Requirements and Architecture for Data Grid Middleware

Kai Nan, Deting Yang

Computer Network Information Center
Chinese Academy of Sciences

PNC 2003 Bangkok

Outline

- Data Grid
- Common Requirements for Data Grid Middleware
- Experiences on SDB
- Design for Architecture of SDG
- Progress Update

Data Grid

- Grid
  - resource sharing
  - collaborative problem-solving
- Data Grid
  - more focus on data
  - (scientific) data become one footstone of modern sciences and research
  - data sharing is crucial to most scientists today

Outline

- Data Grid
- Common Requirements for Data Grid Middleware
- Experiences on SDB
- Design for Architecture of SDG
- Progress Update

Requirements towards Data Grid Middleware

- Identification
- Provenance
- Metadata
  - technical / context / content / management
- Access Control
- Universal Access Interface
- Publishing / Discovery / Retrieval
- Data Lifecycle
- …

Simplified 3 Steps

- find the data
  - and get related info. (metadata)
- obtain proper rights towards the data
- access the data
  - maybe multiple distributed and heterogeneous databases involved within one request
  - maybe not just data, but processing and/or analysis

- these steps seem to be easy, but …

Grid Information Service

- Step 1-- To find the data
- Requirements
  - Define metadata schema
    - resource discovery
      - answer to "*What, How*" – intrinsic properties of data
      - relatively static metadata, generated by man
    - location & monitoring
      - answer to "*Where, When*" – extrinsic properties of data
      - dynamic information, generated by program
  - Define API
    - Publish / Collect
    - Query

Grid Security System

- Step 2 -- To ensure that data be accessed rightly
- Requirements
  - Single Sign-On
  - Delegation
  - Universal credentials
  - **Integration with local policies**
  - **Policy management**
  - **Data-oriented access control**
  - User-based trust/trusteeship
  - Logging
  - Open architecture & Interoperability with other Grids

Uniform Data Access

- Step 3 -- To get the data easily
- Requirements
  - **Uniform access interface to single data resource**
  - **Coordinated access to multiple data resources**
  - **App-oriented, unified and convenient program interface**
  - Schedule policy
  - Data replication
  - Data quality assurance

Outline

- Data Grid
- Common Requirements for Data Grid Middleware
- Experiences on SDB
- Design for Architecture of SDG
- Progress Update

Our Experiences on SDB

- SDB – Scientific Database
  - a project funded by CAS since 1986
  - a collection of scientific databases, which cover multiple disciplines including chemistry, biology, geography, astronomy, ecology, …
- By now, SDB has
  - 45 member institutions across China
  - 296 databases
  - data volume 8.2TB

SDB Characteristics & Challenges

- Characteristics
  - Distributed
  - Heterogeneous
- Challenges
  - Requirements for data sharing
  - More collaborative work across multi-sites and multi-disciplines
  - More collaborations with colleagues across the world under Knowledge Innovation Program of CAS
  - The data are from research, and for research.

Data Grid !

Outline

- Data Grid
- Common Requirements for Data Grid Middleware
- Experiences on SDB
- Design for Architecture of SDG
- Progress Update

Scientific Data Grid (SDG)

- one-sentence statement
  - a grid which focuses on sharing multi-discipline scientific data and advancing cooperative research based on the utilization of scientific data
- more words
  - built upon the Scientific Database (SDB) of CAS
  - started in 2001
  - plan to provide service by 2004-2005
  - for academic and research
  - built by CAS, open to the world

SDG Vision

- Resource Level – sharing and development
  - make scientific data more accessible
  - data integration
  - data → information → knowledge
- App Level – enabling e-Science applications
  - complex problem-resolving with heavy use of data
  - cross multiple databases / cross-disciplinary
  - demand more resources (cycle, storage, bandwidth, instrument, sensor, …)

SDG Middleware

applications

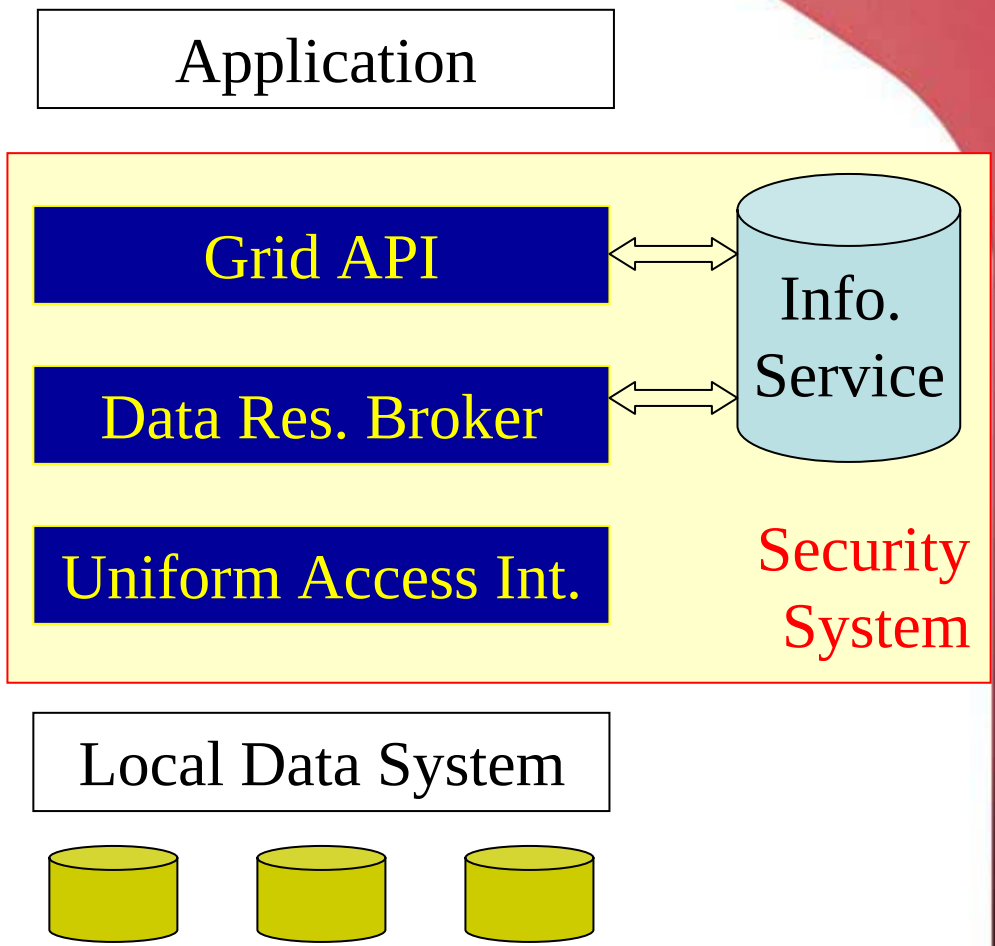
app-oriented, unified program interface

coordinated access to multiple data resources

uniform access interface to single data resource

local data management system, could be DBMS or file system

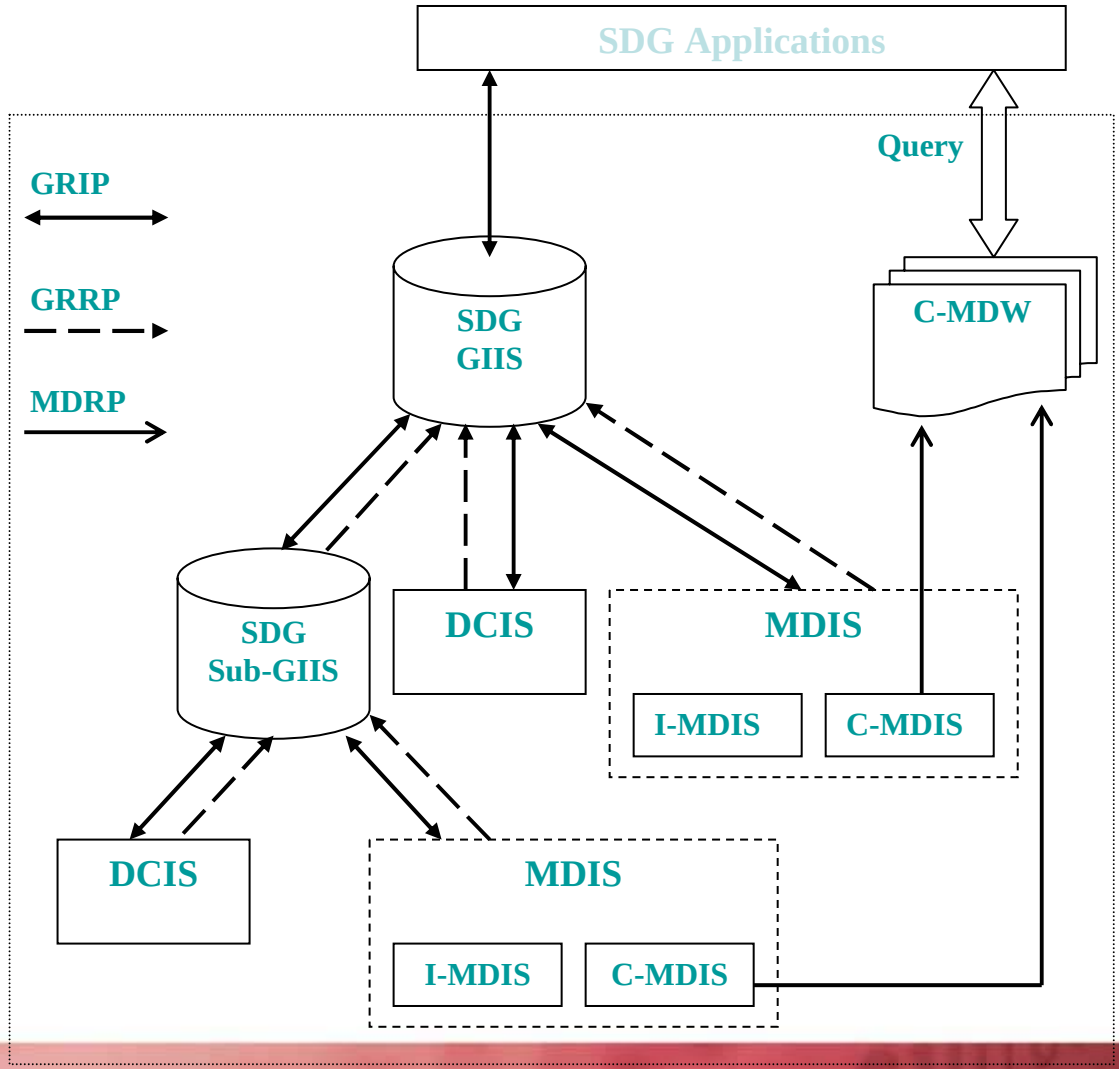
databases



SDG Information Service

- SDG Info. Service
  - DCIS: Data Container Info. Service
    - built on Globus MDS
    - design DIT for SDG (schema, OID, namespace)
    - develop a program which collects information and returns it as LDIF, called *info. provider*
    - configure a new MDS
  - MDIS: MetaData Info. Service
    - actually a normal LDAP
    - add ldbm-backend to MDS in order to store static metadata
    - develop the *metadata tool* to manage MDIS
  - Compatible with Globus MDS 2.1
  - Future plans: extend the infrastructure with Grid Services

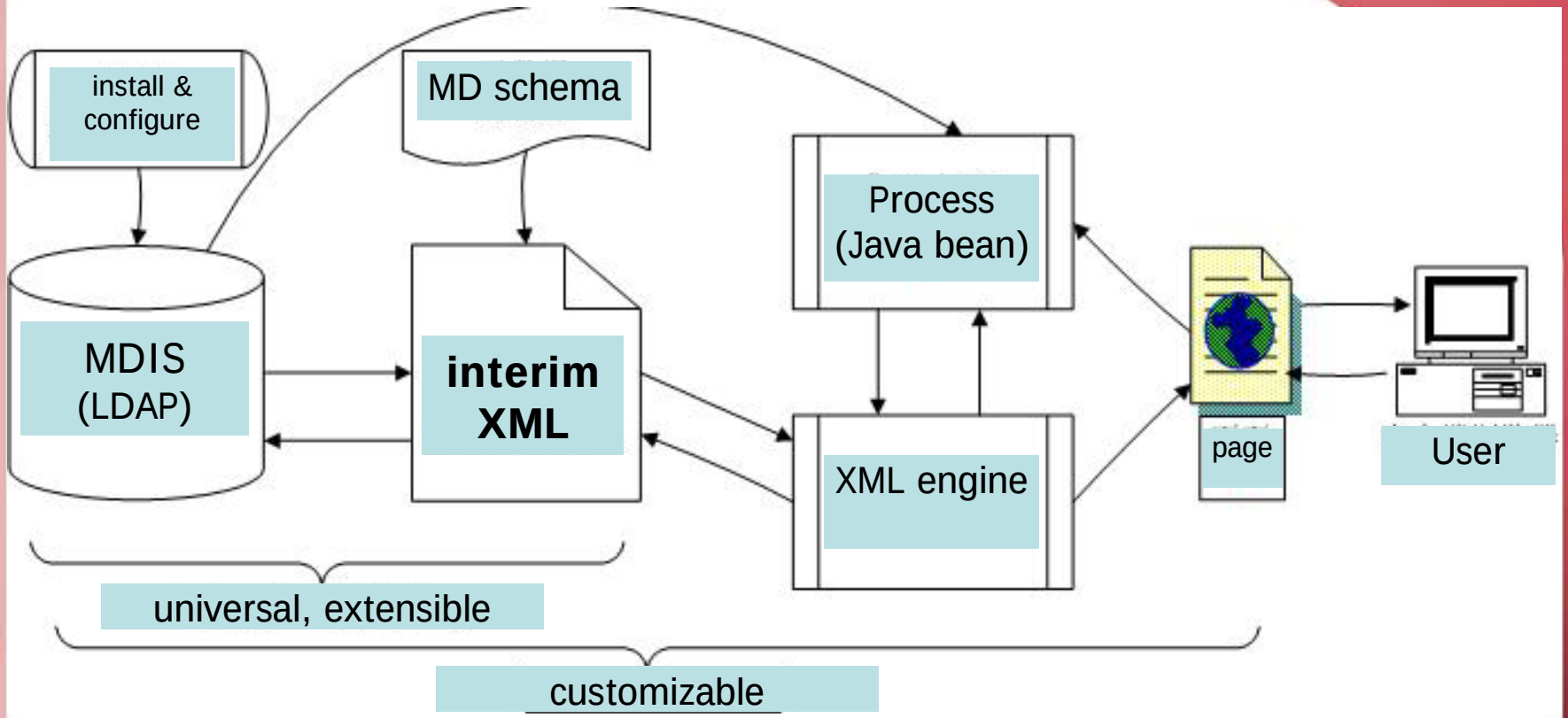
SDG Information Service (cont'd)



SDG Universal Metadata Tool

- Requirements
  - why universal
    - many disciplines in SDG → similarly many or more metadata standards
    - it's not good for us to develop a tool for every metadata schema individually
    - input metadata for existing databases is more bothersome, so an ease-to-use tool might be must-have in practice
  - input: a metadata schema (xml DTD)
  - output:
    - Web-based, customizable UI
    - LDAP-based Storage
    - Management functions (add, delete, modify and query)
  - back-end is MDIS

SDG Universal Metadata Tool



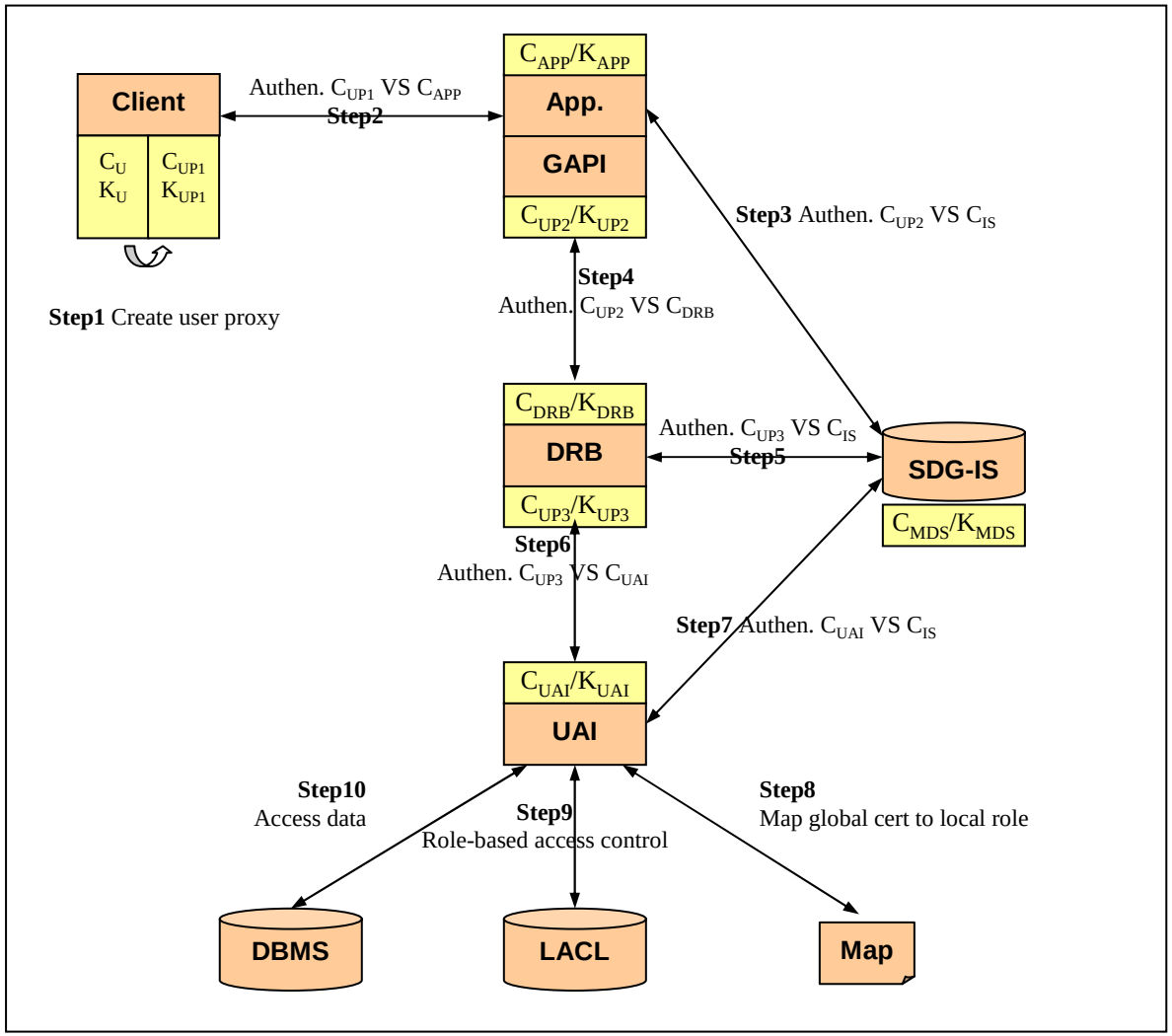
-metadata is tree-like and more flexible than fix-column tables, difficult to deal with on web UI

-use xml files to store interim results

SDG Security System

- Services
  - Authentication (Based on Globus GSI)
    - secure connection
    - user proxy management
  - Authorization
    - mapping global certificates to local roles
    - role-based access control
    - local role management
  - Accounting

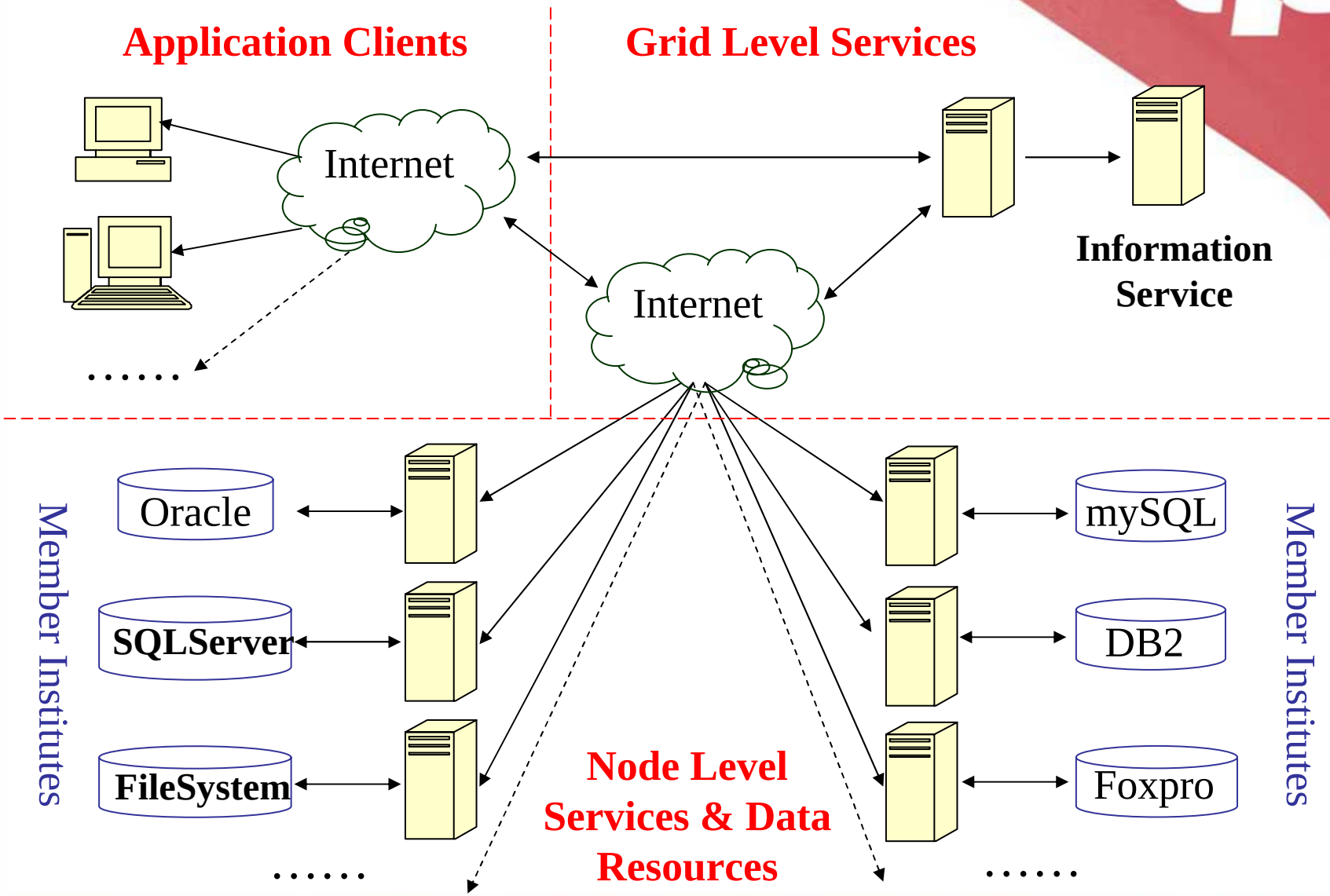
SDG Security System (cont'd)



C_X, K_X	X's Cert & Key
UP1, UP2, …	User Proxy, 2nd-level User Proxy, …

Full Process of security-related operations under SDG Security System

SDG Uniform Access Interface



SDG Uniform Access Interface (cont'd)

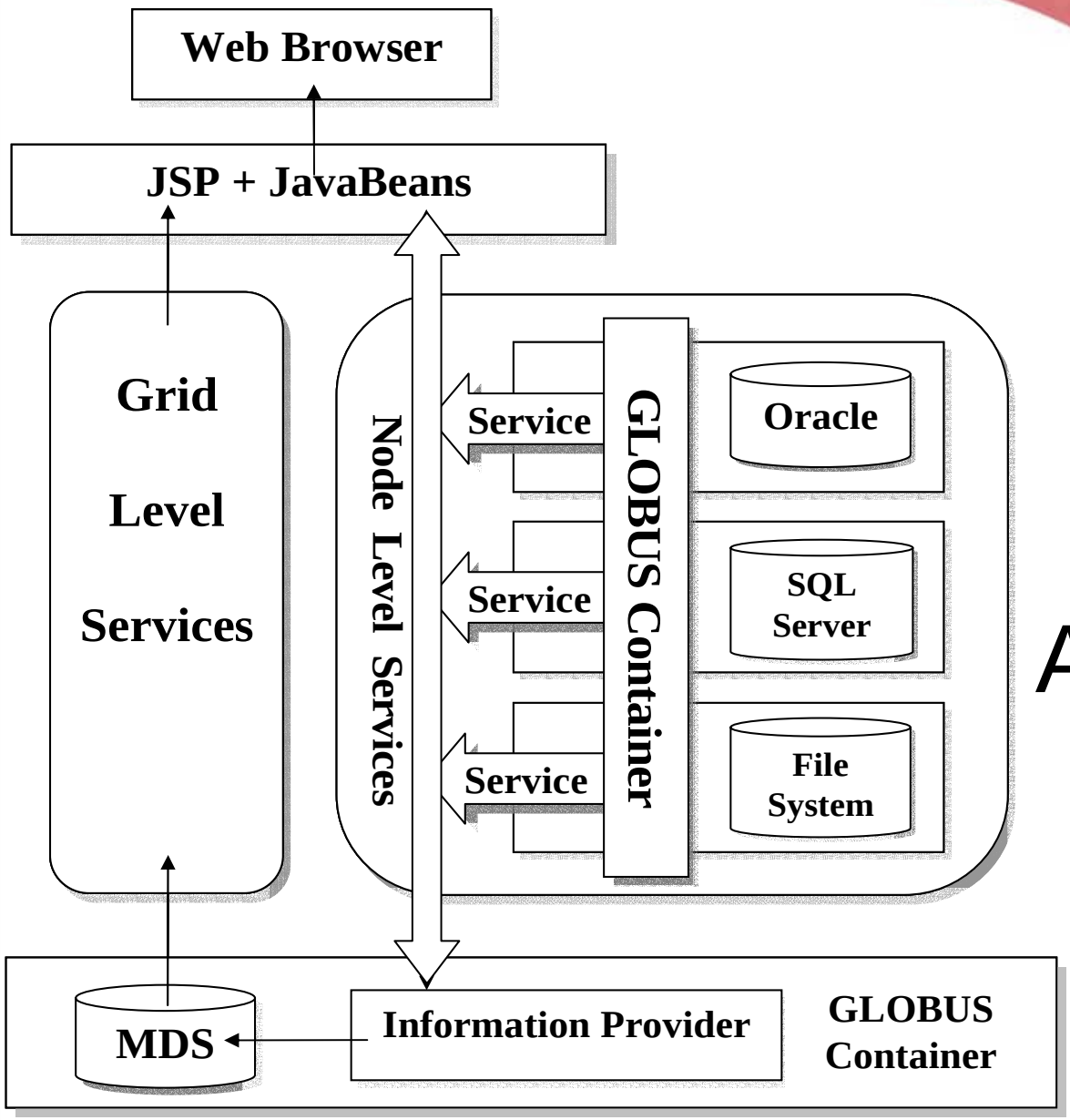
- OGSA-based
- Two Levels Services
  - Node level
    - Data services on single node
  - Grid level
    - Data services cross multiple nodes
- Data services
  - Data Query
  - Data Analysis
  - Data Processing
  - Data Replica
  - ……

Outline

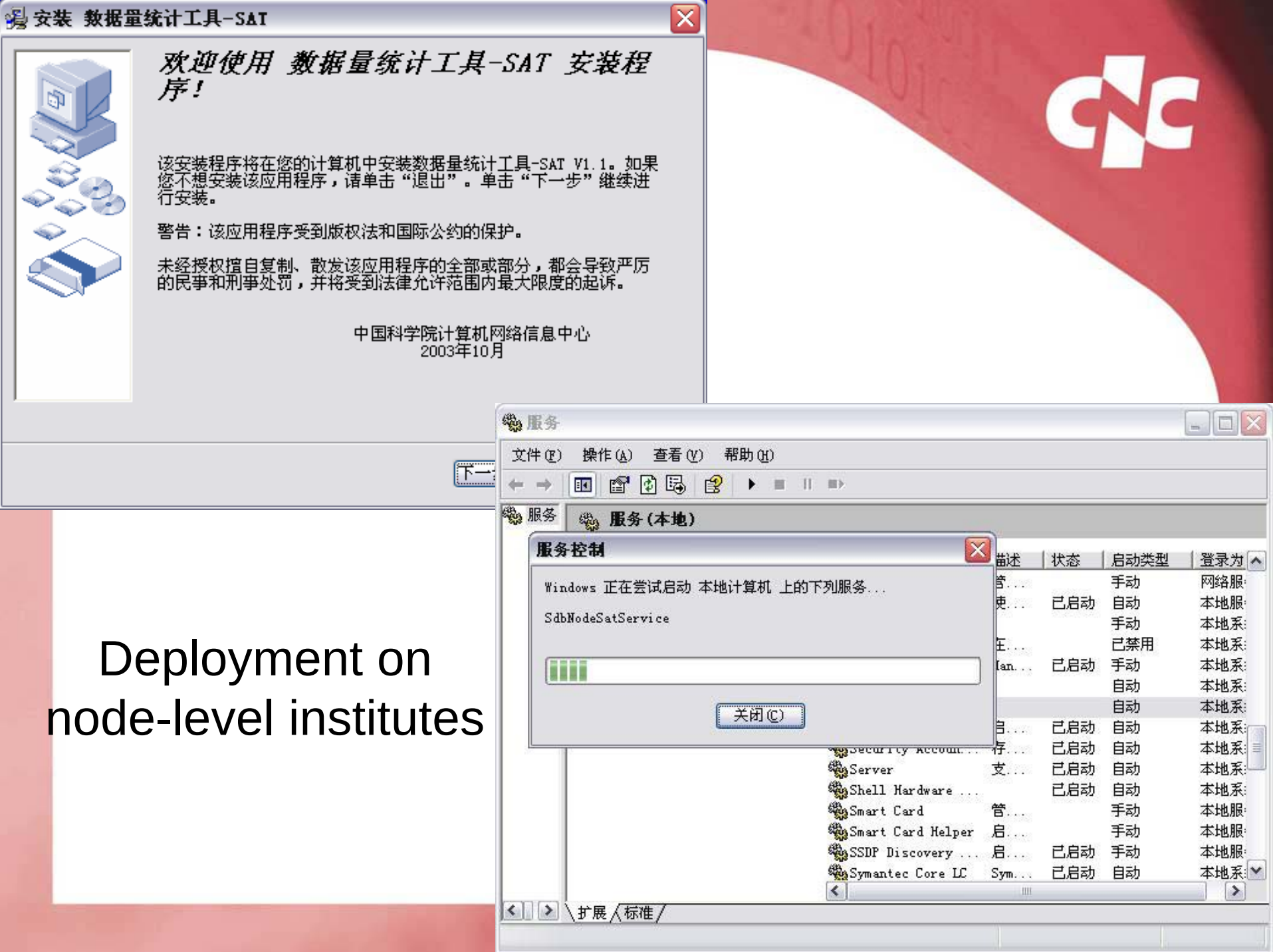
- Data Grid
- Common Requirements for Data Grid Middleware
- Experiences on SDB
- Design for Architecture of SDG
- Progress Update

Progress Update

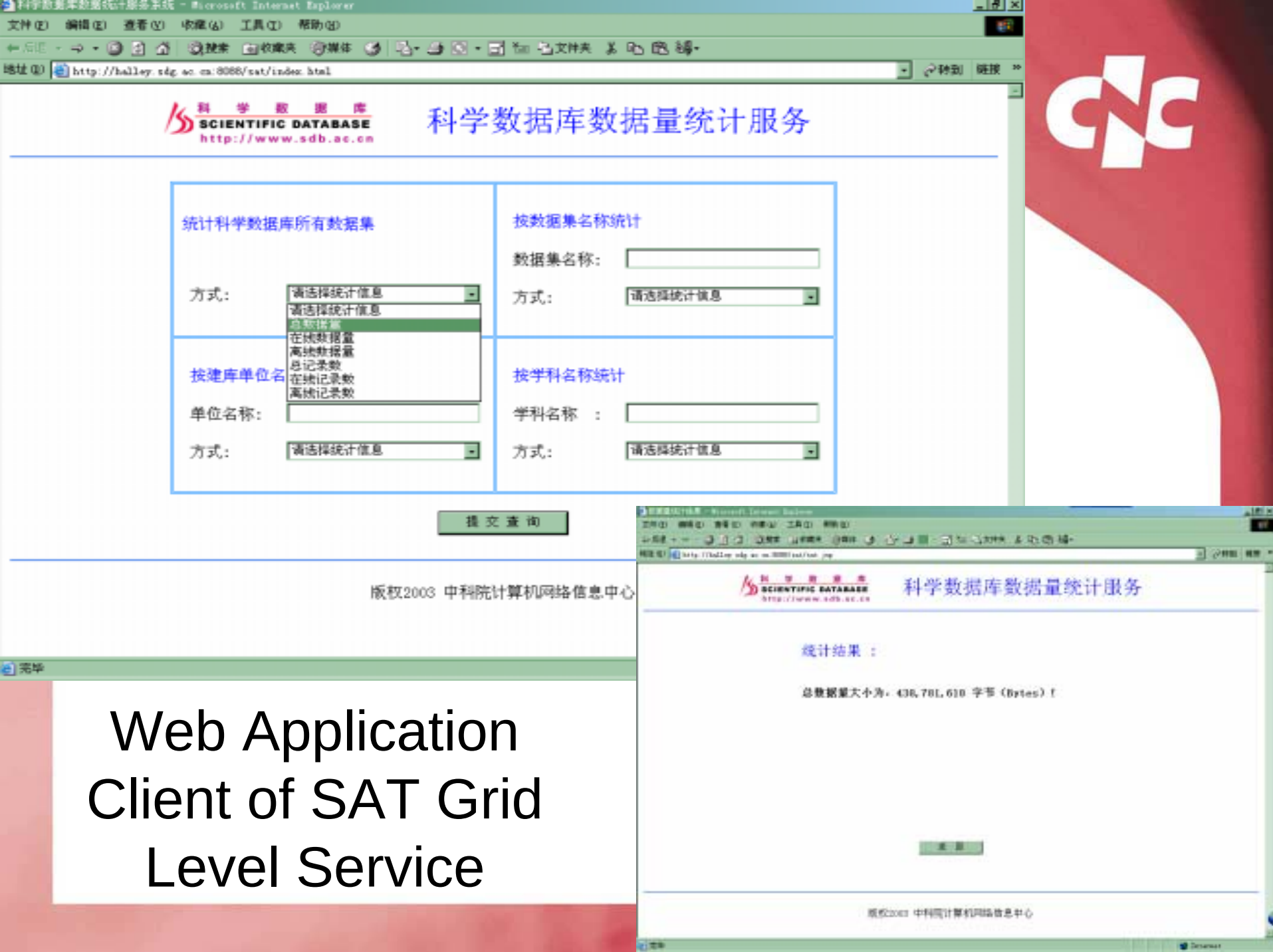
- SDG Middleware Tools and Services
  - Universal Metadata Tool, V2.0
  - Local Access Control Tool, V1.0
  - Certificate Management System, V1.0
  - Statistics Services of Data Volume (SAT), V1.1
  - Image Process Services, V1.0



SAT Architecture



Deployment on node-level institutes


科学数据库
SCIENTIFIC DATABASE
http://www.sdb.ac.cn
科学数据库数据量统计服务
统计科学数据库所有数据集
方式:
请选择统计信息
总数据量
在线数据量
离线数据量
总记录数
在线记录数
离线记录数
按数据集名称统计
数据集名称:
方式:
请选择统计信息
按建库单位名
单位名称:
方式:
请选择统计信息
按学科名称统计
学科名称 :
方式:
请选择统计信息
提 交 查 询
版权2003 中科院计算机网络信息中心
科学数据库数据量统计服务
统计结果 :
总数据量大小为: 436,781,610 字节 (Bytes)!
版权2003 中科院计算机网络信息中心
Web Application
Client of SAT Grid
Level Service


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