive copper-based Gigabit Ethernet technology, CDC completely replaced their LAN in 2000 for approximately $45,000. The new LAN (**Fig. 7.2**) features an Extreme Networks Summit 7i Gigabit Ethernet switch at its heart. The switch provides 28 1000BT (copper wire) ports and four 1000SX (fiber cable) ports. The 7i's backplane can handle all active ports at full wire speed. Most of CDC's Sun servers are connected directly into the 7i to provide high-bandwidth for the NFS (network file system) traffic between the user's running job and the user's data files. Other Gigabit links go to Fast Ethernet concentrators, the Extreme Networks Summit 24 and a Summit 48. The Summit 48 provides 48 Fast Ethernet ports for the desktop PC's, Mac's, and printers. The Summit 24 connects to our 12 Maxtor NAS devices, with plenty of room for expansion. Another Gigabit link from the 7i goes to a Sun Ultra 60 that serves as the gateway to our Compaq Alpha-cluster. This Sun gateway serves to shield the 7i from the Linux operating system on the Alphas. The Alphas are connected to the Ultra 60 using a Cisco Catalyst 5000, our only holdover from our former location on the CU east campus. Finally, two Gigabit links go from the 7i to the two SunRay servers. The SunRay architecture requires a private LAN for communication between the SunRay server and the SunRay desktop clients. CDC utilizes an Extreme Networks Summit 48 for this purpose. The Summit 48 has been configured so that if one of the SunRay servers fails, the SunRay users are still able to log on to the other SunRay server. In 1997, each office user had access to an Ethernet (10-Mbps) line shared with nine other users or devices. Today, each office user has a dedicated Fast Ethernet (100-Mbps) line.

For connections to the outside world, CDC takes advantage of the excellent connectivity provided by the DSRC's ATM backbone and the Boulder Research and Administrative Network (BRAN), a fiber-based metropolitan-area network that CDC helped create (see **Fig. 7.3**). The DSRC ATM backbone provides OC12 (622-Mbps) links between CDC and the other building tenants, including the Forecast Systems Laboratory (FSL), the Aeronomy Laboratory (AL), and the National Geophysical Data Center (NGDC). An OC3 (155-Mbps) link via BRAN connects the DSRC backbone both to the University of Colorado-Boulder and to NCAR. Both of these institutions are connected at OC3 speeds to the Front Range GigaPOP (FRGP), which provides connectivity to the commercial Internet via Qwest and to the Next Generation Internet (NGI) in the form of Abilene. If problems develop at the FRGP, the DSRC backbone can switch to a secondary route though FSL, albeit at fractional T3 (6-Mbps) speeds. A bottleneck for CDC is caused by the connection from the CDC Summit 7i to the DSRC ATM backbone. At present this is only a Fast Ethernet (100-Mbps) connection. Despite this, CDC, which represents only 5% of the DSRC's occupants, has accounted for 28% of the total network traffic through the DSRC's gateway to the outside world over the past

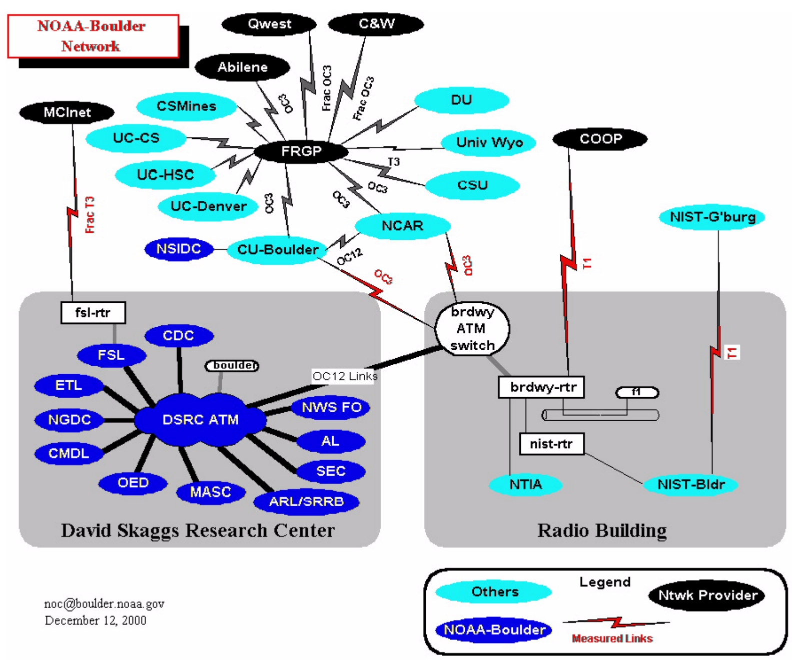

Fig. 7.3. Layout of the NOAA-Boulder Network.

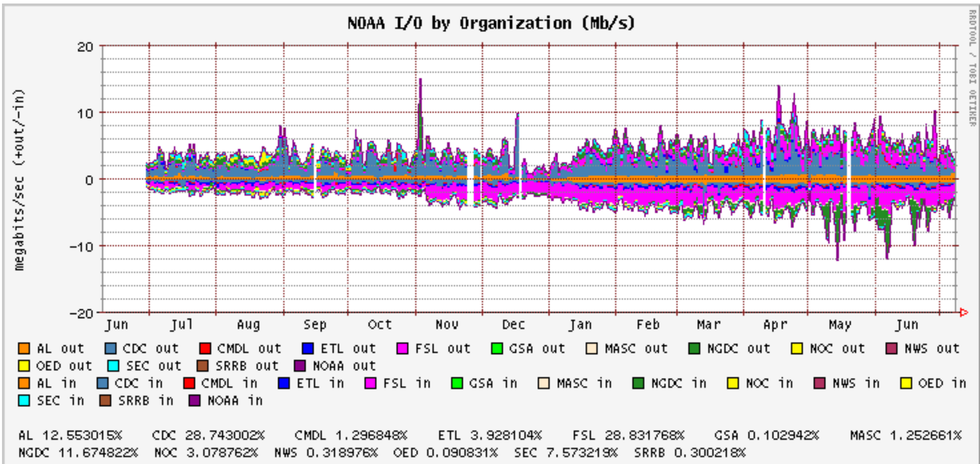

Fig. 7.4. Network traffic into and out of the DSRC over the past year.

year (**Fig. 7.4**). Most of this traffic has been outbound, meaning transfer of information from CDC to the rest of the scientific community and to the general public. Only FSL can equal CDC's network traffic through the DSRC gateway, but most of their traffic has been inbound on that link. In the near future, the DSRC network operations center is planning to upgrade our 100-Mbps bottleneck to a 1000-Mbps Gigabit Ethernet connection.

7.3 Web Services

CDC's systems and data management groups provide the infrastructure and support for CDC's primary and subsidiary web sites. The primary web site (http://www.cdc.noaa.gov/) is generated using a set of document templates which result in a consistent site look-and-feel and ensure compliance with Federal policies and regulations. A recent example of a Federal policy is the prohibition on the use of "cookies," which deposit information on the end user's computer. An example of a regulation is the Federal Acquisition Regulation which implements Section 508 of the Rehabilitation Act of 1973. Currently, the CDC webmaster is responsible for tracking these various directives and translating them into meaningful courses of action. Our document structure provides a central point of change if the document layout and navigation schemes need to be modified to comply with changing policies and requirements. This allows local scientists to concentrate on content generation, and not, for example, on the need to add links to "disclaimer" and "privacy" statements. These high-level links are included in the document templates and hence automatically included in any document which is generated using the templates.

The CDC webmaster, and the broader data management group, are also tasked with staying abreast of technology trends in the rapidly evolving area of web site architecture and management. CDC actively participates in and often helps organize opportunities for web technology exchange, including the OAR Technical Committee on Computing Resources (TCCR), OAR Web Workshops, NOAA Web Workshops, NOAA Tech Workshops, the Federal Webmasters' Conference series, WWW Consortium meetings, and the American Meteorological Society (AMS) Interactive Information and Processing Systems (IIPS) Conferences. CDC has also been a leader within NOAA in the investigation and, if warranted, local deployment of Java, JavaScript, Java Server Pages (JSP), XML, webcasts, and digital certificates. The CDC web site underwent a major makeover this year to improve site navigation and to incorporate better support for some of these new technologies.

Subsidiary web sites are those which are largely autonomous sites hosted on the CDC web server. They serve a broader institutional purpose but are maintained by CDC staff. These sites include the COADS web site (http://www.cdc.noaa.gov/coads/), the Metropolitan Denver Combined Federal Campaign (CFC) web site (http://www.cdc.noaa.gov/cfc/), the AMS Committee on IIPS (http://www.cdc.noaa/gov/iips/), and a NOAA-level site devoted to the NOAA Strategic Goal on "Seasonal to Interannual Climate Prediction" (http://www.cdc.noaa.gov/seasonal).

The success of the CDC web site(s) can be measured to some degree by tracking the number of files served, often referred to as "hits." Since its inception, the CDC web server has seen almost exponential growth in the number of documents

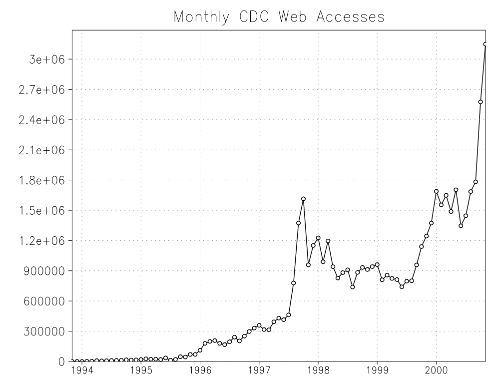

Fig. 7.5. Total number of "hits" per month for all of CDC's web pages.

served to outside users (**Fig. 7.5**). Current "hits" are in the range of two million per month, about six times the rate of four years ago.

7.4 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, ingest, store, and maintain a wide variety of climate-related data sets at CDC. Most data sets, so maintained, are made available to CDC's internal users as directly accessible files and to outside collaborators through anonymous-FTP or tape copies. Over the past four years, the CDC web site has become the primary method by which external users locate, browse, and download CDC's diverse data holdings. CDC's Computer Users Advisory Committee (CUAC) provides advice to the data management group on which data sets to include in the centrally managed on-line data archive. Currently supported data sets are detailed in **Table 7.3**.

CDC data management has standardized much of its data work in the netCDF format. NetCDF was chosen because of its widespread use in the atmospheric sci-

Table 7.3: Summary of Cooperatively Managed Data Sets at CDC

DATASET TITLE	SIZE (GB)
Climate Diagnostics Data Base	1.70
CPC Merged Analysis of Precipitation	0.15
CPC .25 x .25 Daily US Unified Precipitation	1.94
Comprehensive Ocean-Atmosphere Data Set (COADS)	62.85
DOE Gridded Surface Precipitation and Temperature Anomalies	0.05
DAI Palmer Drought Severity Index	0.05
ECMWF (non-public)	1.48
GFDL Consortium Derived Products	0.31
GFDL Consortium (non-public)	1.55
Global Sea-Ice and Sea Surface Temperature (non-public)	0.39
Hadley Sea-Ice and Sea Surface Temperature (non-public)	0.39
NOAA Interpolated OLR	0.22
Kaplan Sea Surface Temperature	0.01
Monterey Marine Real-time Marine Data	0.33
Microwave Sounding Unit (MSU) Data	0.61
NCEP Daily Global Analyses	16.09
NCEP GCM T42 MRF1 (non-public)	0.02
NCEP Pacific Ocean Analysis	1.66
NCEP Real-time Marine Data	0.23
NCEP/NCAR Reanalysis Products (4x daily)	393.45
CDC Derived NCEP/NCAR Reanalysis Daily Averages	80.94
CDC Derived NCEP/NCAR Reanalysis Monthly/Long-term Means	5.99
NCAR Daily Observed SLP daily data (non-public)	0.09
NOAA Highly Reflective Clouds	0.47
NODC World Ocean Atlas 1994	0.58
NODC World Ocean Atlas 1998	1.74
Reconstructed Reynolds SST	0.02
Reynolds Sea Surface Temperature	0.12
Northern Hem. EASE-Grid Weekly Snow Cover and Sea-Ice Extent	0.02
University of Delaware Precipitation and Air Temperature	0.58
TOTAL	**574.22**

ences, especially in academia, and because its files are self-describing and machine-independent. Beyond that, however, CDC has cooperated with data managers at the Pacific Marine Environmental Laboratory (PMEL) and the National Climatic Data Center (NCDC) to further refine netCDF for use with gridded climate data sets. The result has been the Cooperative Ocean-Atmosphere Research Data Standard (COARDS) convention that defines a metadata format, including required variables, variable attributes, and a data packing algorithm.

Use of the COARDS netCDF convention has allowed CDC and cooperating institutions to develop data sets and data access routines that are more easily exchanged. For example, CDC data management has developed COARDS-compliant access routines for GrADS and IDL, while PMEL has developed similar routines for MATLAB and their FERRET software package. In the past four years, the COARDS convention has become the de-facto standard for gridded climate data sets wherever netCDF is in common use. To comply with national policy directives, CDC also provides metadata for all of its data sets in the required FGDC format.

As can be seen elsewhere in this volume, CDC 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 a locally developed web-based search interface or through cooperative initiatives, such as ESDIM's NOAAServer, NESDIS' National Virtual Data System (NVDS), URI/UCAR's Distributed Oceanograhpic Data System (DODS), and PMEL's Live Access Server (LAS). In several cases, CDC provides the largest on-line data archive accessible through these services. CDC is also currently cooperating in the development and installation of a next-generation data service, called the GrADS-DODS server (GDS). The GDS will form the core of the new NOAA Operational Model Archive and Distribution System (NOMADS) that will provide enhanced access to climate model output to researchers. CDC is scheduled to come online as a NOMADS node later this year.

Total usage of these services is tracked to anticipate adequate allocation of resources to pace user demand (**Fig. 7.6**). Data set transfers into and out of CDC using anonymous-FTP (the default for Web downloads) are now in the range of 1000 GB per month, 25 times the rate of four years ago. The NCEP Reanalysis data set has been particularly popular. Since its inception in 1995, CDC's

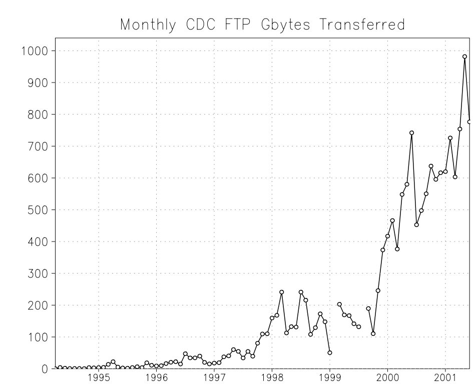

Fig. 7.6. Total gigabytes per month of files transferred using FTP at CDC.

NCEP Reanalysis web pages have received 2,157,250 “hits”, leading to 257,838 downloads that have totaled 11.1 Terabytes. In addition, a total of 360 8-mm and DLT tape orders, often involving multiple tapes, have been filled which total an additional 14.9 Terabytes. Only CDC’s netCDF version of NCEP Reanalysis is provided to external users. The original GRIB-format version of the Reanalysis can be acquired from NCDC or NCAR. The figures quoted above do not include CDC’s own internal usage of the Reanalysis data set, which is considerable.

7.5 Ensuring long-term observed climate records

CDC pioneered the development of the Comprehensive Ocean-Atmosphere Data Set (COADS)—an international resource for historical weather observations over the globe. The ability to detect any future changes in regional, hemispheric, and global climate, and to make improvements in global climate models, depends crucially on the availability of high-quality retrospective data and metadata. As such, COADS forms the central oceanic ingredient for past and present National Scientific Assessments and for the Intergovernmental Panel on Climate Change, as well as for Global Atmospheric Reanalyses by US groups and by the European Centre for Medium-Range Weather Forecasts.

In 1985, the cooperating organizations issued the first version (Release 1) of this comprehensive historical record of marine meteorological data—observations from ships now extending back for over two centuries, supplemented since the late 19th century by data from oceanographic research vessels, and in recent decades from moored and drifting environmental buoys—and made them widely available and readily accessible to users. Both observational and gridded monthly summary products, which offer an unparalleled level of archival and statistical detail, were offered to the research community. These products have been periodically updated and enhanced, with the period of record now covering 1784–1997. To date, the data have been the basis for more than 500 published papers (up to 60 a month in American Meteorological Society Journals for recent years).

In 2001, a major milestone for COADS was the release of higher quality data and products that extend from 1997 back to 1784. Our most recent efforts concentrated on the period prior to 1950, requiring exceptional efforts in data archaeology to locate and digitize millions of historical ship records from US and international archives that were never previously available in digital form. This also required extensive cooperation with a host of international partners, including China, Germany, India, Japan, the Netherlands, Russia, and the United Kingdom. Major input data sources for the pre-1950 period are detailed in **Table 7.4**.

Final observational and statistical products based on these and other data were completed in early 2001, comprising COADS Release 1c (1784–1949). Similar updates to more contemporary periods, Releases 1b (1950–1979) and 1a

Table 7.4: New Pre-1950 Data Sources for COADS

NAME	YEARS	# of OBS
Blend of UK Main Marine Data Bank (MDB) and COADS	1854–1949	12.1M
Maury Collection	1784–1863	1.3 M
Norwegian Logbook Collection	1867–1889	0.2 M
Japanese Kobe Collection	1890–1932	1.0 M
US Merchant Marine Collection	1912–1946	3.5 M
Russian Makarov Collection	1804–1891	3.5 K
World Ocean Database	1998	405 K
Arctic drift shifts and stations	various	16 K
Russian MARMET data (previously known as MORMET)	various	268 K

(1980–1997), were finished during 1996–1997. Thus upon completion of Release 1c, the original Release 1 (1854–1979) data were fully replaced by a consistent set of data and metadata products covering the extended period 1784–1997, with many additions of new and improved data. Web-based capabilities are now under development for subsetting in space and time of the basic 1784–1997 observations and statistics into easily used ASCII formats. When these new capabilities are finalized, COADS Release 2 will be considered complete.

Increases in the number of observations for the pre-1950 period due to new data are illustrated in **Fig. 7.7**. The two panels show the number of observations prior to 1950 previously available in COADS Release 1, plus the new observations recently added. The major contribution of the Maury Collection (left panel) towards extending the start of COADS prior to 1854 is clearly evident. It should be noted, however, that instrumental data are sparse prior to 1854 and most of these observations are wind estimates from sailing ships. Similarly, the right panel shows the importance of additions from the UK MDB, Kobe and US Merchant Marine Collections in the first half of the 20th century.

The increase in SLP observations between Release 1 and Release 1c is dramatic especially in the 19th century (**Fig. 7.8**). Much of the Release 1 data prior to about 1880 were from a “Dutch” source (possibly of international origin), in

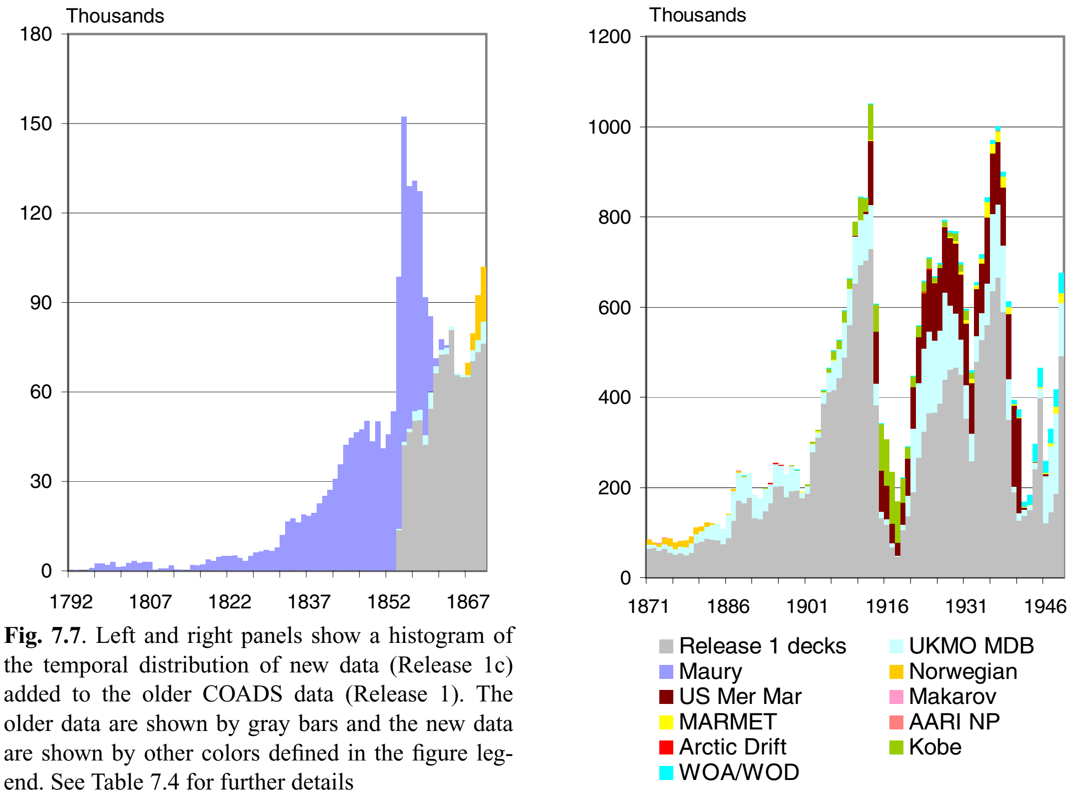

Fig. 7.7. Left and right panels show a histogram of the temporal distribution of new data (Release 1c) added to the older COADS data (Release 1). The older data are shown by gray bars and the new data are shown by other colors defined in the figure legend. See Table 7.4 for further details

which pressure was available among supplemental data but not adjusted for gravity. These SLP observations have been adjusted for gravity and made available for the first time in Release 1c. We believe that these and other important increases in the numbers of historic observations have significantly improved COADS coverage, providing a good basis for future planned improvements.

A "Workshop on Advances in the Use of Historical Marine Climate Data" will be hosted by CDC 18–21 September 2001 in Boulder, Colorado, USA. The workshop, organized by NOAA, the UK Met Office and the Japan Meteorological Agency, and sponsored by the Global Climate Observing System (GCOS) and WMO, will build on the blend of the US Comprehensive Ocean-Atmosphere Data Set (COADS), with the UK Met Office MDB, plus with newly digitized data in the US and from other international partners. The blended product, encompassing 1784–1997, will provide the climate research community with an unprecedented assembly of *in situ* marine data.

Contributed by: *J. Collins, C. Harrop, D. Hooper, S. Lubker, A. McColl, C. McColl, B. McInnes, D. Mock, R. Schweitzer, and S. Woodruff.*

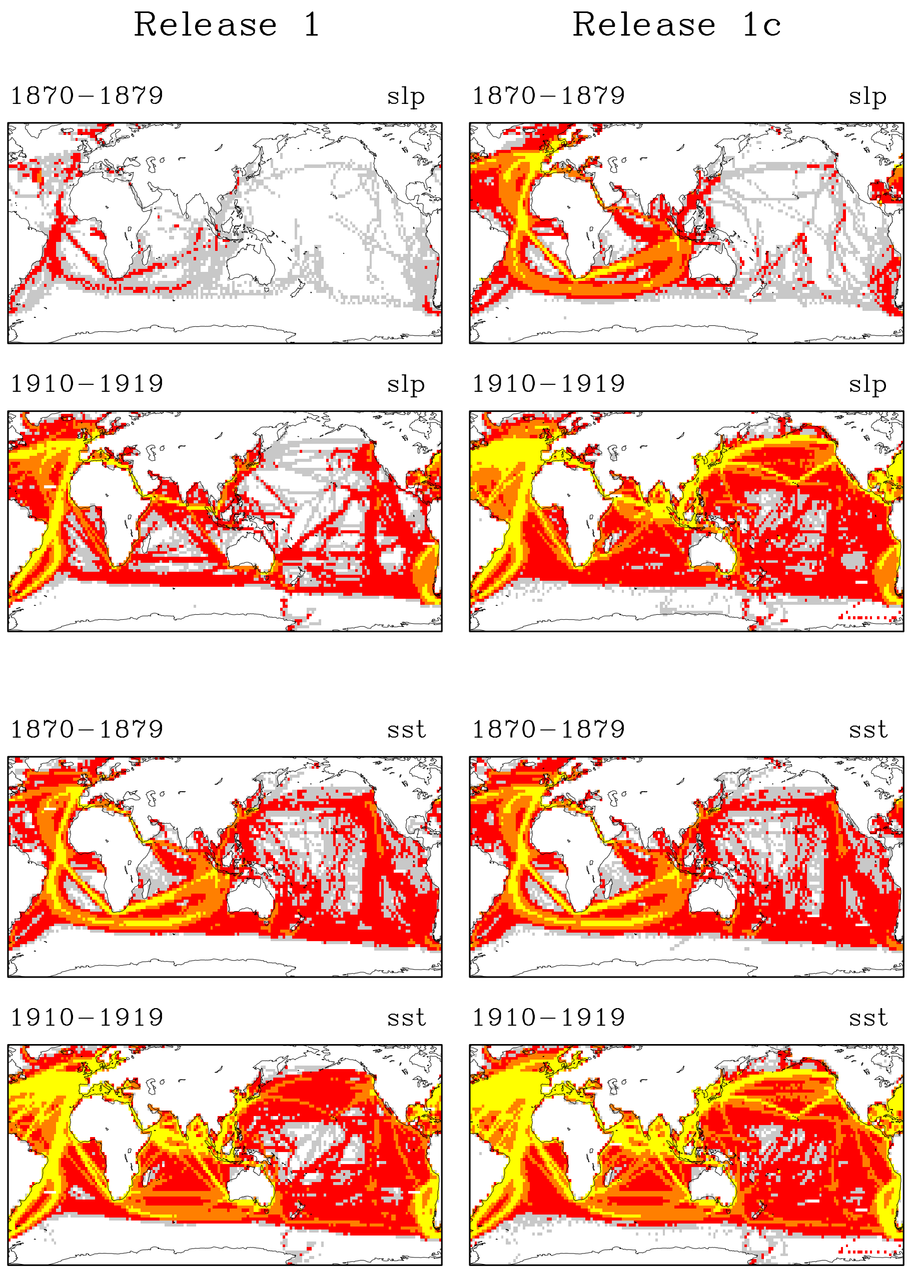

Fig. 7.8. Mapped comparison of decadal totals of SLP and SST observations (70°N–78°S; 68°W–68°W) for two decades (1870–1879 and 1910–1919) in COADS Release 1 and 1c. Older data available in COADS Release 1 are shown in the left column; existing plus new data added for Release 1c are shown on the right column. The colors correspond to the number of observations in each 2° box per decade: blue indicates 10–99, yellow indicates 100–399, orange indicates 400 or more.