CAS' Scientific Database and its Application System

Dr. YAN, Baoping

Principal of SDB Project

Computer Network Information Center (CNIC)
Chinese Academy of Sciences (CAS)

20th CODATA Conference, Oct.24, 2006, Beijing

Agenda

- About CAS
- Background of SDB Project
- SDB in 2001- 2005
- SDB in 2006- 2010
- Conclusion

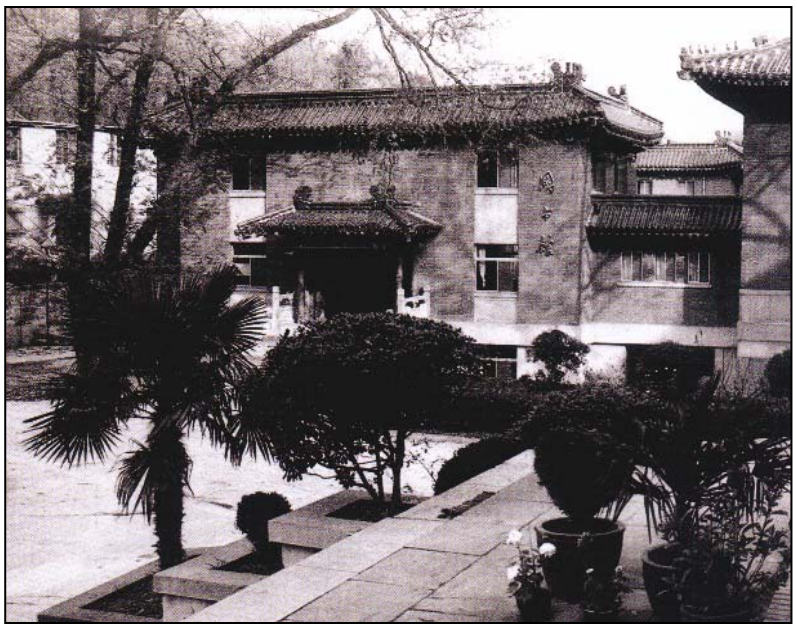
中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Chinese Academy of Sciences (CAS)

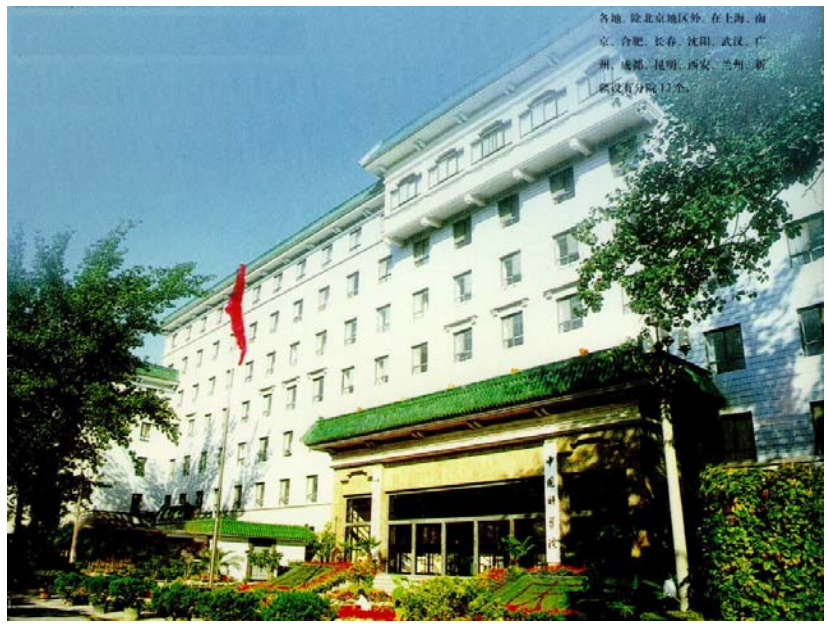
History & Position

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- Founded on Nov. 1, 1949
- Highest academic institution in natural sciences in China
- Most comprehensive R&D center in natural sciences and high-tech development
- Highest national advisory body in S&T



1949



2005

Mission

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

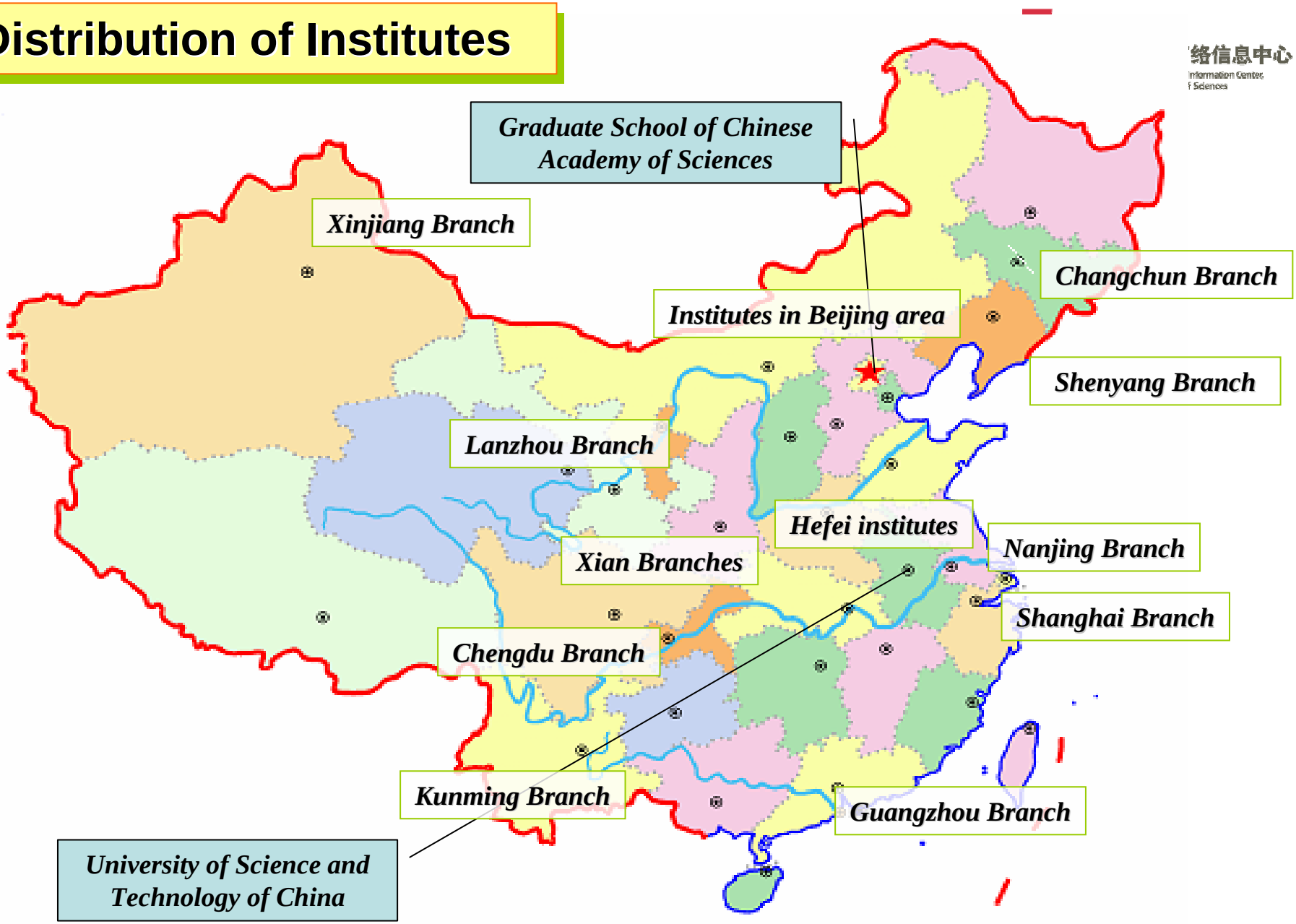
- Target at national strategic needs and world frontiers of science
- Mainly carry out basic and strategic research in an effort to solve major S&T issues of basic, strategic and forward-looking nature in national construction
- Play a key role in the national knowledge innovation system
- Train first-class S&T talents
- Provide scientific bases and tech-innovation sources
- Serve as a national think-tank

Research & Development

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- Total staff: 44,000, of which 13,000 senior and 30,000 other S&T professionals
- Plus 30,000 visiting scholars, post-doctors, and graduates
- 12 branches
- 89 institutes
- Graduate School and USTC
- 9 supporting institutions (tech and docu)
- CAS Holdings Co., 10 major Co. & 490 others

Distribution of Institutes



+200 Wild Field Observatories Distributed

Some Priorities in Basic Research

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- Nano-materials and nano-devices
- Novel quantum phenomena
- Theoretical biophysics, structural and functional of biomacromolecules and bioinformatics
- Brain and cognitive science
- Complex systems
- Functional materials with new structures
- Physics under extreme conditions
- Molecular sciences and engineering
- Particle physics and evolution of universe
- Physics and chemistry in environmental S&T
- Scientific issues in national security
- Interdisciplinary theoretical studies
- Mathematics and interdisciplinary
- Future information sciences
- Space science and technology
- Future energy
- Interior earth and evolution of life in earth
- Large-scale scientific facilities and application of multi-subjects

Priorities in Life Sciences & Biotech

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Biomedical sciences

- System Biology
- Neuroscience
- Brain Function and Cognition
- Reproduction and Development
- Mechanism of Main Diseases
- Immunity and Infection
- Metabolism and Nutrition
- Diagnosis Technique
- Drug Discovery
- Modernization of Traditional Chinese Medicine

Industrial Biotech

- Bio-energy
- Biobased Chemicals
- Biomaterials
- Environmental Biotechnology
- Enzymes, Lipids and Glycose Biology

Agricultural Biology and Biotech

- Crop Design
- Cloning
- Agricultural Functional Genome
- Agricultural Pest Management
- Marine Biotechnology
- Agricultural Resource Management
- Soil Monitoring
- Regional Agriculture

Integrated Biology

- Taxonomy
- Biodiversity
- Ecology
- Global Change Biology
- Conservation Biology
- Gene and Germplasm Bank
- National Botany Garden System

Priorities in Resources and Environment

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- Basic theory and key tech for oil, gas and mines
- Lithosphere evolution
- Qinghai-Tibetan Plateau
- Geo-engineering technologies
- Water resources
- Costal marine ecosystems
- Deep sea environment and life process
- Ocean, continent and atmosphere interaction in Asian monsoon
- Earth system model
- Ecosystem functions
- Biodiversity
- Lake pollution and remediation
- Environment and health
- Eco-environmental effects of key engineering
- Remote sensing monitoring of resources and environment
- Global change

Priorities in High-tech R&D

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Information Technology

- High performance computing
- High performance processor
- Micro electro-mechanical systems
- Wireless sensor network
- Next generation internet
- Information security
- Cognition and computational intelligence
- Quantum information

Space Science and Technology

- Scientific application on the National Spaceflight Program
- Lunar exploration
- Mini and micro satellites
- Space remote sensing
- Geospace environment research and space weather

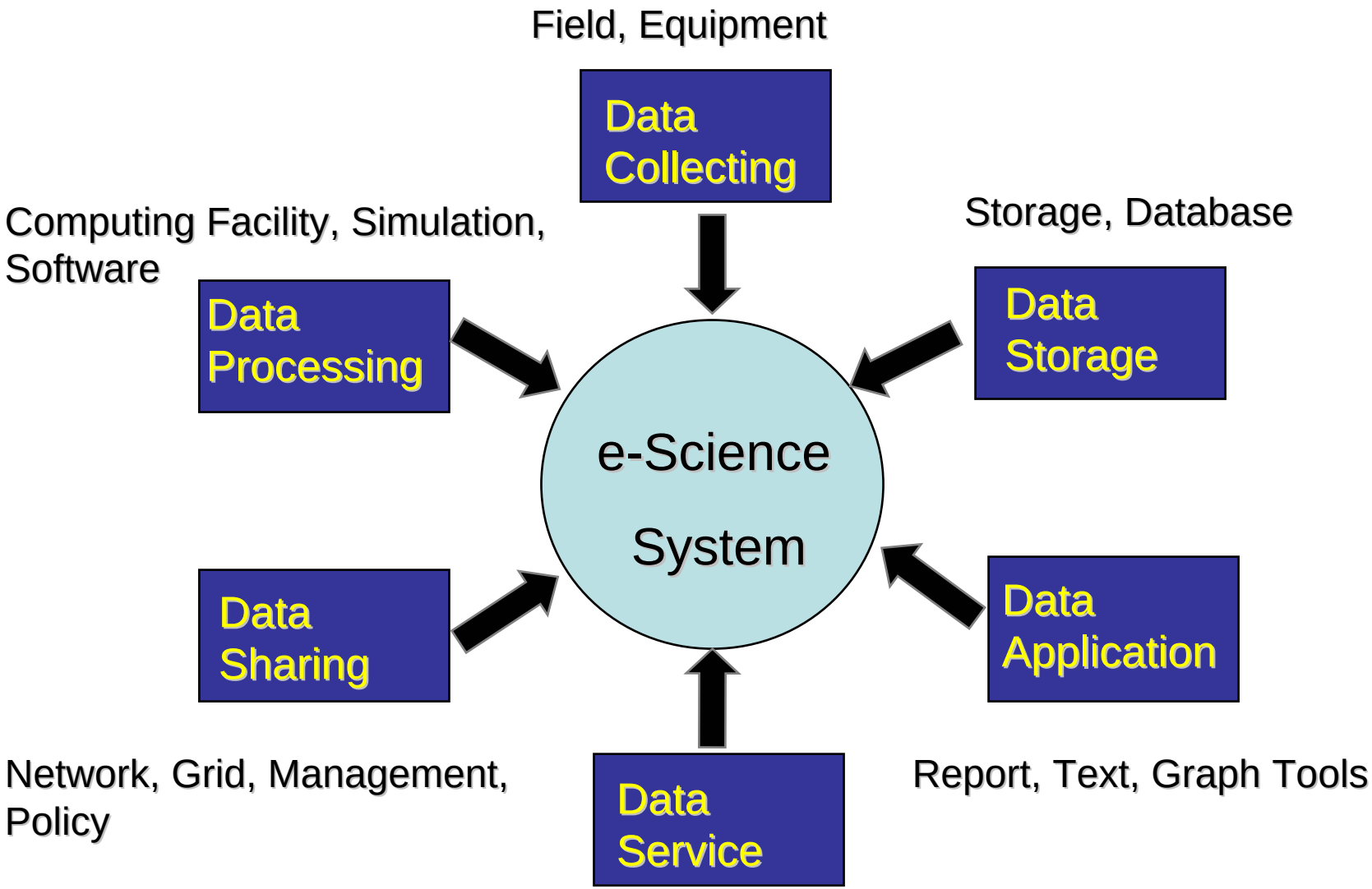
Energy

- Coal based co-production
- Clean coal technology
- Biomass energy
- Solar energy and wind energy
- Hydrogen energy and Fuel cell

Material and Chemical Engineering

- Green production
- Immobilization and utilization of CO_2
- Natural gas conversion
- High performance metallic material
- Advanced non-organic material
- Environment-friendly material
- Bio-material and medical material
- Material designing and computational simulation

SDB in 2001-2005


Field, Equipment
Data Collecting
Computing Facility, Simulation, Software
Storage, Database
Data Processing
Data Storage
e-Science System
Data Sharing
Data Application
Network, Grid, Management, Policy
Data Service
Report, Text, Graph Tools
Search & Retrieval, Content Management

Scientific Database (SDB)

- Data is the one of the foundational elements in e-Science
- data from research, for research, drive e-Science
- SDB is a long-term project since 1982, in which there are multi-disciplinary scientific data accumulated through the course of science activities in CAS
- many institutes involved, long-term, large-scale collaboration

- In 1970s, some chemical institutes under CAS began to build specialized databases
- A large quantity of valuable scientific data have been produced during the long course of research activities at CAS
- In 1982, CAS initiated the idea for establishing “Scientific Database and its Application System”
- In 1986, CAS formally started the construction of SDB, 20th Anniversary this year

Funding

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

As a collection of large-scale, multi-discipline, distributed, scientific databases, SDB is:

- Key engineering project of State Planning Commission （1986-1995）
- Key project of Chinese Academy of Sciences（1986-1990）
- Major project of network application of Natural Science Foundation of China （1995-1996）
- Basic research special support project of Chinese Academy of Sciences（1991-2000）
- Key-project of the 10th five-year planning for information construction of CAS （2001-2005）
- Key engineering Project of National Scientific Data Sharing of MOST（2004-2005）
- Key-project of the 11th five-year planning for information construction of CAS （2006-2010）

CAS Informatization Program 2001-2005

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

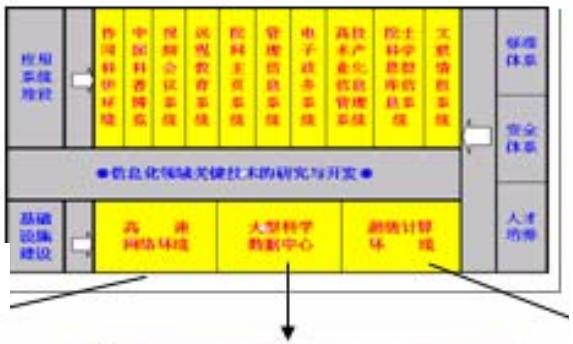
CAS Web Site

Video Conference System

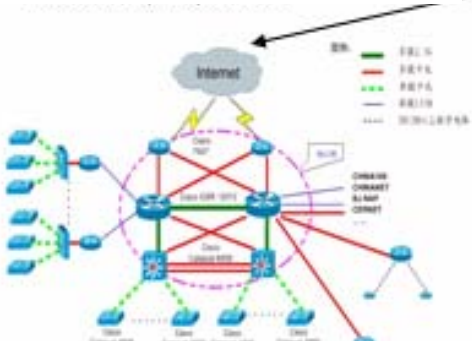
ARP

Distance Education

industry system web site

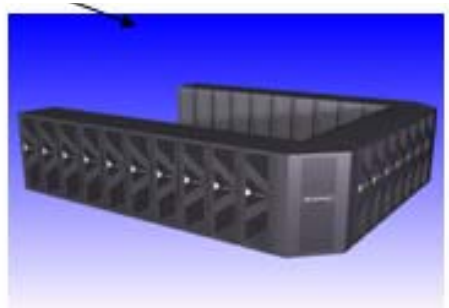


virtual museums



networking

Scientific Database



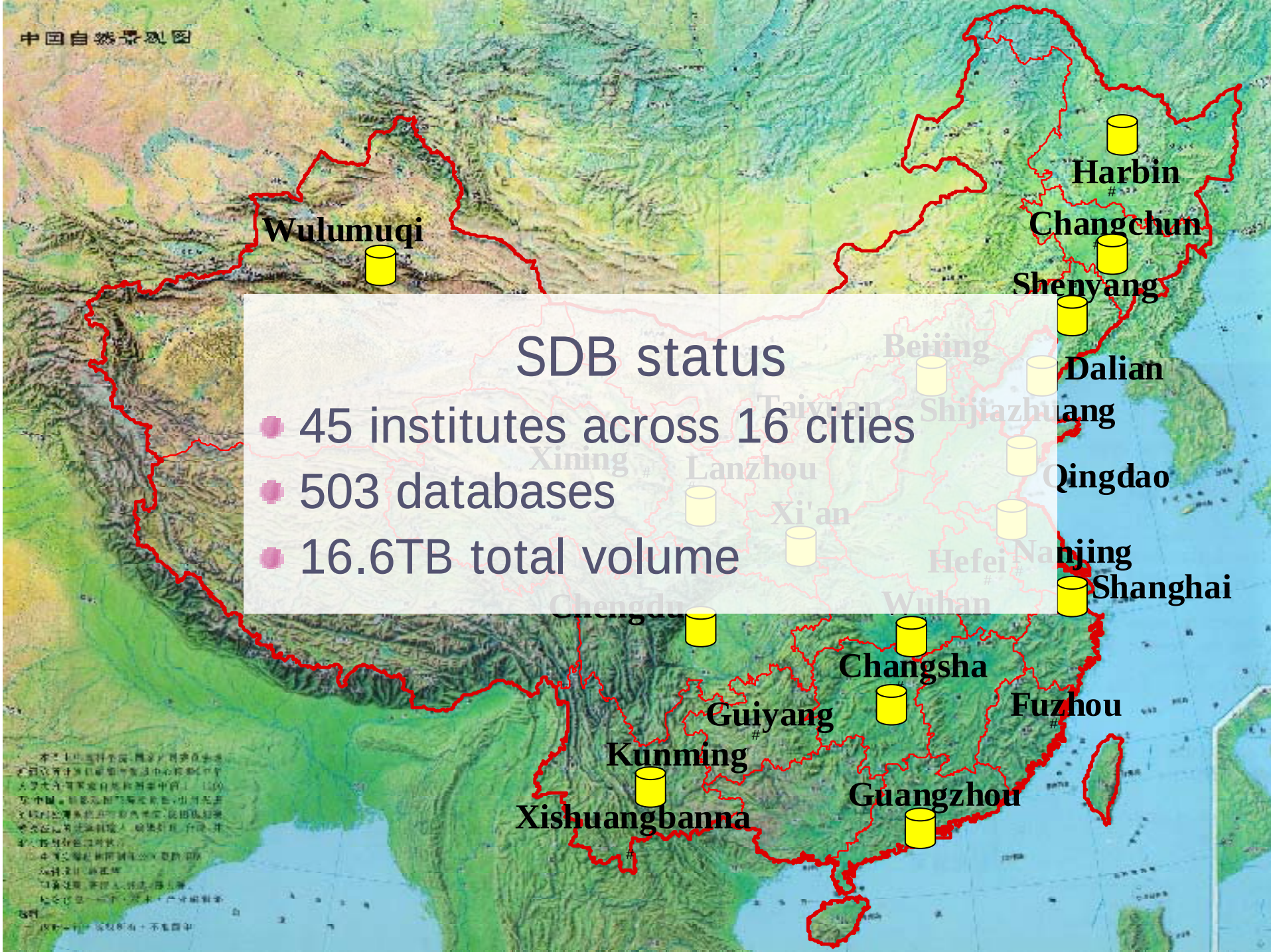
Supercomputing

CAS Cyberinfrastructure Situation

Infrastructure	Item	By 2000	By 2005
Networking	core	1Gbps	**2.5Gbps-10Gbps**
	backbone	2Mbps	**2.5G-155Mbps**
	Oversea link	55Mbps	**620Mbps+15Gbps**
HPC	Peak TFLOPS	0.13	**5.5**
	Linpack TFLOPS	0.05	**4.3**
	Storage	2.1TB	**180TB(Disk+Ta)**
Scientific Database	Member institutes	21	**>45**
	Databases	180	**+503**
	Data volume	725GB	**+16.6TB**

Milestones(2001-2005)

- In 2000, the Scientific Database (SDB) project renewed fund by CAS 10th Five-year Program
- In March 2001, proposed “Scientific Data Grid”
- In October 2002, SDG joined the China National Grid (fund from MOST)
- In Nov 2003, SDG Middleware v1.0 released
- In July 2004, SDG got fund from NSFC
- In Sep 2004, SDG renewed fund from MOST
- In Oct 2004, DeepComp 6800 for SDG installed
- In Nov 2004, SDG Middleware v2.0 released
- In Aug 2005, SDG Middleware v2.1 released
- Now, we’re working for SDG in 11th Five-year Program 2006-2010


SDB status
45 institutes across 16 cities
503 databases
16.6TB total volume
Harbin
Changchun
Shenyang
Wulumuqi
Dalian
Shanghai
Changsha
Guiyang
Fuzhou
Kunming
Guangzhou
Xishuangbanna

Main Tasks in 2001-2005

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- Six main tasks:
  - Database Resource
  - Data & Database Specification
  - IT Infrastructure Constructuring
  - Middle ware Platform - Scientific Data Grid (SDG) Developing
  - SDB & SDG Service
  - Pilot Applications

1.Database Resource

- 45 Institutes and hundreds of researchers have participated in the construction of SDB.
- Data Volume: 16TB+
- The Number of Database:500 +
- Database Content covers Physics, Chemistry, Geosciences, biosciences, Ocean Science, Energy Science, Material Science, Astronomy, Space Science and etc.

Database list（1）

中国科学院
计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

Database Name	Number of sub-DB	Developing Institute
Bio-science(156)		
Chinese Botany Database	10	Institute of Botany
The Database of Tropical and Subtropical Botany	4	South China Botanical Garden
Plants Database of Southwest of China	10	Kunming Institute of Botany
Chinese Pictorial Flora Database	7	Wuhan Botanical Garden
Database of Zoological Science of China	5	Institute of Zoology
The Animal Resource Database of Southwest of China	10	Kunming Institute of Zoology
Microbial Resource Database of China	81	Institute of Microbiology
Database of Virus Resource	5	Wuhan Institute of Virus
Database of Hydrobiology in China	11	Institute of Hydrobiology
China Database of Nucleic Acid and Protein	2	Shanghai Institute of Biological Sciences
Bioinformatics Database	2	Beijing Genomics Institute
The Rice Genome Database	3	National Center for Gene Research
The Database of E-Clone-Genes and Genome Polymorphism	6	Institute of Biophysics

Database list（2）

中国科学院
计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

Database Name	Number of sub-DB	Developing Institute
Geosciences(151)		
Database of China Geotectonics	4	Changsha Intitute of Geotectonics
China Remote Sensing Satellite Image Catalogue System	1	China Remote Sensing Satellite Ground Station
Scientific Database of Subtropical Agri-ecology	4	Institute of Subtropical Agriculture
Mountain Environment & Hazards Database of China	3	Institute of Mountain hazards and Environment
Lake Database of China	2	Nanjing Institute of geography and Limnology
Soil Database of China	7	Nanjing Institute of Soil Science
Natural Resource Database of China	55	Institute of Geographical Sciences and Natural Resource Research
Data Library of Dynamic Geodesy and Resource Environment	2	Institute of Geodesy and Geophysics
The Mineralogy, Petrology and GeoChemistry Database of China	9	Guangzhou Institute of Geochemistry
Database of Soil and Water Conservation in the Loess Plateau	7	Institute of Soil and Water Conservation

Database list（3）

Database Name	Number of sub-DB	Developing Institute
Geosciences(151)		
Xinjiang Ecological and Environmental Resources Database	4	Xinjiang Institute of Ecology and Geography
Database of Chinese Resources and Environment Remote Sensing	11	Institute of Remote Sensing Application
China Wetland Database	11	Northeast Institute of Geography and Agricultural Ecolgoy
Cold and Arid Region Characteristic Database	18	Cold and Arid Regions Environmental & Engineering Research Institute
Agricultural Ecology Database for China's Northeast Black Land	5	Northeast Institute of Geography and Agricultural Ecolgoy
Paleontological and Stratigraphic Database of China	4	Nanjing Institute of Geology and Palaeonotology
Atmospheric Science Database	4	Institute of Atmospheric Physics
Ocean Science(10)		
China Marine Science Database	3	Institute of Oceanology
Ocean Database of South China Sea	7	The South China Sea Institute of Oceanology

Database list（4）

中国科学院
计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

Database Name	Number of sub-DB	Developing Institute
Physics and Chemistry(45)		
High Energy Physics and Related Subject Database	2	Institute of High Energy Physics
Optical System Database	6	Changchun Institute of Optics, Fine Mechanics and Physics
Professional Chemical Database	19	Shanghai Institute of Organic Chemisty
Engineering Chemistry Database	8	Institute of Process Engineering
Database of Applied Chemistry	4	Changchun Institute of Applied Chemistry
理化性能及分析数据库	3	大连化物所
Chemical Substance Toxicity Database	3	Computer Network Information Center
Astronomy & Space Science(12)		
Astronomical Database	8	National Astronomical Observatories
Space Environmental Database	4	Center for Space Science and Applied Research

Database list（5）

Database Name	Number of sub-	Developing Institute
Energy Science(17)		
China Energy Database	5	Energy Research Institute
New Energy Source & Environmental Protection Database	7	Guangzhou Institute of Energy Conversion
Gas Hydrate Database	5	CNIC and etc.
Material Science(12)		
Materials Database	7	Institute of Mental Research
Nano Science and Nano Technology Fundamental Database	5	Nano Science and Technology Center
Others(1)		
Specialists of Chinese Academy of Sciences	1	

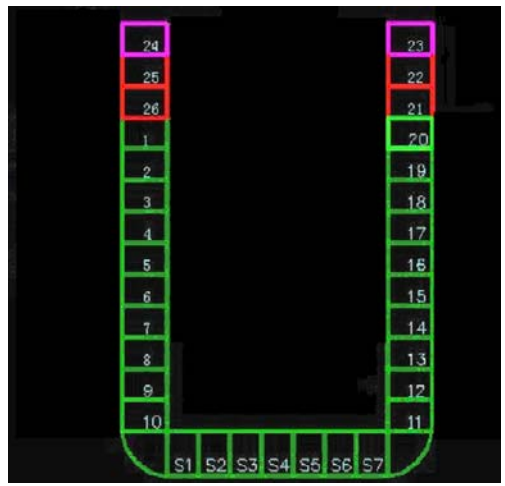
2.Data & Database Specification and Standard

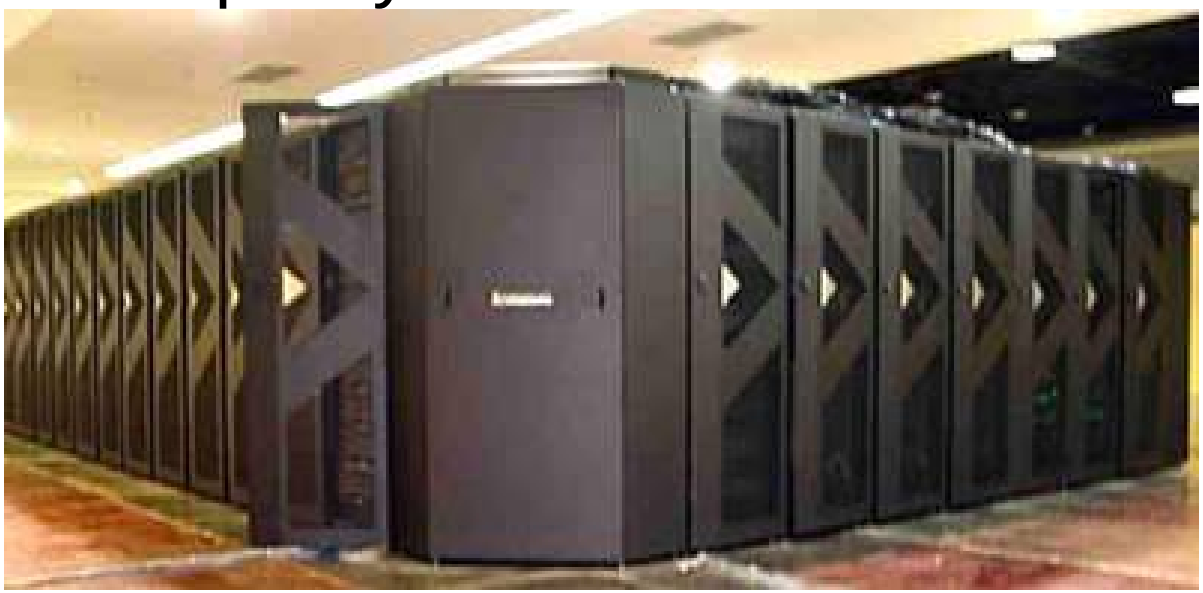
In order to Standardize the process of database construction and database Schema for data integration, Series of specifications for SDB have been published .

- The standard process of scientific database construction and document specification
- Data Sharing Policy and specification for data sharing statement
- Core Metadata Specification for SDB（Ver2.0)
  - A metadata repository and clearing house has been established in the Scientific Data Center
- Some metadata specification for special domains
  - Flora Images, Ecological Data, biological species and so on.
- The Framework for Data quality control and evaluation

3.IT Infrastructure Construction

- Data Center
  - 20TB SAN Storage
  - 50TB Tape Storage
  - TFLOPS-scale computing capacity

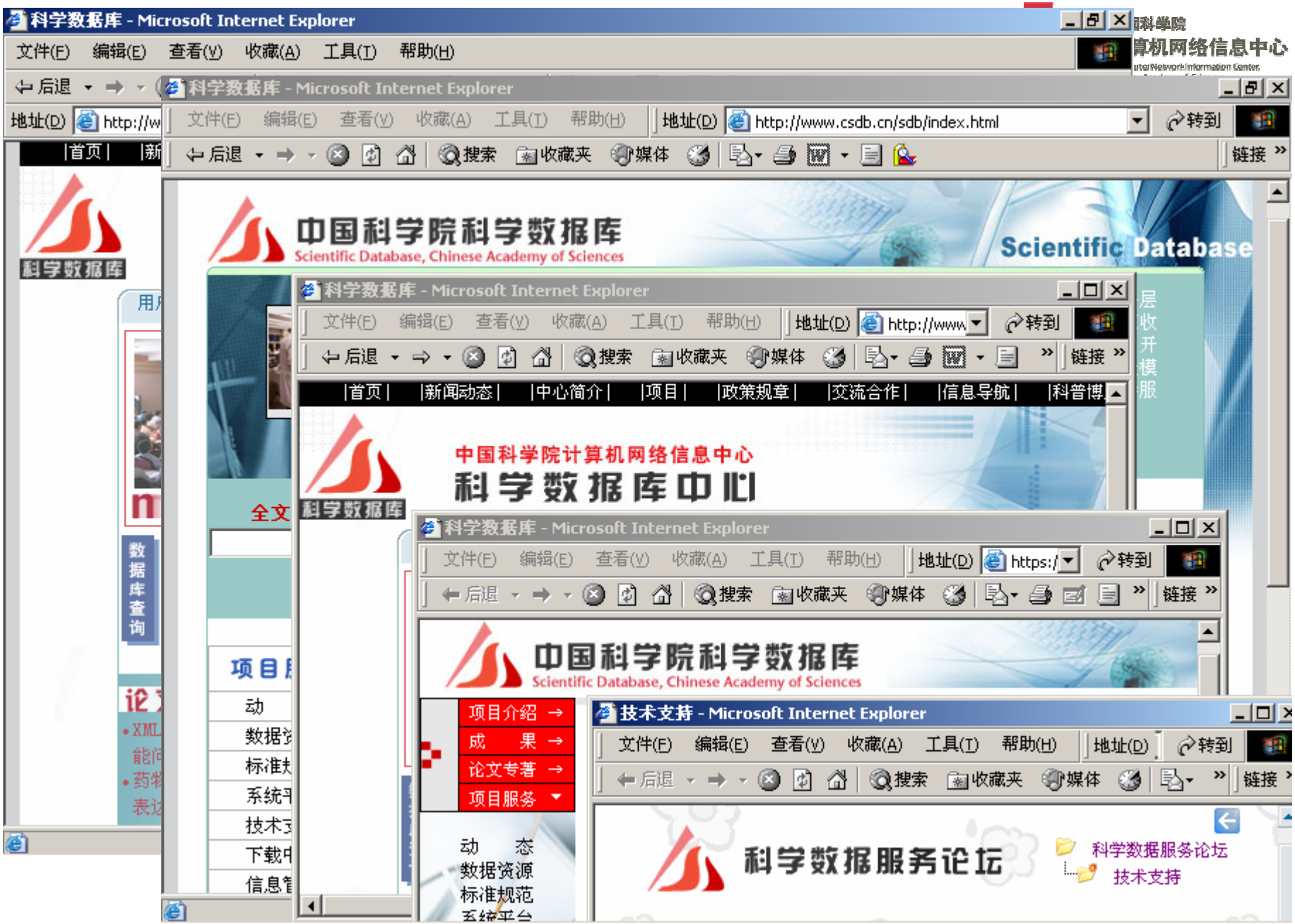




Lenovo DeepComp 6800

4.Data Service

- A Portal website of SDB has been established and put into service at http://www.csdb.cn
- Over 40 distributed data service websites have been built
- A portal website for technique communication and supporting in SDB community has been established, https://support.csdb.cn


科学数据库 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
地址(D) http://www.csdb.cn/sdb/index.html
转到
后退 搜索 收藏夹 媒体 链接
中国科学院科学数据库
Scientific Database, Chinese Academy of Sciences
Scientific Database
|首页| |新闻动态| |中心简介| |项目| |政策规章| |交流合作| |信息导航| |科普博
中国科学院计算机网络信息中心
科学数据库中心
科学数据库
数据库查询
项目介绍
成 果
论文专著
项目服务
动 态
数据资源
标准规范
技术支持 - Microsoft Internet Explorer
科学数据服务论坛
科学数据服务论坛
技术支持

5.Scientific Data Grid (SDG)

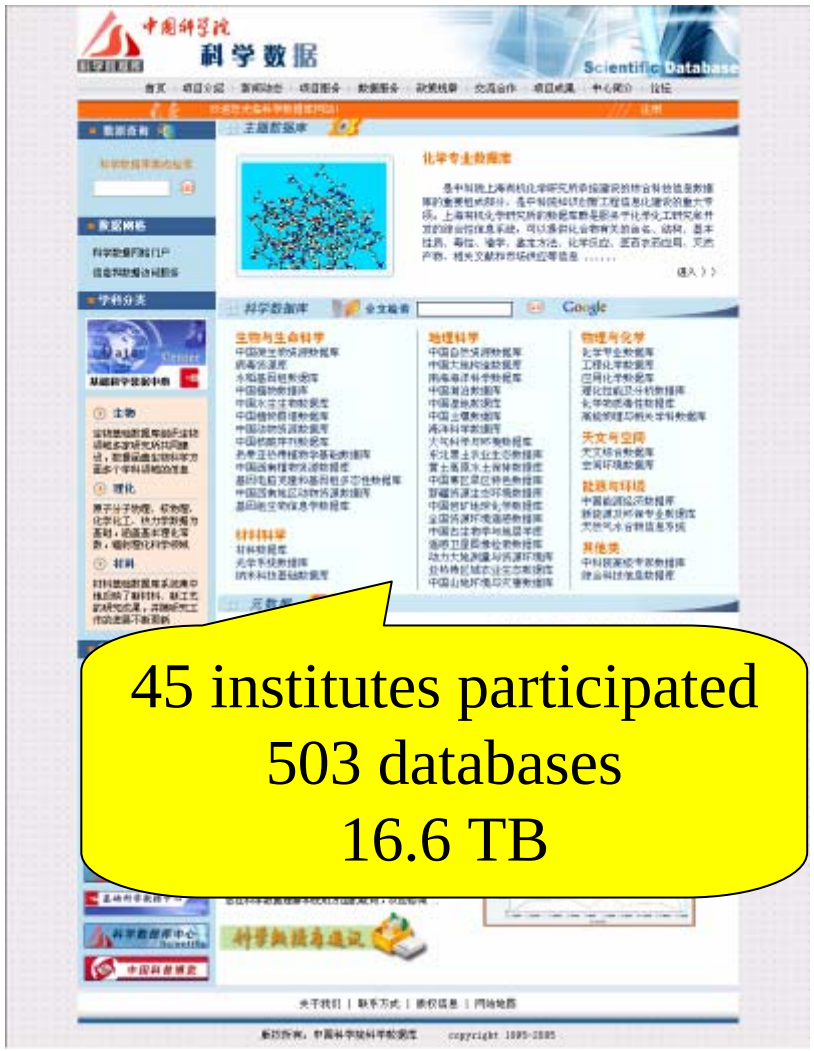
- Scientific data is one of three poles of the cyber infrastructure of CAS
  - Networks
  - Computing
  - Database
- SDG is a sub-project of SDB

Scientific Data Grid

- SDG is built upon the mass scientific data resources of the Scientific Database (SDB).
- Scientific Data Grid (SDG) is a typical project of CAS e-Science based on SDB, also a pilot.
- The vision of SDG is to take valuable data resources into full play by benefiting from advanced information technologies, in particular, the Grid technology.

Scientific Database (SDB) & Scientific Data Grid (SDG)

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences



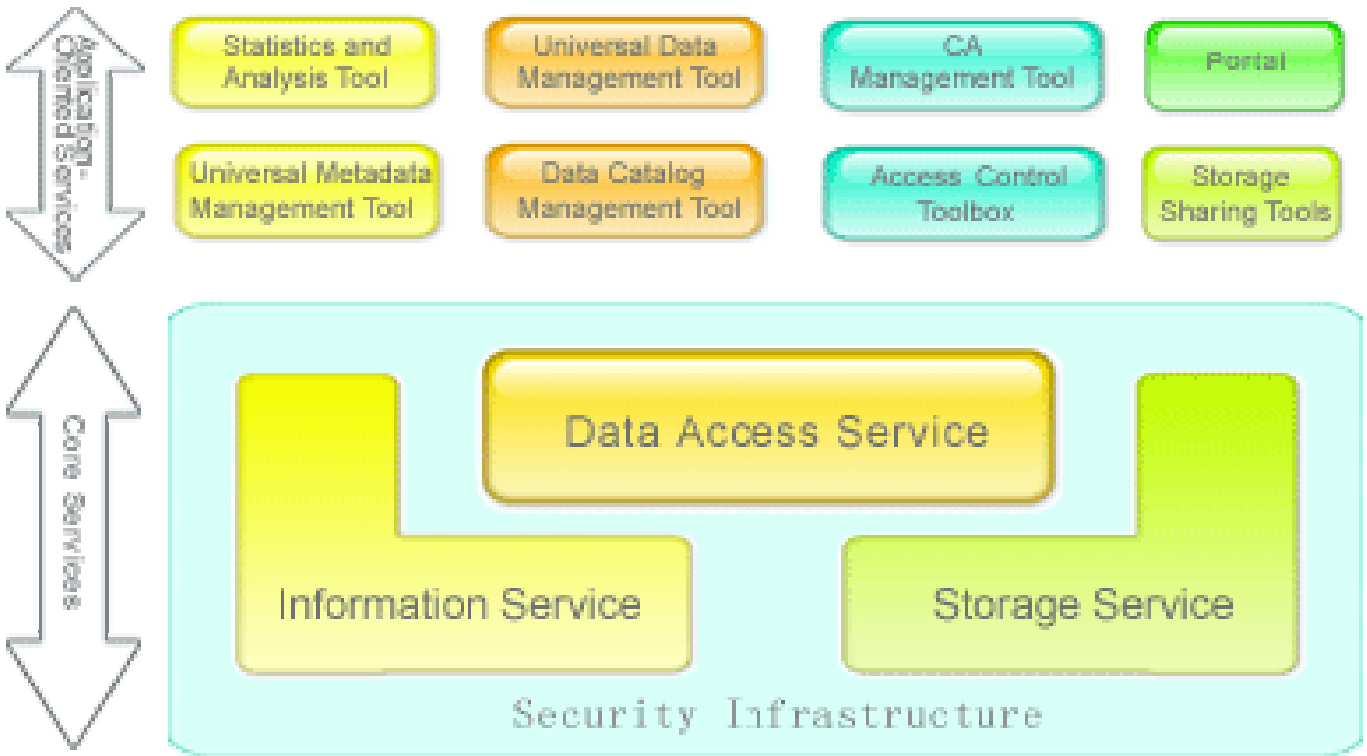
45 institutes participated
503 databases
16.6 TB

236-CPU Superserver (1TF)
20TB Disk Array
50TB Tape Library
VizWall & Access Grid

Requirements and SDG

- How to FIND the data I want from hundreds or thousands of databases
- How to ACCESS large-scale, distributed and heterogeneous scientific data uniformly and conveniently
- How to make sure all this goes always in a SECURE and proper way

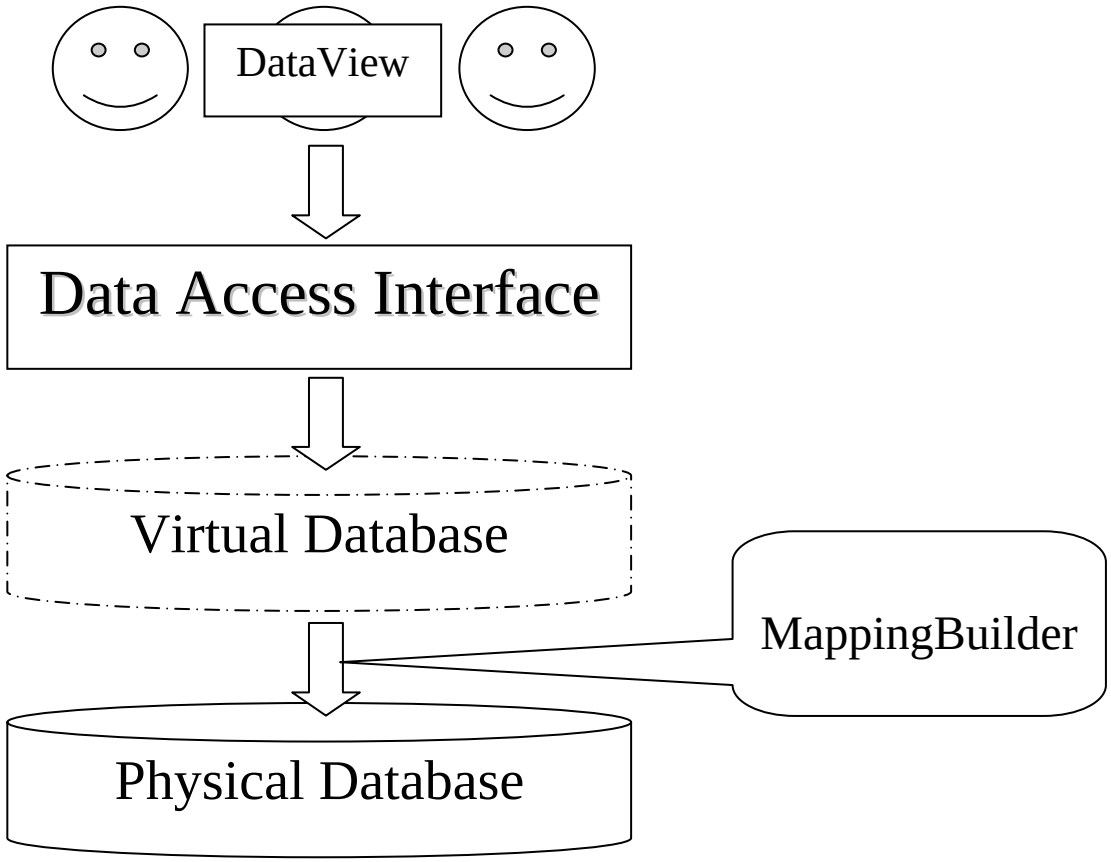
SDG Software Architecture



Data Access Service (DAS)

- Uniform Access Interface (read-only)
- Rich metadata
- Easy publish on web
- flexible configuration and extensibility

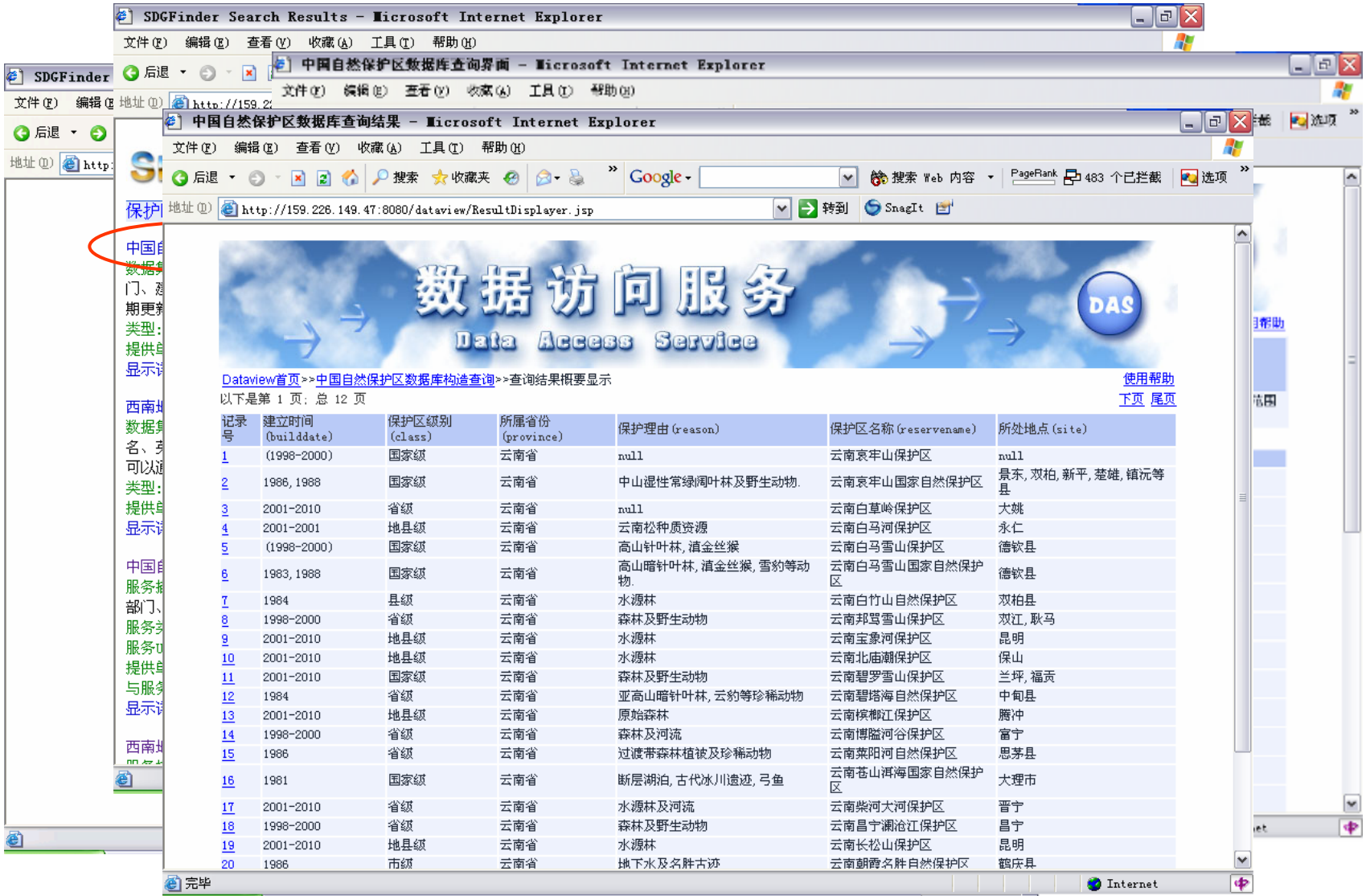
DAS modules


DataView
Data Access Interface
Virtual Database
MappingBuilder
Physical Database

SDG Services

中国科学院
计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

Discovery and Access



记录号	建立时间 (builddate)	保护区级别 (class)	所属省份 (province)	保护理由 (reason)	保护区名称 (reservename)	所处地点 (site)
1	(1998-2000)	国家级	云南省	null	云南哀牢山保护区	null
2	1986, 1988	国家级	云南省	中山湿性常绿阔叶林及野生动物.	云南哀牢山国家自然保护区	景东, 双柏, 新平, 楚雄, 镇沅等县
3	2001-2010	省级	云南省	null	云南白草岭保护区	大姚
4	2001-2001	地县级	云南省	云南松种质资源	云南白马河保护区	永仁
5	(1998-2000)	国家级	云南省	高山针叶林, 滇金丝猴	云南白马雪山保护区	德钦县
6	1983, 1988	国家级	云南省	高山暗针叶林, 滇金丝猴, 雪豹等动物.	云南白马雪山国家自然保护区	德钦县
7	1984	县级	云南省	水源林	云南白竹山自然保护区	双柏县
8	1998-2000	省级	云南省	森林及野生动物	云南邦骂雪山保护区	双江, 耿马
9	2001-2010	地县级	云南省	水源林	云南宝象河保护区	昆明
10	2001-2010	地县级	云南省	水源林	云南北庙潮保护区	保山
11	2001-2010	国家级	云南省	森林及野生动物	云南碧罗雪山保护区	兰坪, 福贡
12	1984	省级	云南省	亚高山暗针叶林, 云豹等珍稀动物	云南碧塔海自然保护区	中甸县
13	2001-2010	地县级	云南省	原始森林	云南槟榔江保护区	腾冲
14	1998-2000	省级	云南省	森林及河流	云南博隘河谷保护区	富宁
15	1986	省级	云南省	过渡带森林植被及珍稀动物	云南菜阳河自然保护区	思茅县
16	1981	国家级	云南省	断层湖泊, 古代冰川遗迹, 弓鱼	云南苍山洱海国家自然保护区	大理市
17	2001-2010	省级	云南省	水源林及河流	云南柴河大河保护区	晋宁
18	1998-2000	省级	云南省	森林及野生动物	云南昌宁澜沧江保护区	昌宁
19	2001-2010	地县级	云南省	水源林	云南长松山保护区	昆明
20	1986	市级	云南省	地下水及名胜古迹	云南朝霞名胜自然保护区	鹤庆县

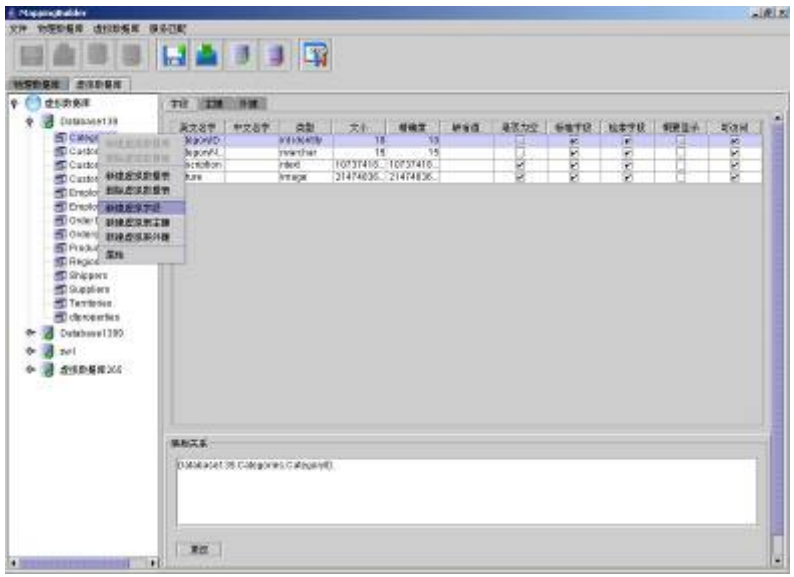

数据访问服务
Data Access Service
DAS
中国自然保护区数据库


数据访问服务
DAS
县级地名数据库

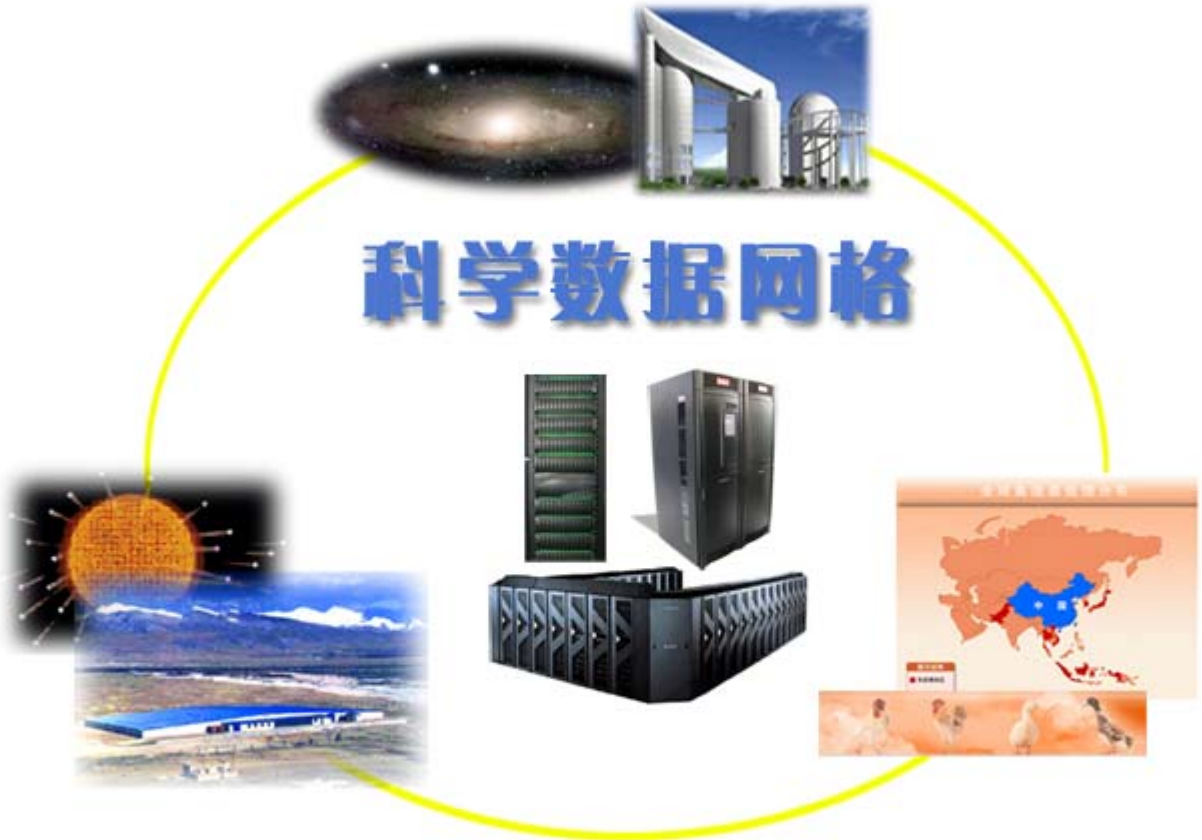

数据访问服务
Data Access Service
DAS
脊椎动物标本数据库


数据访问服务
DAS
云南森林昆虫分布名录数据库

MappingBuilder & Dataview



SDG Today



sdb6800 Superserver

- 59 nodes /236 CPUs
- official service started in Apr. 2005
- node usage 79.7% storage usage 87% (by Sep 2005)

SDG Storage System

Visualization System

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

portal.sdg.ac.cn

Collaborations

- PRAGMA
  - www.pragma-grid.net
- EUChinaGrid
  - www.euchinagrid.org *Interconnection and Interoperability of Grids between Europe & China*
- IGTF / ApGrid PMA
- …

5.e-Science applications5

- High Energy Physics
- Astronomy
- Biology
- Natural Resources
- Disaster Reduction
- …

YBJ-ARGO/ASγ

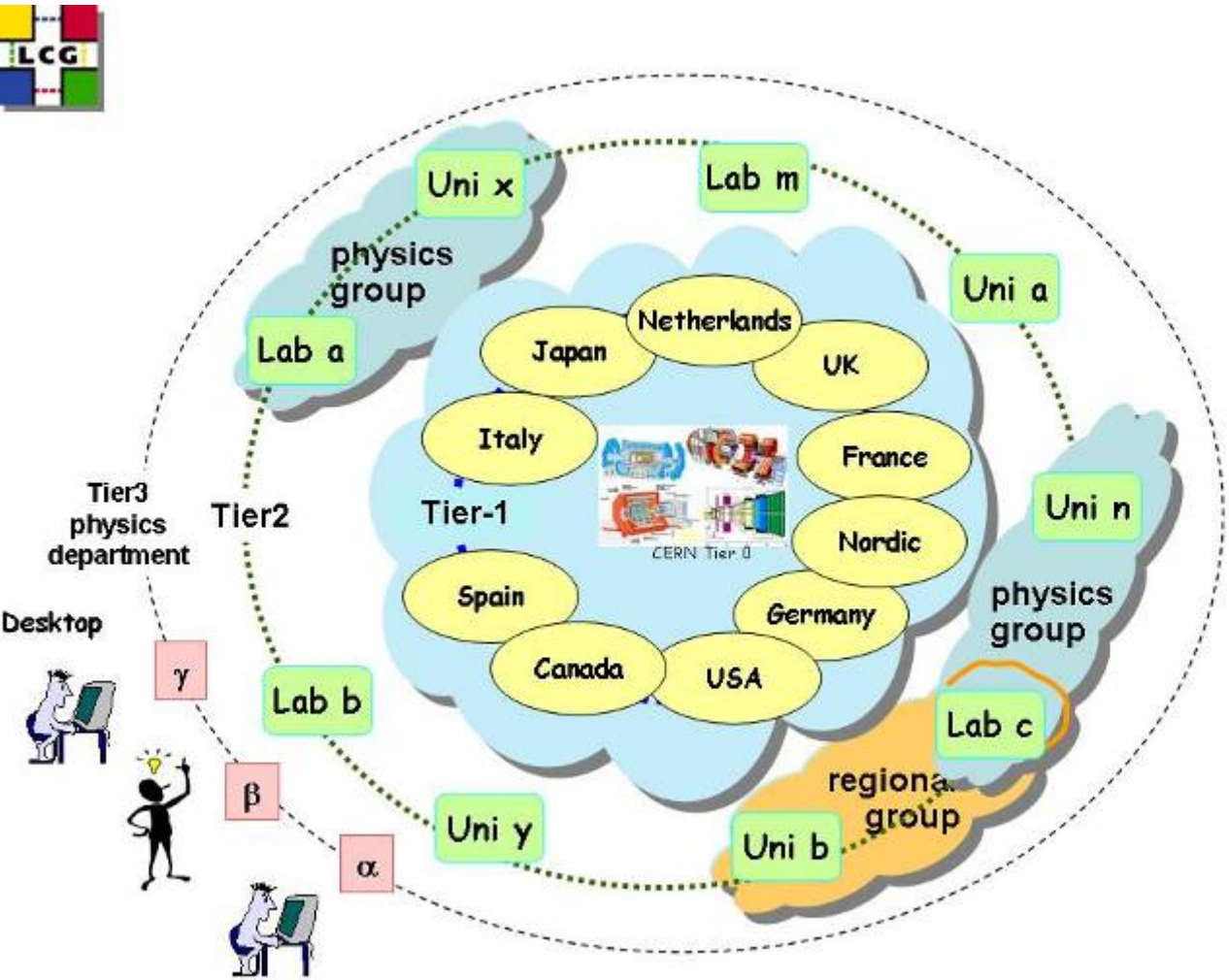
- Italy,Japan-China cosmic ray observatories in Tibet.
- 200TB raw data per year.
- Data transferred to IHEP and processed with 400 CPUs.
- Rec. data accessible by collaborators.

YBJ-ARGO

- Established a 8Mb/s link from Tibet to Beijing in March 2005, by CNIC of CAS. Upgraded to 155Mb/s in March 2006.
  - Stopped bringing tapes half year ago.
- Building a computing system based on LCG, collaboration of IHEP of CAS, CNIC of CAS, INFN of Italy , EU-China Grid application under EU FP6.

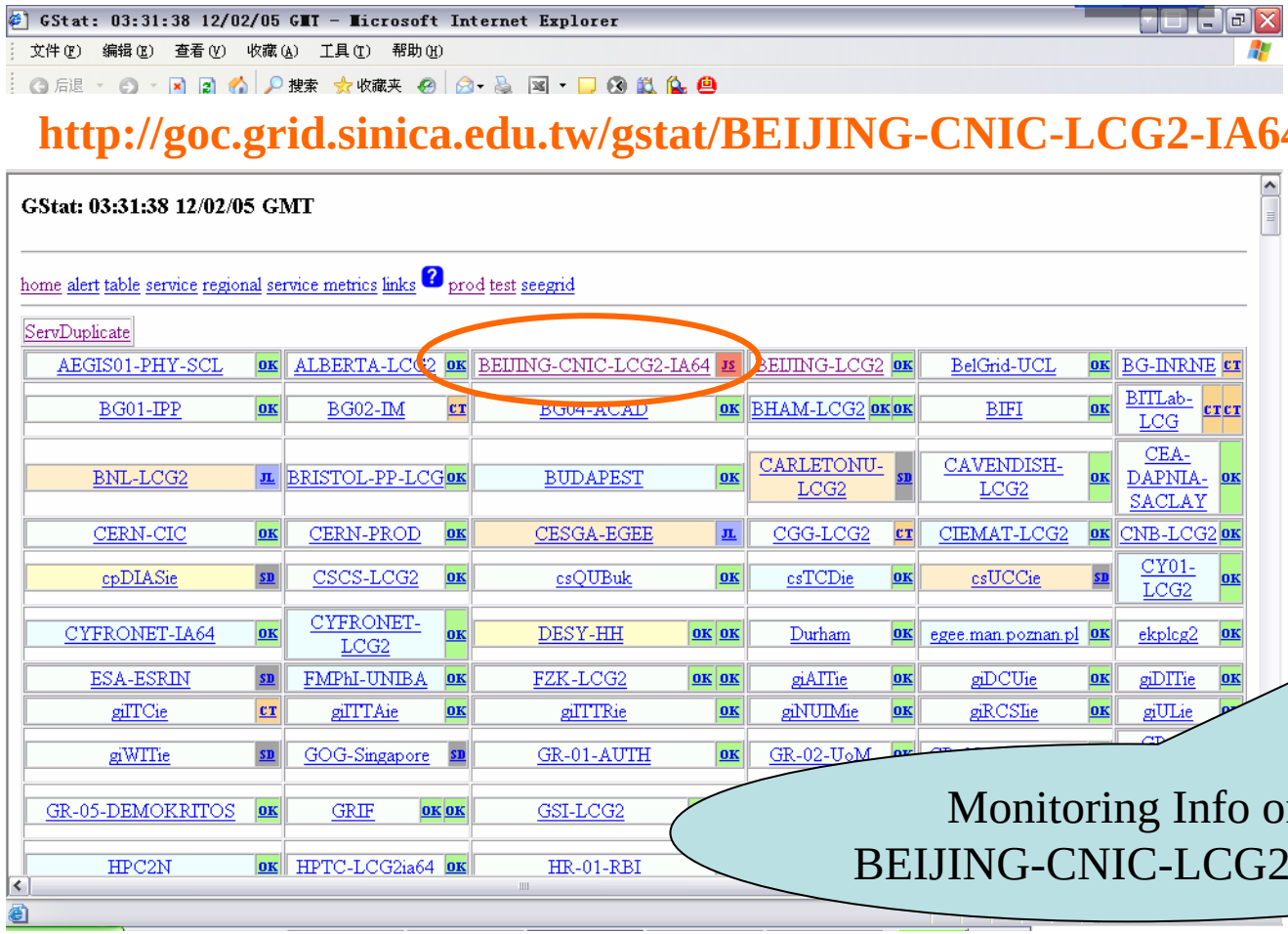
中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

LCG Tier-1/2

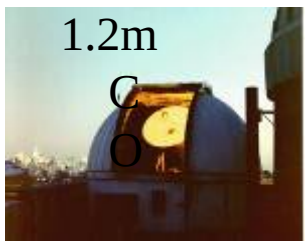


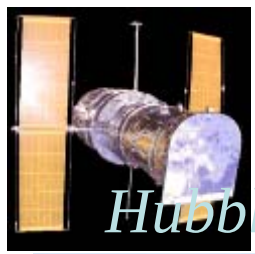
- **to build a LCG Tier-1/2 node in China**
- Institute of High Energy Physics of CAS
- CNIC providing support and working together with IHEP

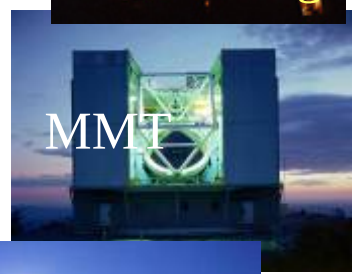
LCG2 production site @CNIC



VO = World Wide Telescope







China Virtual Observatory at SDG Portal

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Avian Bird Flu Alarming & Predicating System

By: Institute of Microbiology, CAS
Institute of Zoology, CAS
Institute of Virology, CAS
CNIC, CAS

Avian Bird Flu in Gangcha, Qinghai Province, May 2005

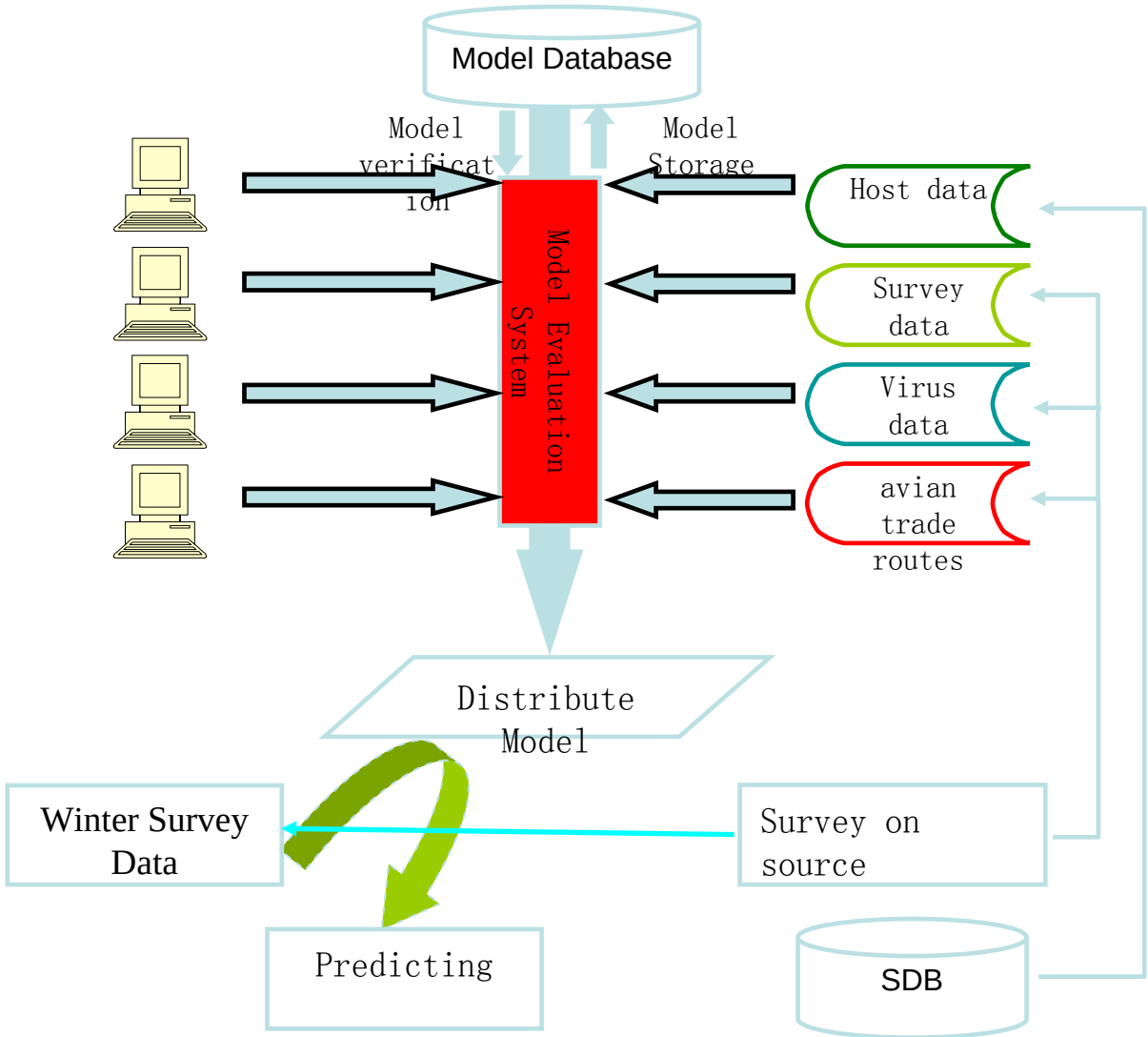
Tasks

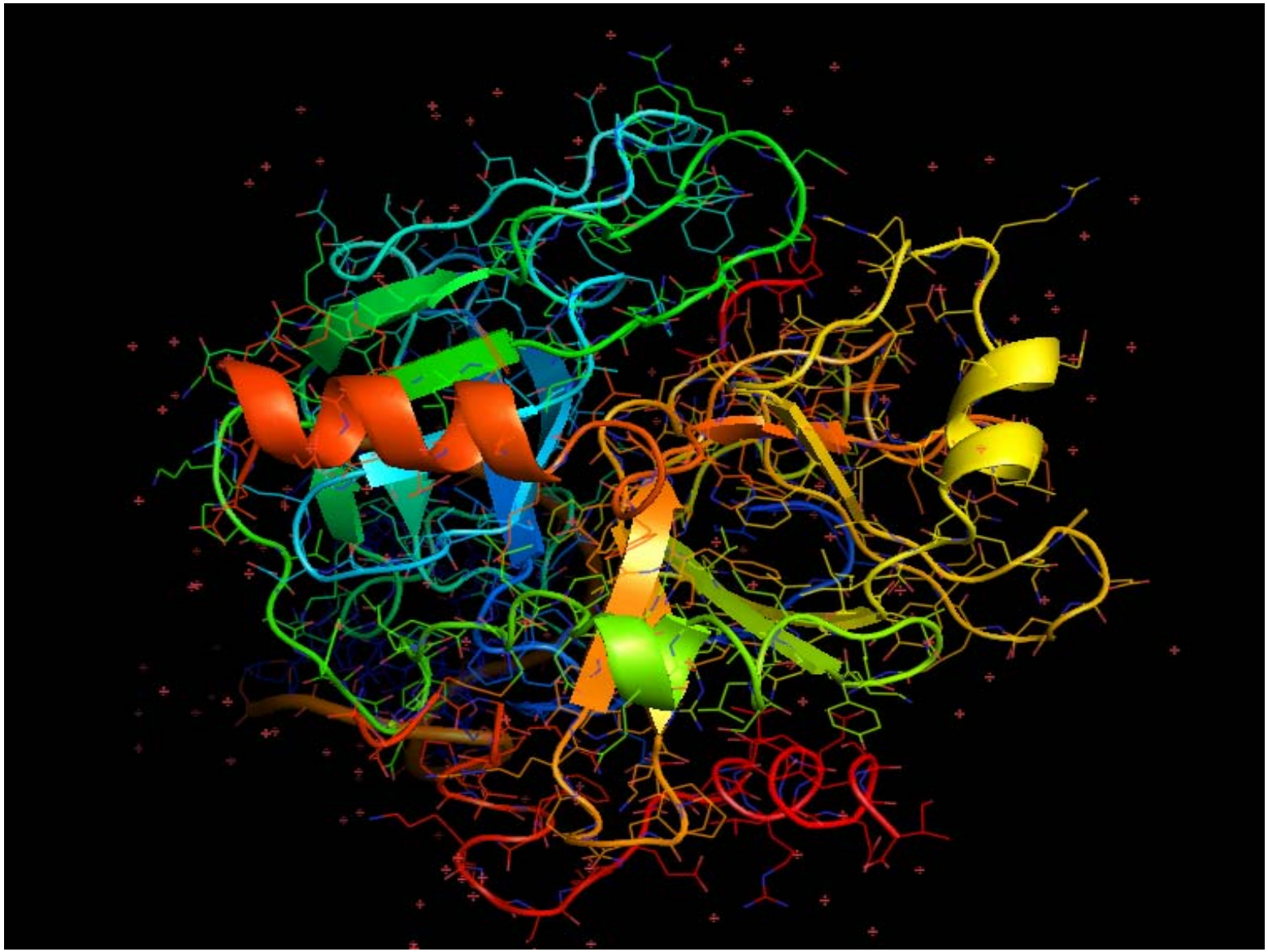
- Integrate bird-flu basic databases from multiple institutes
- Field survey on bird-flu
- Establish bioinformatics comprehensive analysis system for bird-flu
- Establish bird-flu alarming and predicting system
- Establish international cooperative work environment
- Establish information publishing system (web)

Bird-flu basic databases

- Standards
  - Bird-flu basic database's model and data standard
  - Metadata specification and description language of bird-flu information
- Data resources
  - Bird-flu virus resource database
  - Bird-flu virus inherent resource database
  - Bird-flu history database
  - Bird-flu dynamic monitoring database
  - Bird-flu host database
  - Bird-flu information database
  - Bird-flu international DNA database
  - Bird-flu international research progress database

Technical architecture





禽流感信息网

http://www.avian-flu.info

新闻动态 (565)

- 盲目扑杀家禽并不能阻止禽流感的传播 (02-01)
- 越南今年12人死于禽流感 (02-01)
- WHO：关注禽流感新动向 (01-28)
- 越南26省发现禽流感 两名女孩已被证实... (01-28)
- 亚太经合组织将在美举行禽流感研讨会 (01-28)
- 《新英格兰医学杂志》：禽流感病毒可通... (01-28)
- 泰国研究证实：禽流感可通过人际传播 (01-24)
- 世卫称可能发生通过人类接触传染的禽流... (01-23)
- 泰国又发现禽流感 (01-21)

更新新闻...

新闻检索： 提交查询

学术文献 (228)

- Avian H5N1 Influenza in Cats. (09-02)
- H5N1 chicken influenza viruses displ... (09-01)
- Molecular determinants within the su... (09-01)
- H5N1 influenza: a protean pandemic t... (05-25)
- Assessment of risk to human health a... (05-21)
- Public health. Public health risk fr... (05-14)
- Use of the DNA flow-thru chip, a thr... (05-01)
- Reemerging H5N1 influenza viruses in... (05-01)
- Antiviral chemotherapeutic agents ag... (05-01)

更新新闻...

文章检索： 提交查询

最新疫情 (112)

- 越南今年12人死于禽流感 (02-01)
- 越南26省发现禽流感 两名女孩已被证实感染 (01-28)
- 泰国又发现禽流感 (01-21)
- 越南禽流感患者增至7人 其中5人已经死亡 (01-20)

知识介绍 (76)

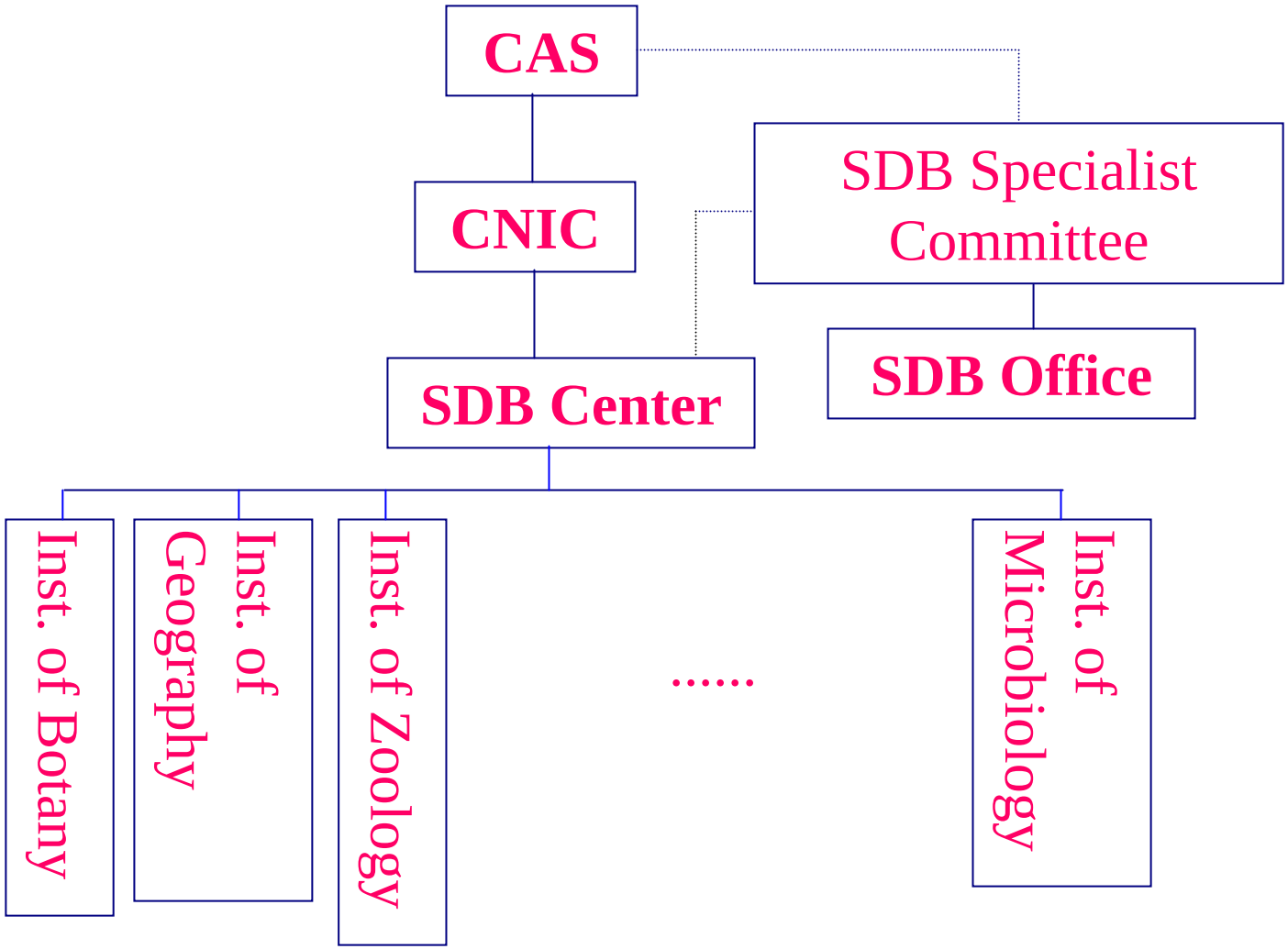
- 人畜共患病之不完全统计 (12-21)
- 台湾今年将建成人畜共通传染病信息库 (08-22)
- 日本专家：吃经过加工的鸡肉不会感染禽流感 (03-01)
- 解读禽流感：禽流感溯源 (02-27)

6. Cooperation & Communication

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- CODATA
  - Secretariat of China CODATA
  - Scientific data & database development and sharing

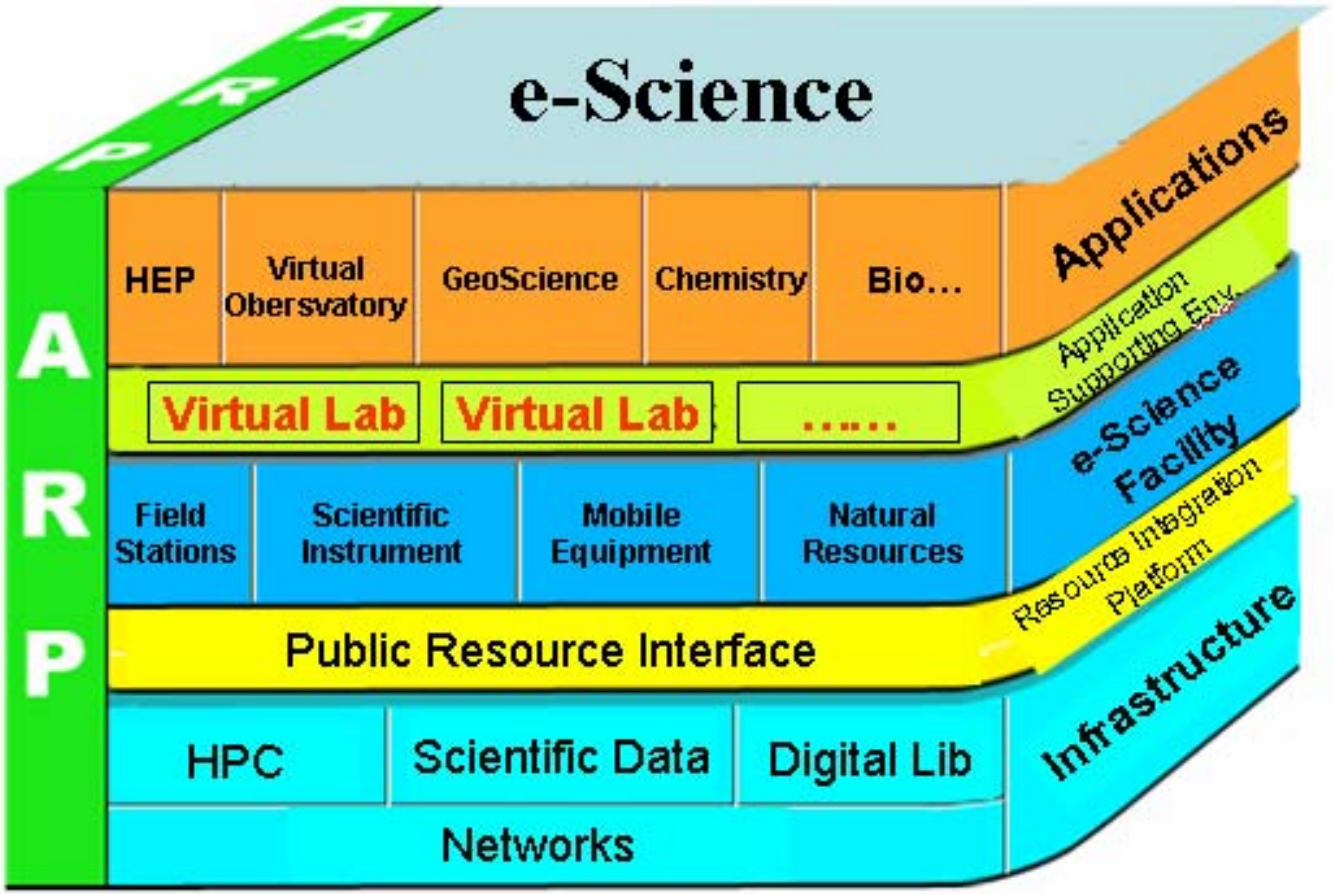
7.SDB Organization chart



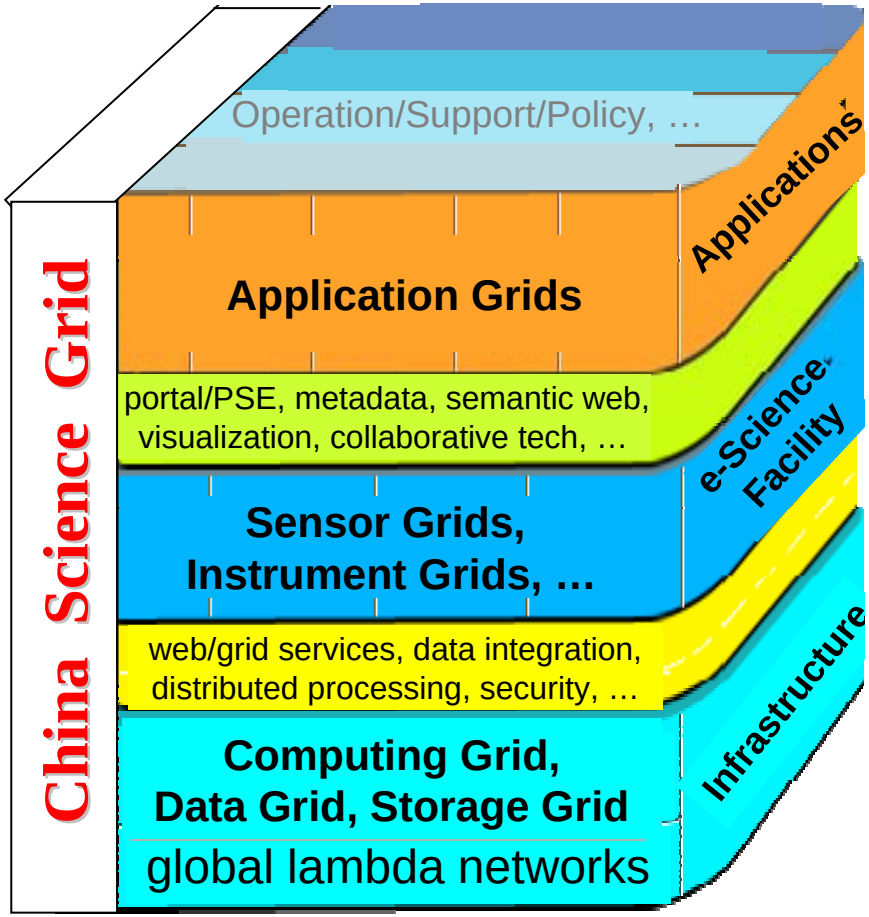
SDB in 2006 - 2010

"SDB Driving e-Science of CAS"

Framework of CAS e-Science



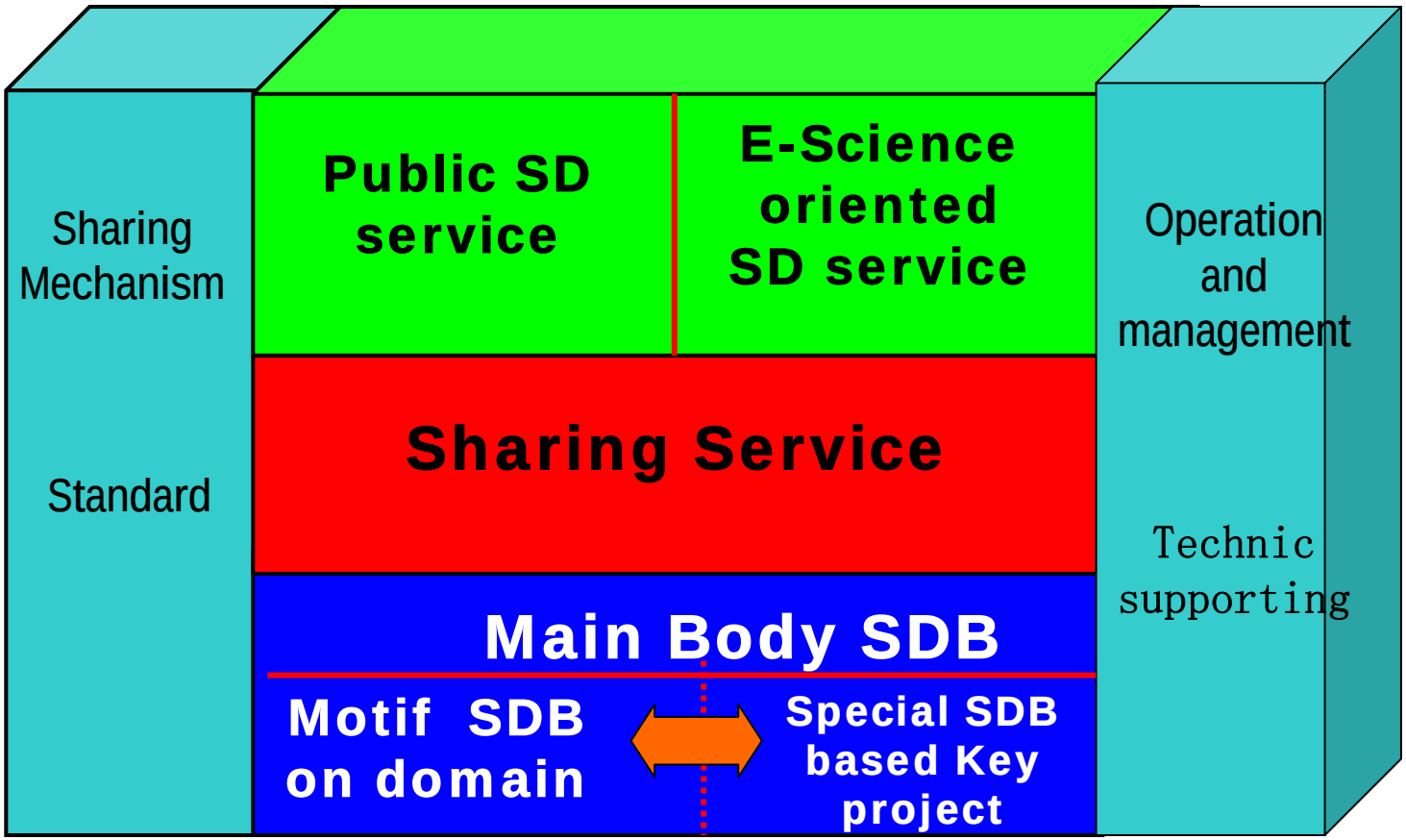
Technical View of CAS e-Science -- China Science Grid



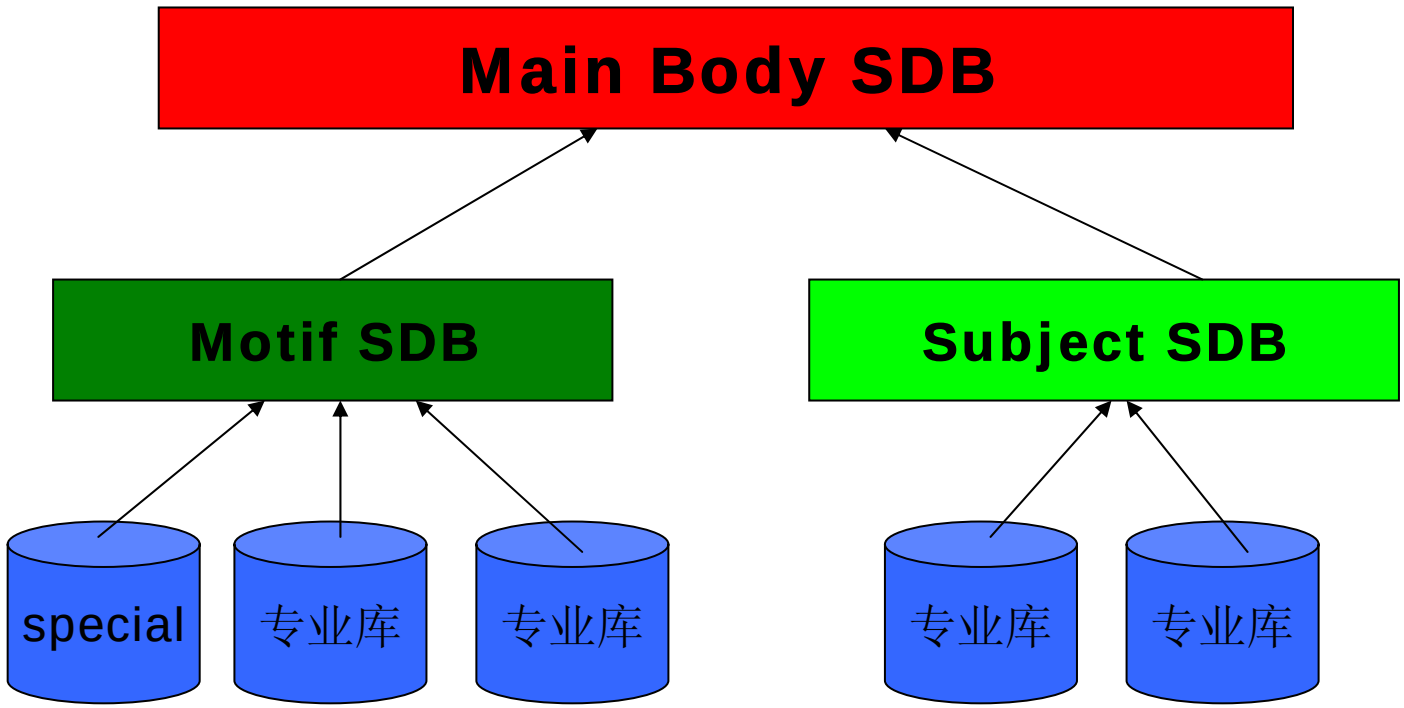
- **Grid-oriented**
- Open
- Sharing
- Collaboration and Virtual Organization
- Security

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

SDB Architecture



SDB Resource Architecture



Main Tasks on SDB

- 60 motif SDBs, 600 special SDBs，60TB sharing
- Continuing standard
- Platform for sharing service
- Platform for running: 300TB disk, 2-3PB tape, parallel wall visualization based LCD, software, ….
- Pilot applications

summary

- SDB is a key foundation for e-Science of CAS
- New challenges:
  - Data technic, data engineering, data science
- Data producing, data management, data service, data using
  - Data quality and maturity
  - Data security
  - Data Policy: sharing and property right, ….
- Drive pilot applications
- Sharing and international cooperation

Thank You !