e-Science & Scientific Data Grid

Baoping Yan (CNIC, CAS)
Kai Nan (NTARL of CNIC, CAS)

CIAPR4  May 21, 2004
Agenda

• Science Activities in CAS
• CAS e-Science
• An exploration towards e-Science -- Scientific Data Grid
• Conclusion
Science Activities in CAS
A Glance at CAS

• Conduct R&D related to almost all aspects of science and technology
• Now about 100 institutes, 1 univ., 2 colleges
• Total staff: 51,000; 37,000 scientists & engineers
• Now a total of 25,000 graduate students
• Around 400 S&T companies and enterprises
The Mission of CAS

- To provide sound scientific and technological solution to social and economic development in China
- To make innovative contribution to the advancement of human civilization.
Disciplinary layout of CAS

- Information Science & Technology
- Life Science & Biotechnology
- Physical Science & New Materials
- Energy Science & Technology
- Marine Science & Technology
- Resource, Environment & Ecology
- Mathematics, Mechanics & Systems Science
- Astronomy, Space Science & Technology
- Research for the Developmental history and Policy of Science & Technology
85 institutes in 24 cities
Nano Science

- (Under) Nanometric metal material manufacturing and super elasticity under normal temperature
- (Right) Long carbon nano tube array
1% sequencing of the International Human Genome with USA, UK, France, Japan, Germany

The first physical mapping of rice genome was successfully completed.
Discovery of bird origin

The most important evidence of the bird origin

*Confuciusornis* 孔子鸟  
*Caudipteryx dongi* 董氏尾羽龙
Different types of robots

high-tech R & D

LBO crystal device

sparkling crystal material
New cotton demonstration in Xinjiang, a new system of cotton cultivation, setting a record of 3.8 ton/ha. in 1999
Soil and water conservation

Loess Plateau
Desertification control
Field survey in Tibetan Plateau
Remote sensing application on flood control (Boyang Lake, Aug 1998)
南极地磁台

长城（1984-5）
62.2S, 59.0W

中山（1991）
69.4S, 76.4E
中国生态系统研究网络站布局规划

CERN

现有站

增补站

省区界
特殊环境与灾害监测网络布局

- 阿克苏水平衡试验站
- 平凉雷电和雹暴试验站
- 东川泥石流观测研究站
- 动力大地测地中心实验站
- 香河综合观测站
- 天山冰川观测研究试验站
- 格尔木青藏高原综合观测研究站
- 北京遥感试验站
中国科学院大气区域本底观测网络布局规划
CERN 野外站分布图

- Agriculture Sites: 14
- Forest Sites: 9
- Grassland Sites: 2
- Waterbody Sites: 4
- Desert Sites: 3
- Wetland Sites: 1
禹城综合试验站

1993年国家科技进步特等奖：黄淮海平原中低产地区综合治理的研究与开发
建站前的沙坡头

1950's SPT

建于1956年，是我院最早建立的长期定位综合野外试验站之一。
沙坡头沙漠试验站
优化模式走向世界
贡嘎山站--长江上游森林植被恢复研究
三江 湿地站--湿地农田生态系统演变

Mire Observation field
海北高寒草地站

- 高寒草甸对全球气候变化的贡献与响应
太湖站
安塞 黄土高原农业生态站

Experiment of water & soil Conservation
红壤站—水土流失保护试验

Red Soil Station

Compare Experiment plot of water and soil Conservation-Erosion
CAS e-Science
What's e-Science

- e-Science – Informatization of Research Activity
e-Science System

Field, Equipment

Data Collecting

Data Processing

Data Sharing

Data Storage

Data Application

Data Service

Search & Retrieval, Content Management

Computing Facility, Simulation, Software

Storage, Database

Network, Grid, Management, Policy

Report, Text, Graph Tools
Background of CAS e-Science

• CAS launched Knowledge Innovation Program in 1998, it’s time NOW to push it forward in all aspects.
• Scientists demand a higher level Informatization to meet their requirements in research activities.
• Informatization will make great effects on promotion of technology innovation and knowledge innovation.
• CAS starts up the Informatization Program in the 10th Five-year Plan (2001-2005)
Informatization of Research Activities

• Bridge the gaps of time, space and environment, enable global, cross-discipline, large-scale collaboration between scientists

• Change the way how scientists do research, greatly improve communication and collaboration, advance the development of science and technology

• Informatization of Research Activities is the pioneer of Informatization of the whole society
Infrastructure for e-Science

• Computing resources
• Data resources
• Software resources
• Communication resources
• Human resources
• Scientific Instruments
  – particle accelerators, telescopes, sensors, …
e-Science and Application

- e-Science provides an informatized environment and platform for research
- Individual applications for fields and areas should be developed case by case
- Application is key
Milestones of CAS e-Science

• In 2000, proposed “Informatized Research Environment” in the SDB project
• In March 2001, proposed “Scientific Data Grid”
• In August 2001, the project funded by CAS Informatization Program
• In December 2001, proposed “China Science Grid”
• In October 2002, “Scientific Data Grid” became an application grid of China National Grid (CNGrid), which funded by National Hi-Tech R&D Program (863)
CAS e-Science Activities (2001-2005)

- Upgrading IT Infrastructure
- Constructing Scientific Research Environment
- Developing Key IT Technologies
- Demonstrating Science Applications
Upgrading IT Infrastructure

- Networks
  - CSTNET
    - Domestic links – 155M-2.5G
    - International links – 310M
  - CNGI (China Next Generation Internet)
    - Supported by National Development and Reform Commission
    - 12 GigaPoPs, 2.5-10G links will build by CAS

- Scientific Database
  - 10TB

- Supercomputing Environment
  - 5.5 TFLOPS

- Mass Storage System
  - 100TB

- Visualization Environment
  - SGI Oynx350

Lenovo 6800, Installed at CNIC
Constructing Scientific Research Environment based on Internet

- Network of Field Observatories
  - Ecology network
  - Astronomical Observatories
  - Weather stations
  - Mountain disaster stations
  - .......

- Network of Digital Libraries of Specimen
  - 24 (zoology, botany, fossil, mineral, ...), 80% of the whole country
  - Digital Library of Specimen is starting

- Network of Digital Libraries
  - National Science & Technology Digital Library

- Network of Scientific Instruments
  - LAMOST, BEP-II, Electron Microscopes, ...
Key Technologies

- NGI Technology
  - IPv6/IPv4 Transition
  - Network Measurement
  - IPv6 Root DNS
  - Multicast
  - Hierarchy Network
  - Security
  - ......
- Resource Location & Discovery
- e-Num

- Grid Computing
- Data Grid Middleware
  - Data Integration
  - Grid Information Service
  - Grid Security
- Metadata
- Grid-enabled application
- ......
Grid-enabled Applications

- Scientific Data Grid (SDG)
- Experimental Applications
  - Virtual Observatory
  - HEP Data Grid
  - Bio Grid
  - Chemical Integrated Information System
  - Geoscience Virtual Research Environment
  - ……
China Science Grid

• By 2005, “Scientific Data Grid” will be built. Sharing of scientific data resources and collaboration based on it are achieved.
• Then, computing resources and scientific instruments will be integrated into. China Science Grid will be built on the SDG.
• Also, develop grid-enabled applications and establish application grids: bio grid, astro grid, … etc.
• China Science Grid – an instance of e-Science
e-Science Planning in Future

• Started to plan 11th Five-year Informatization Program (2006-2010)
  – Focus on e-Science in CAS
  – Work with CNGI (China Next Generation Internet)
  – International Collaboration
    • PRAGMA
    • GLORIAD
    • APAN…

• Potential Killer Science Applications
  – Virtual Observatory
  – High Energy Physics
  – Bioinformatics…
Scientific Data Grid
Overview

• Scientific Data Grid (SDG) is built upon the mass scientific data resources of Scientific Database (SDB).

• SDB is a long-term project since 1983, in which there are multi-disciplinary scientific data accumulated through the course of science activities in CAS.

• The vision of SDG is to take valuable data resources into full play by benefiting from advanced information technologies, in particular, the Grid technology.
Data Resources

- **Scientific Database (SDB)**
  - 45 institutions across 16 cities
  - 313 databases
  - 8.2TB total volume
  - Cover a lot of disciplines

- Chemistry, Biology, Geosciences, Environment, Astronomy, High energy physics, …
SDG Platform

• Data Center
  – 20TB SAN Storage
  – TFLOPS-scale computing capacity

Lenovo
DeepComp
6800
SDG Middleware Architecture

- Application
- Grid API
- Data Res. Broker
- Uniform Access Int.
- Local Data System
- Info. Service
- Security System

- coordinated access to multiple data resources
- uniform access interface to single data resource
- local data management system, could be DBMS or file system
- databases

- applications
- app-oriented, unified program interface
SDG Software Modules
SDG Middleware

- Grid Information System
- SDG Uniform Access Interface
- SDG Security System
- SDG Toolkits

SDG GIS 1.0
Universal Metadata Tool 2.0
Statistics Tool 1.1
SDG Middleware

- Grid Information System
- **SDG Uniform Access Interface**
- SDG Security System
- SDG Toolkits

Data Access Subsystem 1.0
SDG Middleware

- Grid Information System
- SDG Uniform Access Interface
- **SDG Security System**
- SDG Toolkits

SDG CA 1.0
Access Control Toolkit 1.1
SDG Middleware

• Grid Information System
• SDG Uniform Access Interface
• SDG Security System
• SDG Toolkits

SDG Portal (prototype)
Image Process Tool 1.0
Storage Sharing Service (alpha)
Pilot Applications

- Virtual Observatory
- High Energy Physics
- Global Climate Data Integration
- Bioinformatics Integration
- Resources and Environment Monitoring
China Virtual Observatory Demo
Scientific Data Grid (SDG) is an application grid that focuses on scientific data resources sharing and collaboration. Its goals include connecting and integrating data resources in more than 40 institutes under Chinese Academy of Sciences, enhancing effective sharing of distributed and heterogeneous data in Scientific Database.

- Normal Service
- Abnormal Service
- No Service

Refresh on the hour
Information Service

Universal Metadata Management Tool

Universal Metadata Management Tool is provided to facilitate data providers to register the metadata information of their data resources into metadata directory, which is the basis of data access. This tool supports the operations of addition, insert, modification and query. The web interface can be created dynamically according to different xml schemas.

Statistics and Analysis Tool (SAT)

Statistics and Analysis Tool (SAT) has been developed as a prototype implementation of SDG middleware architecture. The volume of data resource is an important attribute for database managers and users. Because the data volume is dynamic, the attribute value stored in the metadata catalog should keep consistent with the actual data volume in order to provide correct information to users. SAT adopts Globus Toolkit 3 (GT3) as the hosting environment and currently all the SAT services are deployed as grid services in Globus container.

The latest stable release of SAT contains installation packages for Windows and Linux platforms. Currently, SAT...
### 元数据工具个性化定制

#### 查询结果显示界面设定

#### 元数据添加

#### 网格信息服务系统

通用元数据管理工具

首页 | 选择所有元数据 | 用户密码修改 | 单位内部查询元数据 | 添加元数据 | 设定界面风格

### 查询结果

#### 共找到相关记录 2 条

<table>
<thead>
<tr>
<th>ID</th>
<th>部门</th>
<th>数据集名称</th>
<th>描述</th>
<th>数据集来源</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>开发部</td>
<td>国外纳米专利数据库</td>
<td>国外纳米专利库的数据收集了从1986年至2001年12月，主要来自...</td>
<td>计算机网络信息中心</td>
</tr>
<tr>
<td>2</td>
<td>信息部</td>
<td>古代名人库</td>
<td>古代名人：包括白居易，邓世昌等</td>
<td>计算机网络信息中心</td>
</tr>
</tbody>
</table>

技术支持：manager@satl.ac.cn  推荐分辨率：1024*768
版权所有：中国科学院计算机网络信息中心 Copyright 1995-2003
Statistics & Analysis Tool (SAT) for Data Volume

• Features
  – Win2000/XP, Linux
  – Java 1.4
  – Globus Toolkit 3 Core
  – Oracle, SQL Server, Flat files

• Deploy
  – Data nodes: 45 institutes at CAS, across 16 cities in China
    • Normal Service: 13 nodes
    • Abnormal Service: 11 nodes
    • No Service: 21 nodes
  – Mediator: CNIC
  – Service Monitor
科学数据网格门户

数据访问

化学品毒性数据库

化学品毒性数据库是中科院计算机网络信息中心承担建设的综合科技信息数据库的重要组成部分，是中科院知识创新工程信息化建设的重大专项。

化工产品已经深入人类生活，社会经济活动中的生产、经营、贮存、运输、使用、报废等各个环节都涉及到成千上万的化学物质，国家安全的品质也与化学物质息息相关。化学品毒性数据库的内容是关于化学物质的安全管理、化学事故的预防和化学灾害的应急响应等物理、化学的特性数据（如易燃性、易爆性、毒性、环境标准和作业标准等数据），可广泛应用于军队、化工、医药、卫生、农业、环境、环境管理和化学生科等领域。

目前，化学品毒性数据库包含“化学品安全特性数据库”、“化学品毒性效应数据库”和“RTECS数据库（英文版）”三个子库，数据库约600MB、60,000多条记录，用户可以通过专用的域名www.toxic.csdb.cn访问化学品毒性数据库的内容。

纳米科技基础数据库

纳米科技基础数据库是中科院计算机网络信息中心承担建设的综合科技信息数据库的重要组成部分，由中科院纳米科技中心合作建设。纳米科学是21世纪高技术的标志性领域，纳米科技基础数据库自2000年5月开始策划实施，以实现纳米数据的网络化管理和数据化信息共享，为我国纳米科技工作者提供权威及时的纳米科技数据服务，支持国家纳米科技创新。
# 纳米科技基础数据库

## 当前查询库：中国纳米专利公开库

<table>
<thead>
<tr>
<th>申请号</th>
<th>发明名称</th>
<th>申请人</th>
<th>国家/省市</th>
<th>国际公布号</th>
<th>操作</th>
</tr>
</thead>
<tbody>
<tr>
<td>87103208</td>
<td>用黄光预<strong>光</strong>增加感光高分子的感光速率</td>
<td>纳米尔杜邦公司</td>
<td>美国</td>
<td>9503317</td>
<td>详细信息</td>
</tr>
<tr>
<td>90100083</td>
<td>处理薄膜的方法</td>
<td>罗姆和哈斯公司</td>
<td>美国</td>
<td>9503317</td>
<td>详细信息</td>
</tr>
<tr>
<td>80109547</td>
<td>含有吸收紫外光<strong>光<strong><strong>光</strong></strong>光</strong>的<strong>光<strong><strong>光</strong></strong>光</strong></td>
<td>伊斯曼科达公司</td>
<td>美国</td>
<td>9503317</td>
<td>详细信息</td>
</tr>
</tbody>
</table>
科学数据网格CA(Certificate Authority)的介绍

科学数据网格的用户是通过证书来标识身份，科学数据网格的证书是采用X.509证书格式编码。科学数据网格CA是针对证书进行管理。

- 建设CA的目的

为保证科学数据网格的安全性，提供身份认证服务，提供信息保密性，数据完整性和收发数据的不可否认性服务。

- CA的主要功能

证书申请：代替用户生成私钥对，同时生成证书请求文件，从安全角度考虑，秘钥对最好由用户来生成。

证书颁发：CA为科学数据网格用户颁发证书。

证书申请：该功能是核查用户的证书申请是否符合规范，主要由证书的注册机构RA(Registrar Authority)来实现。

证书恢复：因为证书都有使用期限，所以对损坏或过期的证书，要重新给用户颁发新的证书。

证书撤销：CA对所有废止的证书要实时地公布于众，便于用户查询，安全地对证书进行
Deployment

- SDB Technical Training
  - Sep. 23-27, 2003, Qingdao
  - 120+ persons from 45 institutes at CAS

- SDB Work Evaluation Online
  - Nov. 2003
  - Deployment of SDG middleware
Distance Training with UCSD/SDSC

- February 20th, 2004
- Longjiang Ding, UCSD/SDSC
- On Grid technology
- By Virtual Room Videoconference System (VRVS)
- The first distance training experiment
Conclusion
Conclusion

• Today and tomorrow’s research demands global collaboration – e-Science.
• The progress of Information Technology make this possible.
• CAS is making great efforts on e-Science with its Informatization Program in the 10th Five-year Plan.
• The e-Science Program in the 11th Five-year Plan (2006-2010) is being worked out. E-Science will become the groundwork of research in the future five years.
• Scientific Data Grid is the first experimental project for CAS e-Science. A few of science applications on SDG would be our exploration towards e-Science.
Thank you!