



Data Management Components for a Research Data Archive

Steven Worley and Bob Dattore
Scientific Computing Division

Computational and Information Systems Laboratory

National Center for Atmospheric Research

E-mail: worley@ucar.edu

**Presented at the THIC Meeting at the National Center for Atmospheric
Research, 1850 Table Mesa Drive, Boulder CO 80305-5602**

July 19-20, 2005

THIC Inc.

The Premier Advanced Recording Technology Forum

1



NCAR

Data Management Components for a Research Data Archive



Steven Worley and Bob Dattore

Scientific Computing Division

Computational and Information Systems Laboratory

NCAR



Outline



- Research Data Archive (RDA) definition
- Components
 - MSS
 - Online Data Server - traditional service
 - Databases
 - Community Data Portal - evolving service
 - SAN
 - Media for I/O



Research Data Archive (RDA) definition



- Collection of reference datasets used in atmospheric and related sciences
- Over 600 datasets
- 10-20 new datasets added annually
- First established about 40 years ago
- Basic metrics
 - 548K files
 - 100.5 TB
 - 2-3K unique users annually



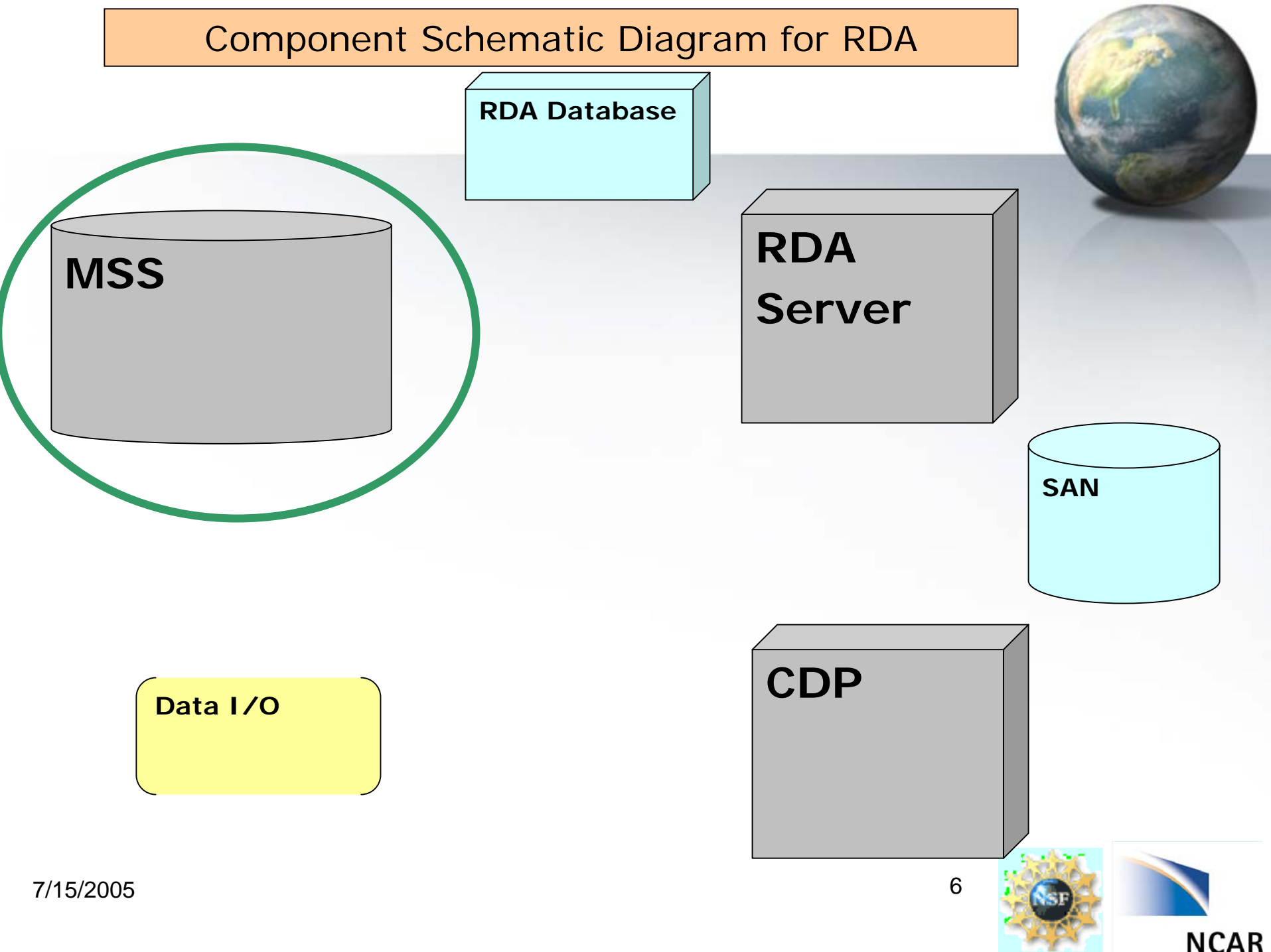
What makes a dataset?



- Elements of a dataset
 - Data files (1 ~ 20K)
 - Syntactic and semantic metadata
 - Publications
 - Documentation
 - Lineage
 - Data preparation, QC, analysis methods, etc



Component Schematic Diagram for RDA



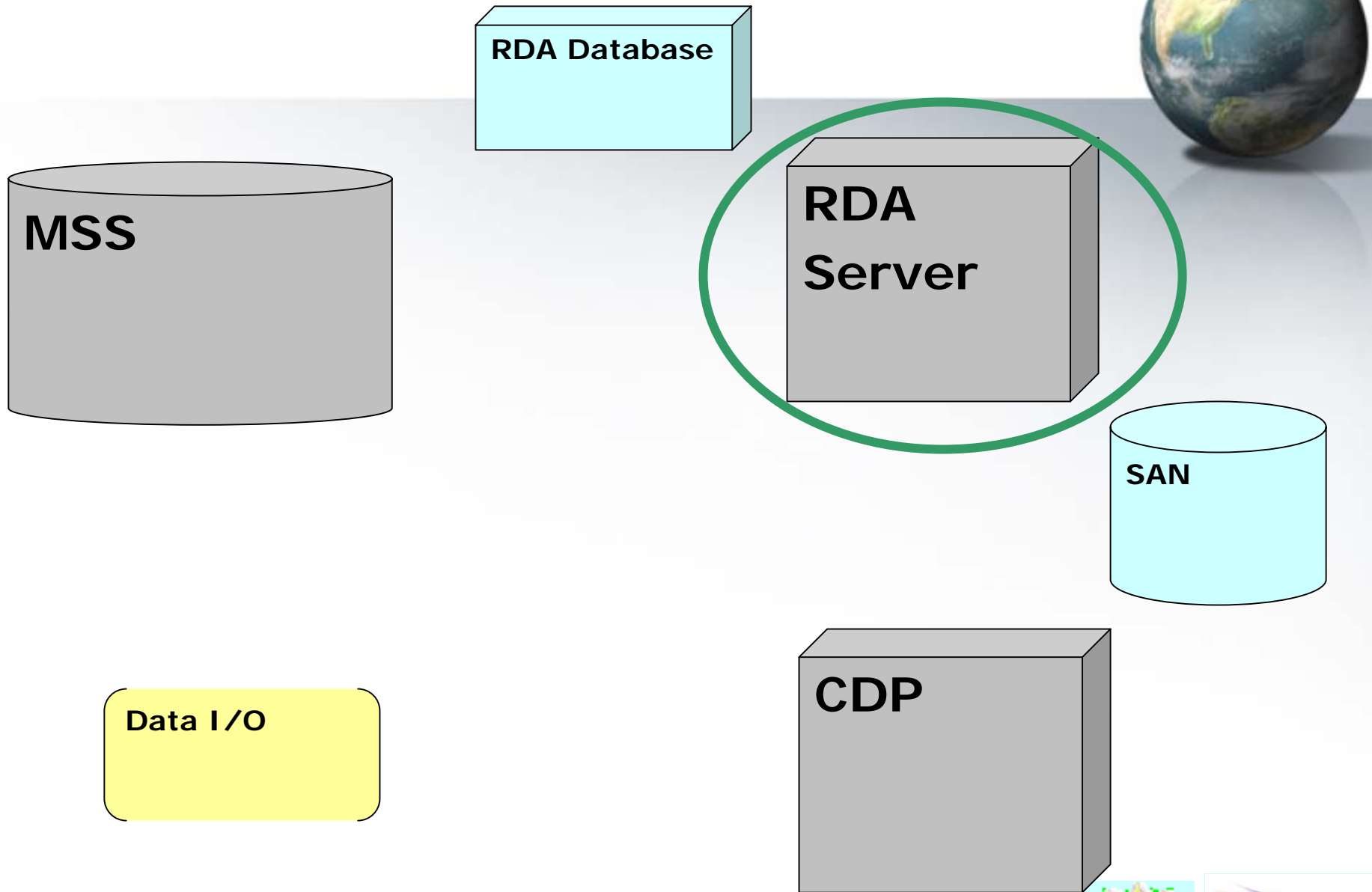
MSS



- Features
 - Archive for all data
 - Including backup files
 - Local users can access all data
 - Local = anyone with SCD computing account
 - Only need file name!
 - Usage logs are generated
 - When, what, who accessed the data



Component Schematic Diagram for RDA



Online Data Server - Traditional



- Features
 - Exclusive dedication to the RDA
 - Single point for all information
 - Project web pages and catalogues
 - Home web page for each dataset
 - General Description
 - MSS File Lists
 - Search/Discovery
 - Software
 - Documentation
 - Consultant contact
 - Most readily needed data, (~ 15TB)
 - FTP and Web access
 - User request forms for one-off data requests



Online Data Server - Traditional




EMAIL: PASSWORD: [Forgot Password?](#) [Data User Registration](#) [Do I have to register?](#)

UCAR | NCAR | SCD | **DSS** SEARCH | SITE MAP | INTERNAL | CONTACT US NCAR

FAQS GO TO DATASET:

HOME | SERVICES | **DS SOFTWARE** | INVENTORIES | MSS FILES | FREE DATA

DS HOME | DS D

 **ds540.0 HOME PAGE** [Help](#)

International Comprehensive Ocean-Atmosphere Data Set (ICOADS), Global Marine Surface Observations

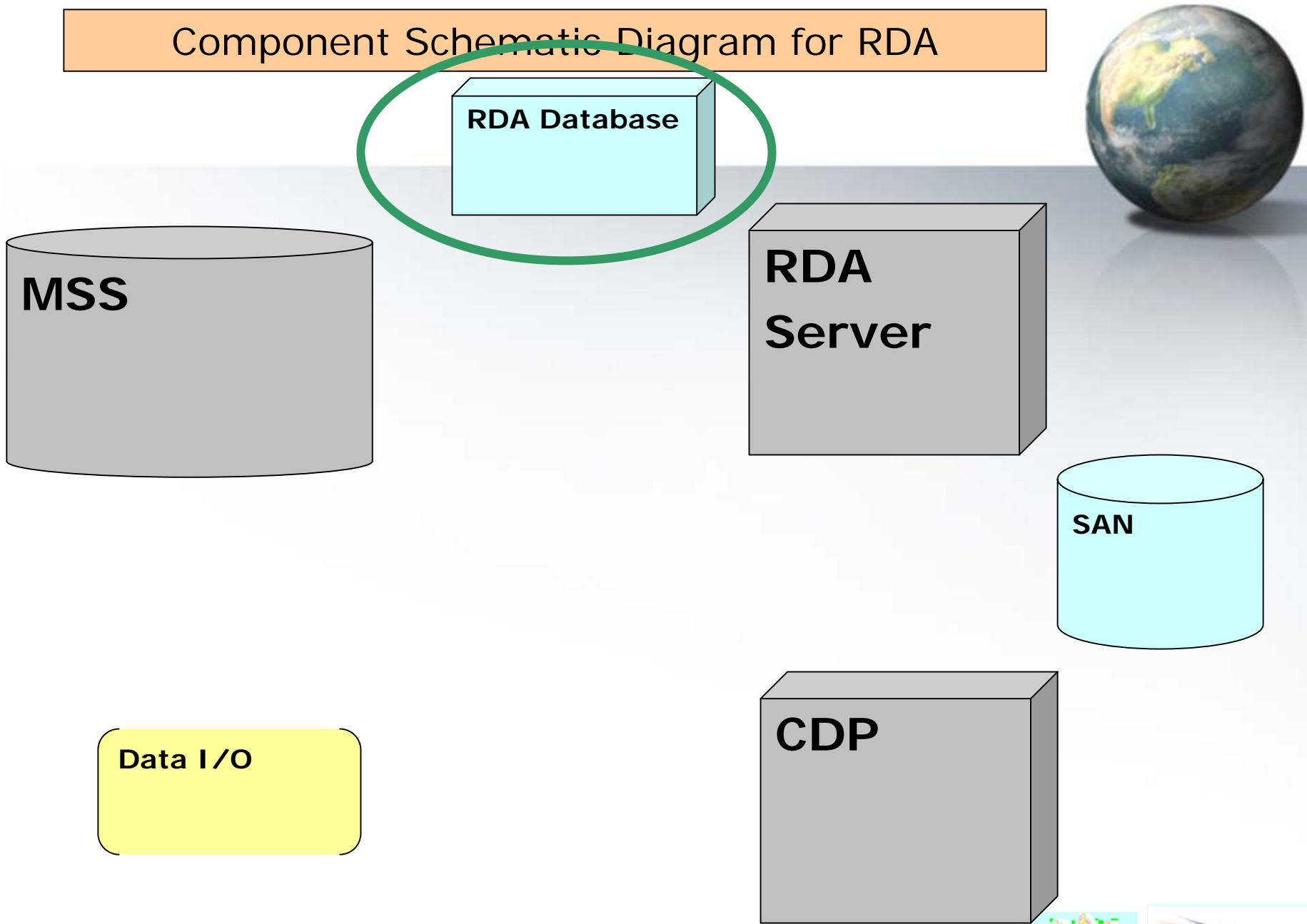
ICOADS is the world's largest collection of marine surface in situ observations with 185 million records for 1784 through 2002. The collection is formed from many international data sources. The records are merged together and originate from ships (merchant, navy, research), moored and drifting buoys, coastal stations, and other marine platforms. Each report contains individual observations of meteorological and oceanographic variables, such as sea surface and air temperatures, wind, pressure, humidity, and cloudiness. The coverage is global and sampling density varies depending on date and position relative to major ocean shipping routes.

ICOADS is the result of a long-standing (beginning in 1981) cooperative project among the NOAA Climate Diagnostics Center, NOAA National Climatic Data Center, and NSF National Center for Atmospheric Research. Systematic quality assurance and processing into a uniform data format make the dataset a significant and often cited archive.

In an important, related dataset, ICOADS observations are statistically summarized on a monthly basis and in one-degree or two-degree latitude by longitude squares. The [monthly summary statistics are also available](#), as is much more information on the [ICOADS project website](#)



Component Schematic Diagram for RDA



RDA Database



- RDA management tool
- Metadata server



RDA Database (Management tool)



Current Capabilities

Future Capabilities

DATA SOURCES

- SCD computer user account data
- MSS and Data Server file descriptions
- MSS and Data Server file usage logs
- RDA dataset - file relationships

DATA SOURCES

- Expanded** RDA metadata for datasets
- Syntactic metadata for files
- Individual data order request information

Research Data Archive Database (RDADB)

APPLICATIONS / SERVICES

- MSS and Data Server usage reports
 - By time, dataset, user, file, ...
 - From command or web view
- MSS RDA file integrity audit:
 - Dataset, password, retention, ...

APPLICATIONS / SERVICES

- MSS filename assignment and dataset registration
- Data order request processing



RDA Database (Metadata Server)



Future Capabilities

Research Data Archive Database (RDADB)

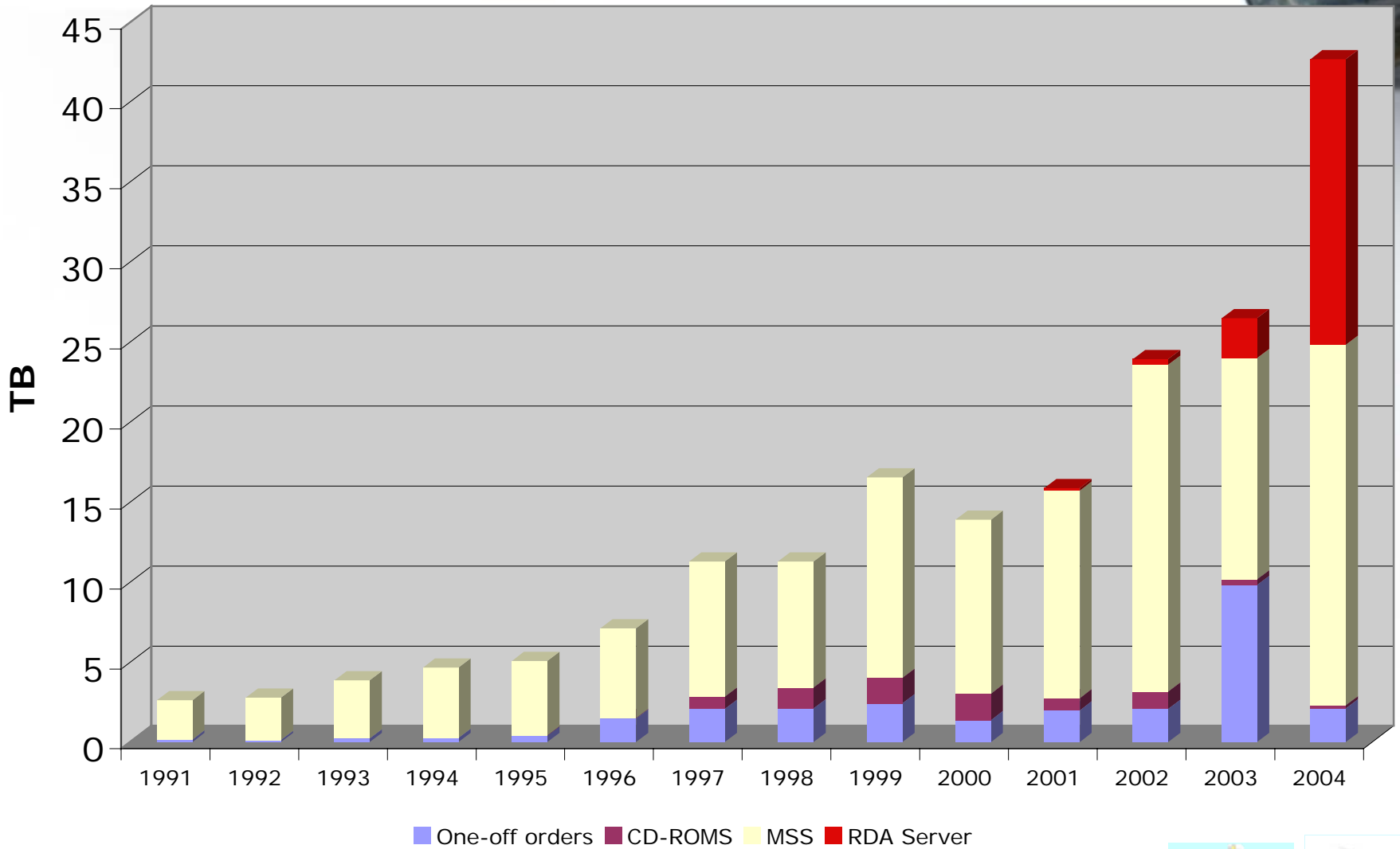


USER UTILITIES

- File selection from search criteria
 - Semantic and syntactic metadata
- Provide pointers to data location
 - MSS, Data Server and CDP
- Provide pointers to documentation and software
- Support MSS file access
 - Pre-form MSS access commands
 - Account for blocking, compression, etc
- Receive and initiate data requests to DSS staff



Research Data Accessed



7/15/2005

15



RDA Server and MSS Example for One Dataset



ERA-40 Data Services, January –June 2005

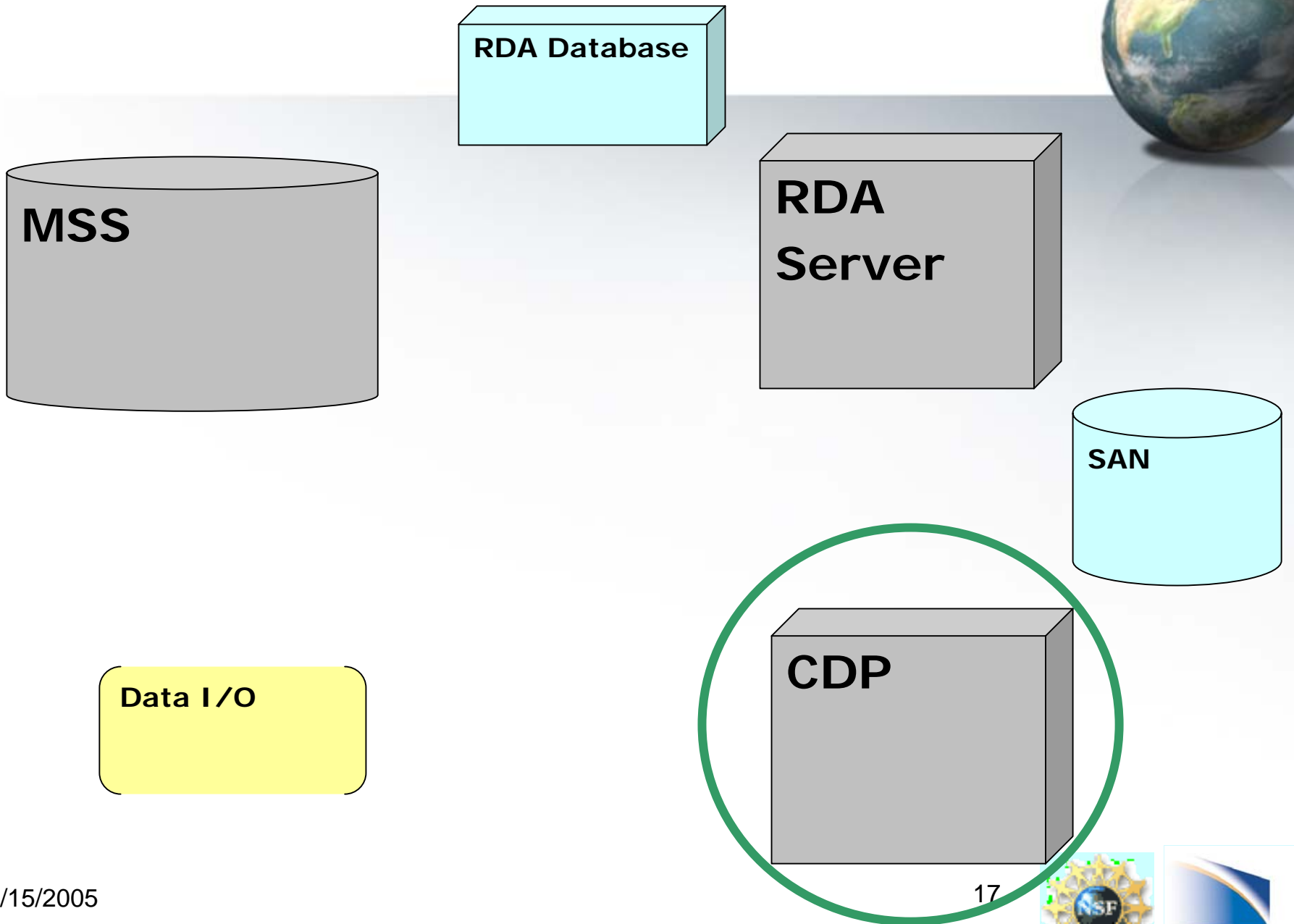
	Web & FTP	NCAR MSS	Total
Unique Users	42	57	95
Number of Data Files	12351	21053	33404
Data Amount (GB)	4949	15600	20549

Note:

- 95 Unique users, total
- 33K files delivered
- 20.5 TB accessed



Component Schematic Diagram for RDA



Community Data Portal (CDP)



- Features
 - Organization-wide facility
 - RDA **plus** many other groups
 - Standard metadata - [minimum requirement](#)
 - CF and GCMD keyword compliant
 - <XML> format
 - Build catalogues
 - Other optional elements
 - Data files, images, movie clips, documentation, model codes, etc....



Community Data Portal (CDP)



- Objectives

- Dissolve cooperate structure from user view and facilitate one stop data discovery
- Enable:
 - Client/server network data access
 - OPENDAP, GDS, LAS interactive access
 - Scientific collaborations between remote groups
 - Easy to use environment
 - A robust system that serves many
 - Eliminate the need for individual groups similar systems



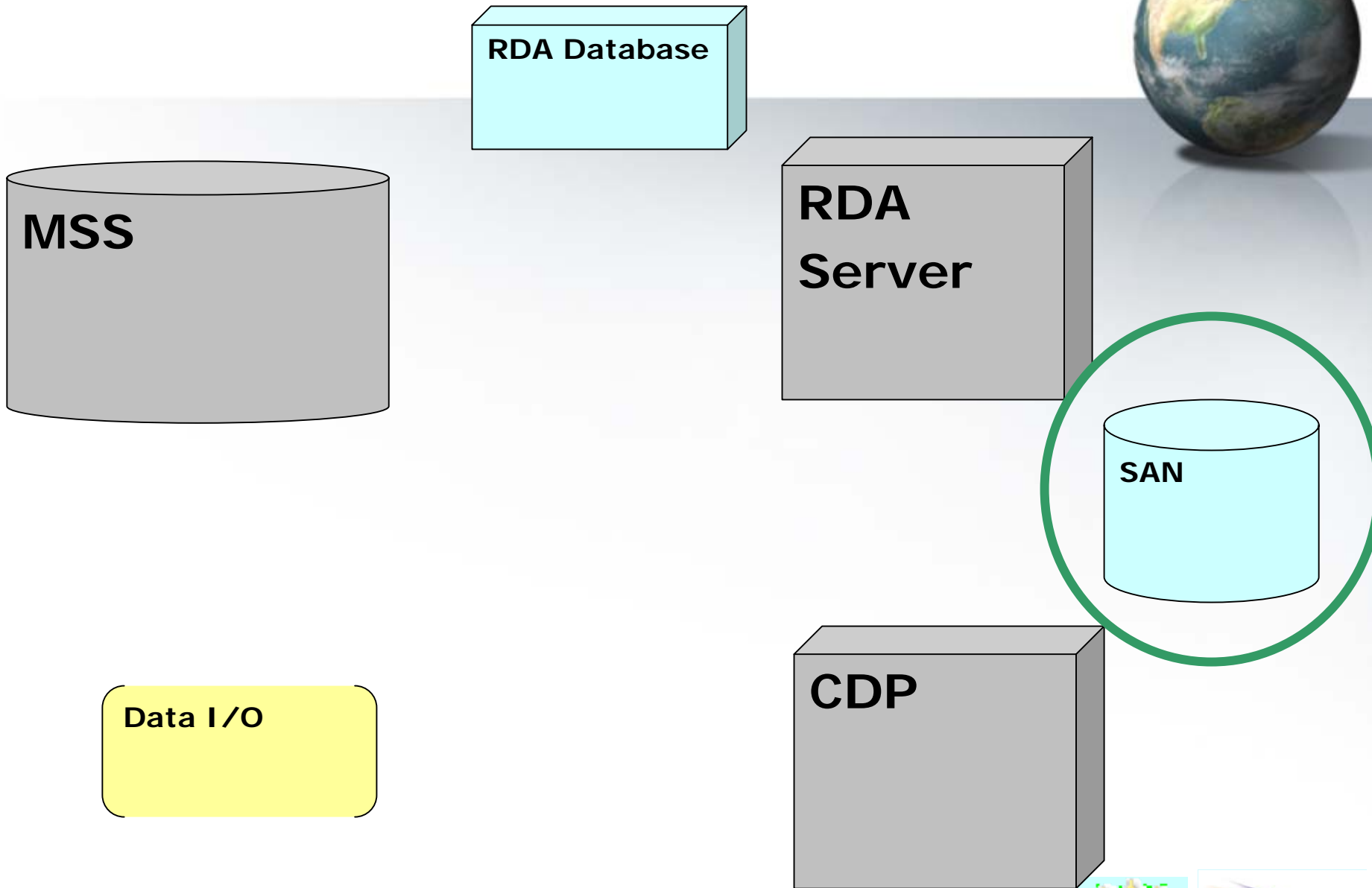
Community Data Portal (CDP)



- Earth System Grid, a CDP subsystem
 - Features
 - Multi-organization (NCAR, DOE, LLNL) shared resources
 - Now, data access only. Future to include computing.
 - Very tight security
 - High level authorization and authentication
 - Advanced software, Globus Toolkit, GridFTP, etc
 - Successful for current AR4 IPCC assessment
 - U.S. contribution to global climate evaluation



Component Schematic Diagram for RDA



SAN



- Features

- New and growing area

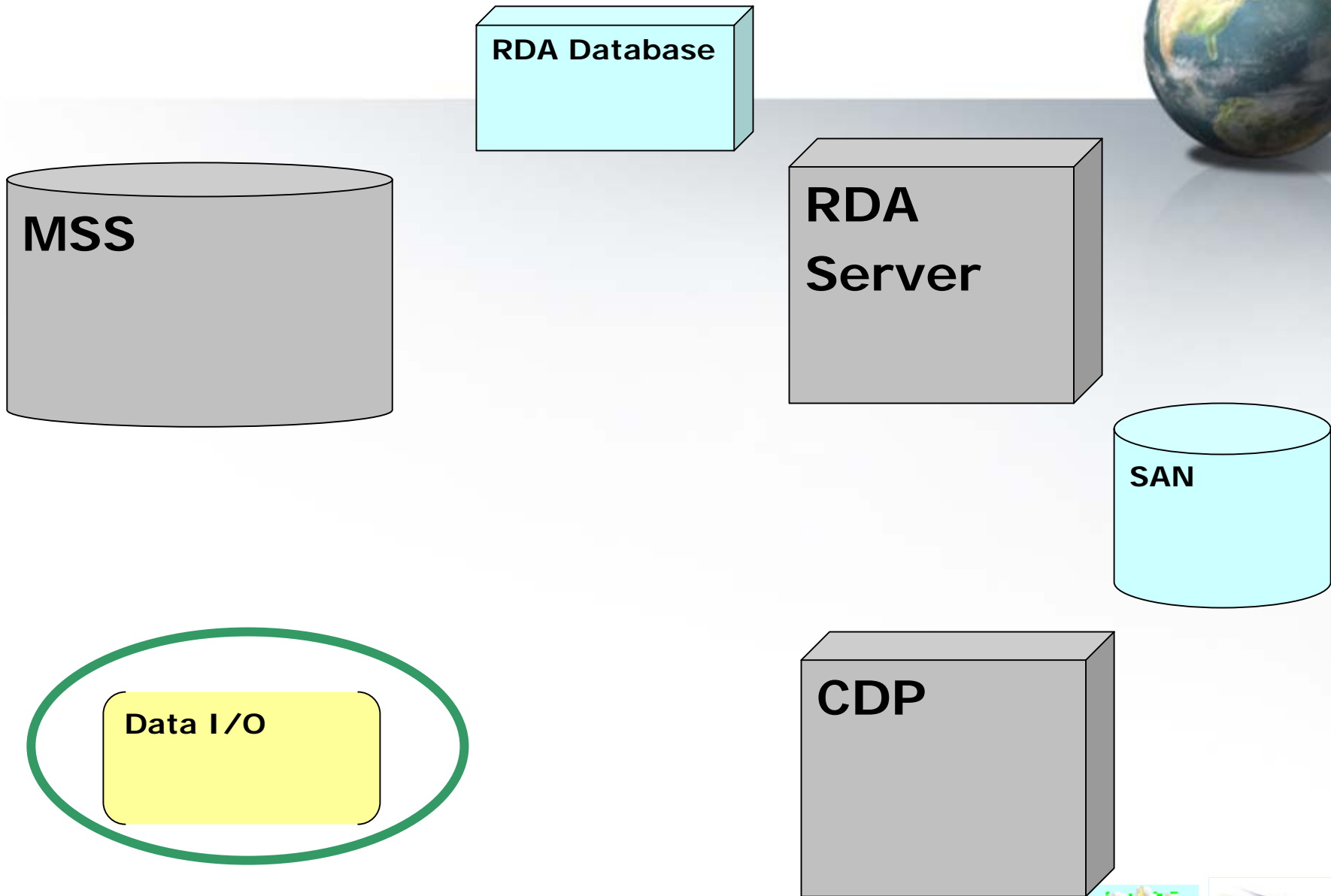
- 32 TB ATA disk with ADIC software
 - Current connections - two data servers
 - RDA and CDP (same architecture, SUN)

- Future

- More ATA storage - target to 60-120TB
 - Heterogeneous servers, e.g. LINUX cluster, SGI, etc



Component Schematic Diagram for RDA



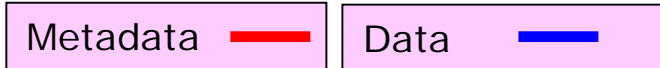
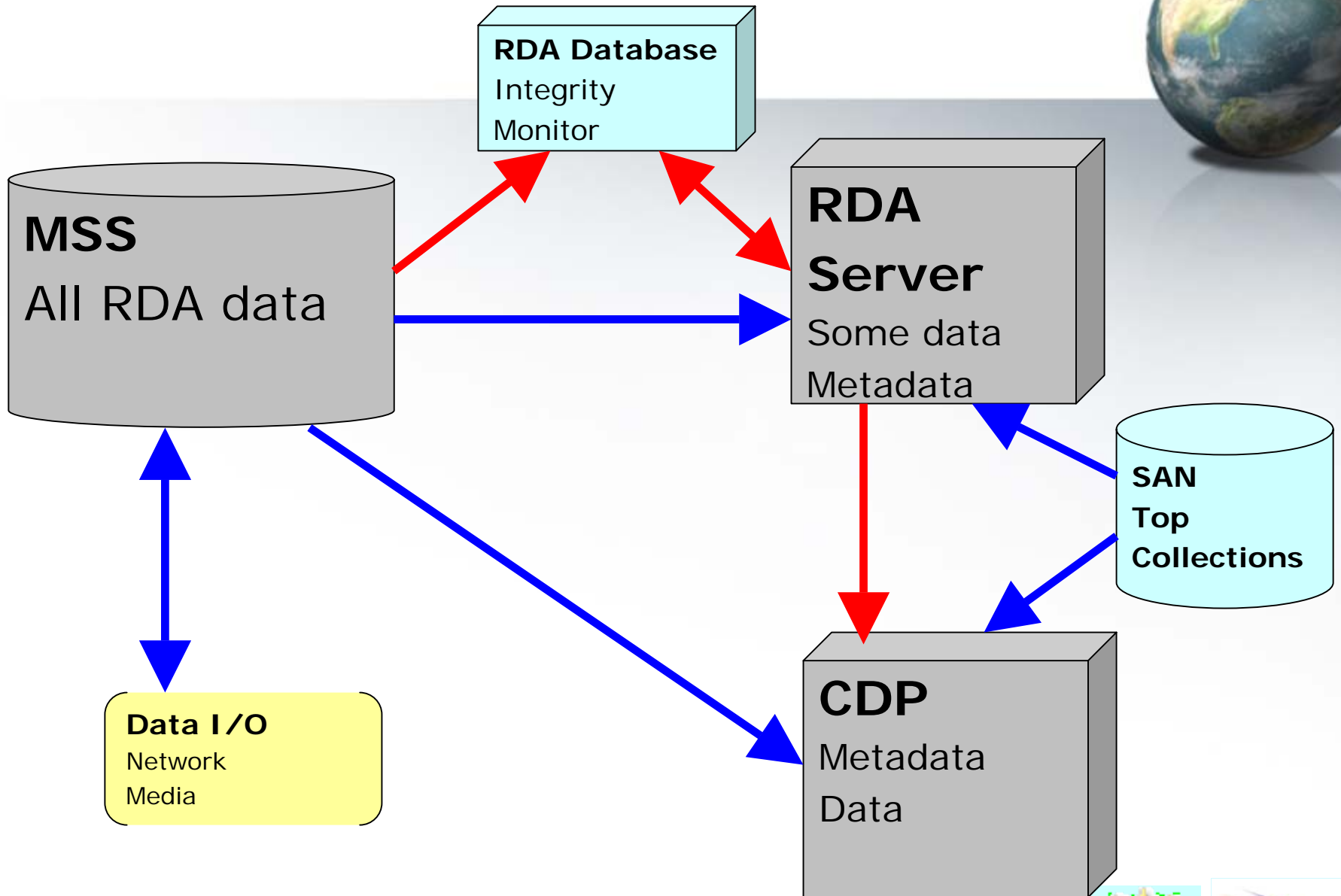
Data I/O



- Objectives
 - I, Build archive content
 - O, Deliver data outside NCAR
- Network transfers I/O
 - used most often
- Media I/O - still important
 - Tapes: LTO, DLT, DAT, Exabyte
 - Disks: CD-ROM, DVD
 - Devices: USB mountable drives
- For data rescue from outside sources
 - Still have 9 and 7 track tape drives



Operational Schematic Diagram for RDA



Conclusion



- Many system component are necessary to manage a RDA
- Components
 - MSS
 - Online Data Server - traditional service
 - Databases
 - Community Data Portal - evolving service
 - SAN
 - Media for Data I/O

