清華大學
Tsinghua University

Bioinformatics Grid (BioGrid)

Prof. Weimin Zheng, Dr. Yongwei WU
Tsinghua University
Email: wuyw@tsinghua.edu.cn

Content

- Introduction
  - Why BioGrid?
  - Bioinformatics Research Status of China
  - Funding Programs
- BioGrid middleware
- Benefits of BioGrid for Users
- Current Status and Next Step

清華大學
Tsinghua University

Background of Bioinformatics

- Merges biology, computer science and information technology together;
- Purpose: Grasps biological meaning of plentiful biological data and enable the discovery of new biological insights ;
- Method: Uses database, data processing method and software to get results through mass computation

Bioinformatics—Grid Compliant

- Research Content: Sequence alignment, fragment assembly, sequence analysis, protein structure analysis, signature recognition…
- Divided into multiple sub-tasks which do not/little communicate each other
- Data intensive and little communicated: suitable for the grid computing.

Bioinformatics Research Status in China

1. Many institutes hold their own computing resources distributed over the Internet;

2. Many researchers in PRC universities have no such research resources to use;

3. It will be a very significant work to integrate all these resources together.

清華大學
Tsinghua University

Purposes of BioGrid

1. Gather heterogeneous large-scale computing and storage devices, computing tools and related databases together through grid technology.
2. Provide bioinformatics supercomputing services for bioinformatics researchers through Web interface.
3. Service define tools for Service Providers:

   Workflow and One step Service
4. For end users, Input the computing requests according to submission form, and get the computing result from the BioGrid Portal.

清華大學
Tsinghua University

Funding Programs of BioGrid

- ChinaGrid (MoE and MoST)
  - 2.5M RMB; (2003-2006)
  - Build by 7 top Chinese Universities (Tsinghua, Peiking, HUST, SCUT, XJTU, Shandong, NUDT)
- National Network Computing Environment (Ministry of Science and Technology)
  - 7.7M RMB; (2006-2008)
  - 10 partners including SCBiT, Beijing Genomics Institute……

清華大學
Tsinghua University

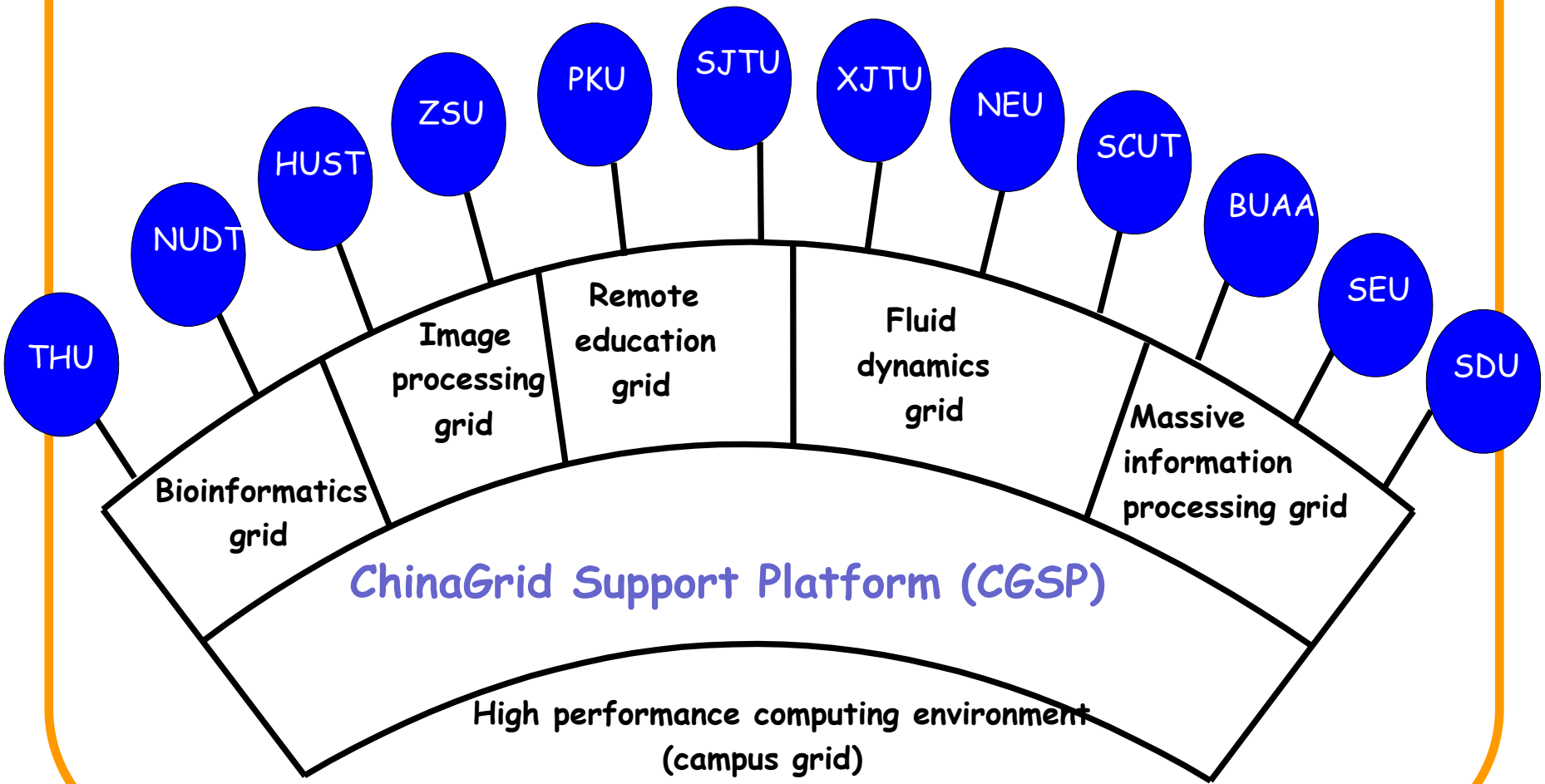
ChinaGrid Program

- As an important 211 project in the Tenth Five-Year Plan Period of Chinese Ministry of Education, ChinaGrid aims at constructing public service system for higher education.
- Based on CERNET (China Education and Research Network)
- First Phase
  - From 2003-2006
  - 12 key universities as initiative
  - 20 key universities now
  - More than 15Tflops and 150TB
  - 5 typical grid application (Bioinformatics, image Processing, Mass Information Processing, CFD, Remote Education)

清華大學
Tsinghua University

ChinaGrid Members (till now)

Layered Infrastructure of ChinaGrid



National Network Computing Environment Program

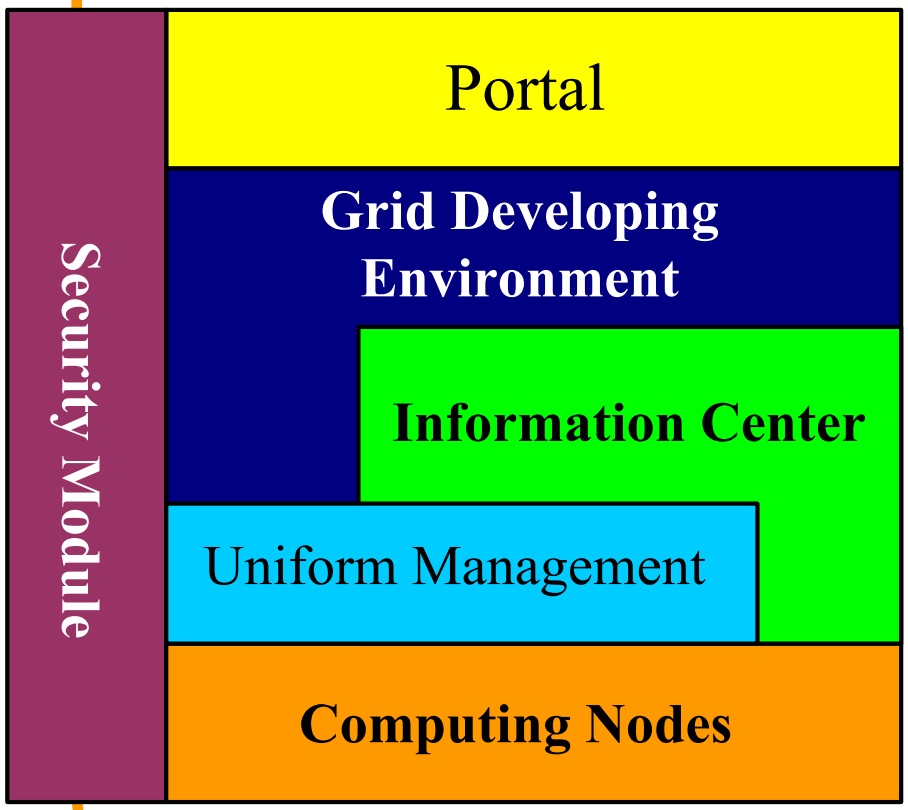
- Purpose: Build public research platform for various discipline and form the national science research platform.
- First phase: 20 platforms for different subjects will be constructed from 2005-2007

Middleware of BioGrid

- ChinaGrid Support Platform (CGSP) is a grid middleware developed for the construction and evolution of ChinaGrid;
- Based on OGSA, CGSP is developed according to the latest grid specification WSRF;
- CGSP supports localized requirement and autonomy requirement of ChinaGrid;
- Scalability of CGSP satisfies the demand of expansion of ChinaGrid.
- CGSP V1.0 is issued on the Jan. 2005.

清華大學
Tsinghua University

5 Function Modules in CGSP



Portal: Grid entry for submitting & monitoring job, querying resources' info, user management and accounting;

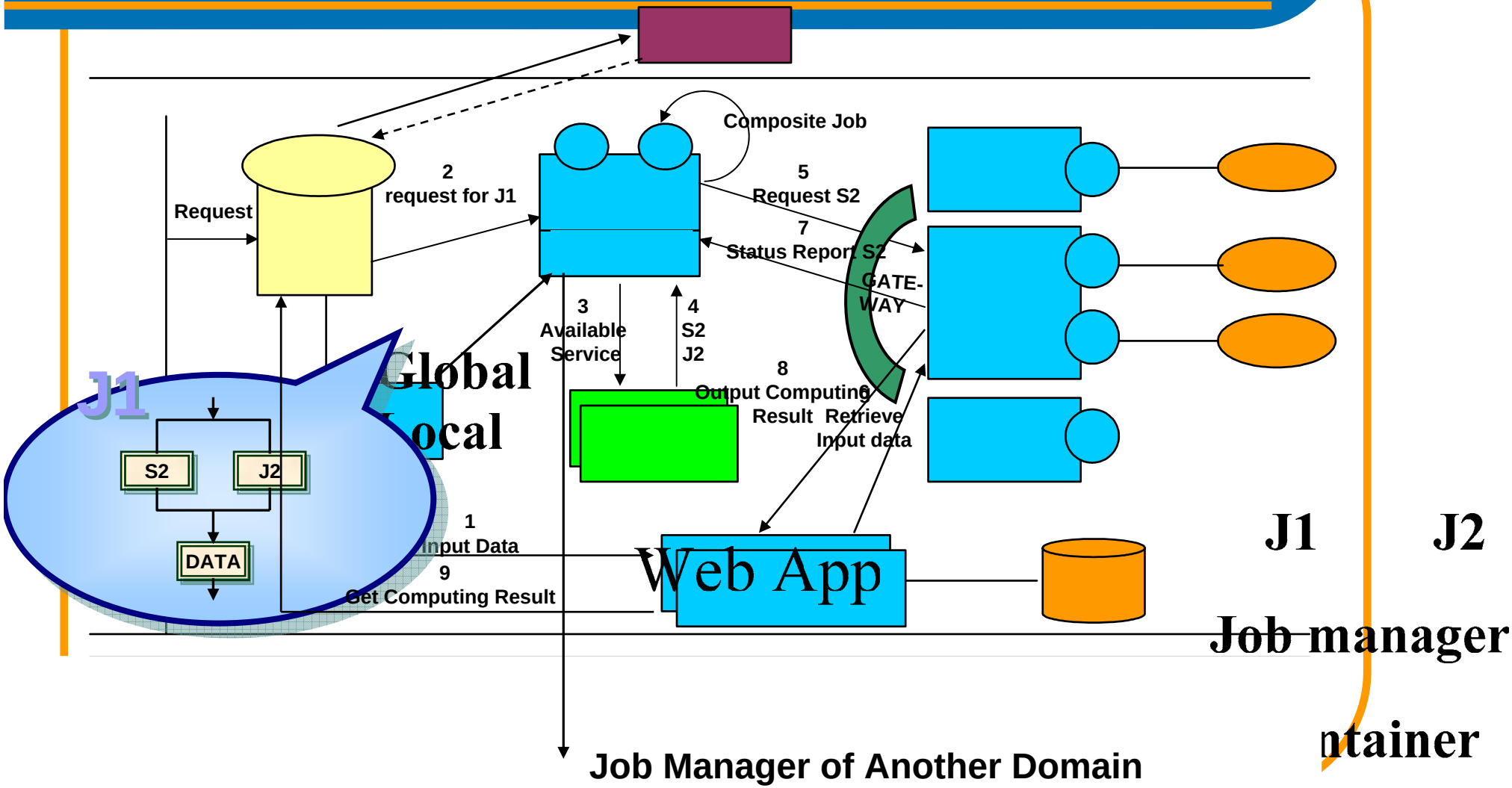
Grid Developing Environment: a set of toolkits including portal development tools, resource encapsulation tools, programming tools and job generating tools etc.

Information Center: the manager of resource & service information

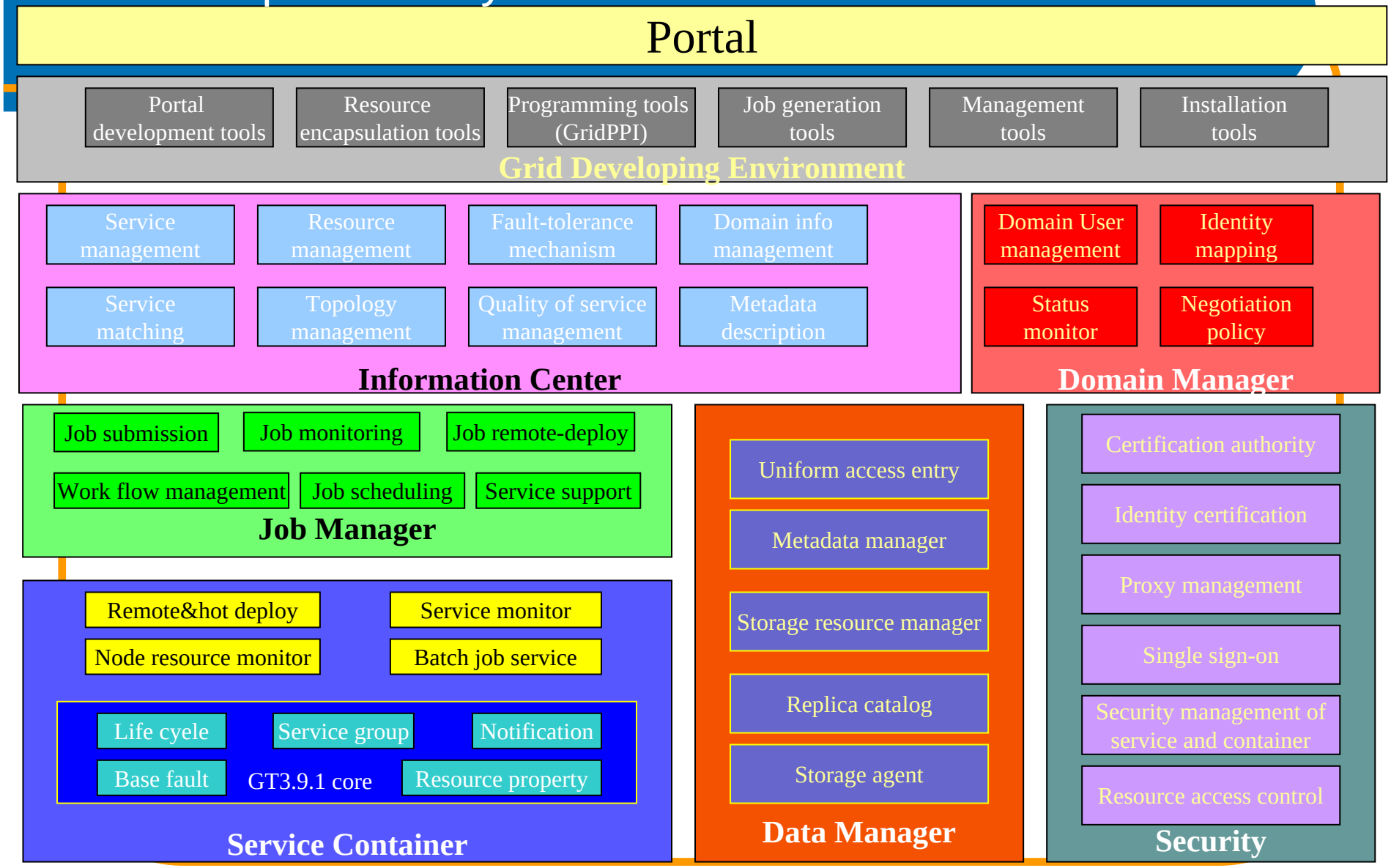
Uniform Management: a set of managers including job manager, data center, domain manager and service container.

Security: Identity authentication and mapping, service and resource authorization, secure transferring.

Running Flow of BioGrid



Functions provided by Each Module in CGSP



Benefits of BioGrid

1. Improve service quality of resources, more quick…

2. Improve utilization of resources, more users…

3. Make bioinformatics supercomputing power Internet accessible easily;

4. Provide coordinated use of heterogeneous computing and research tools.

清華大學
Tsinghua University

Sequence Matching without Grid Computing

1. Search severs providing sequence matching
2. Login server
   - **Study the manual to use the software**
   - **Visit database update time**
3. Submit job, get computation results after looong time
4. Repeat above 3 steps, until get all the possible results from the servers visited
5. Analyze the results, delete the duplicated or abundant results based on database update time (very time consuming)

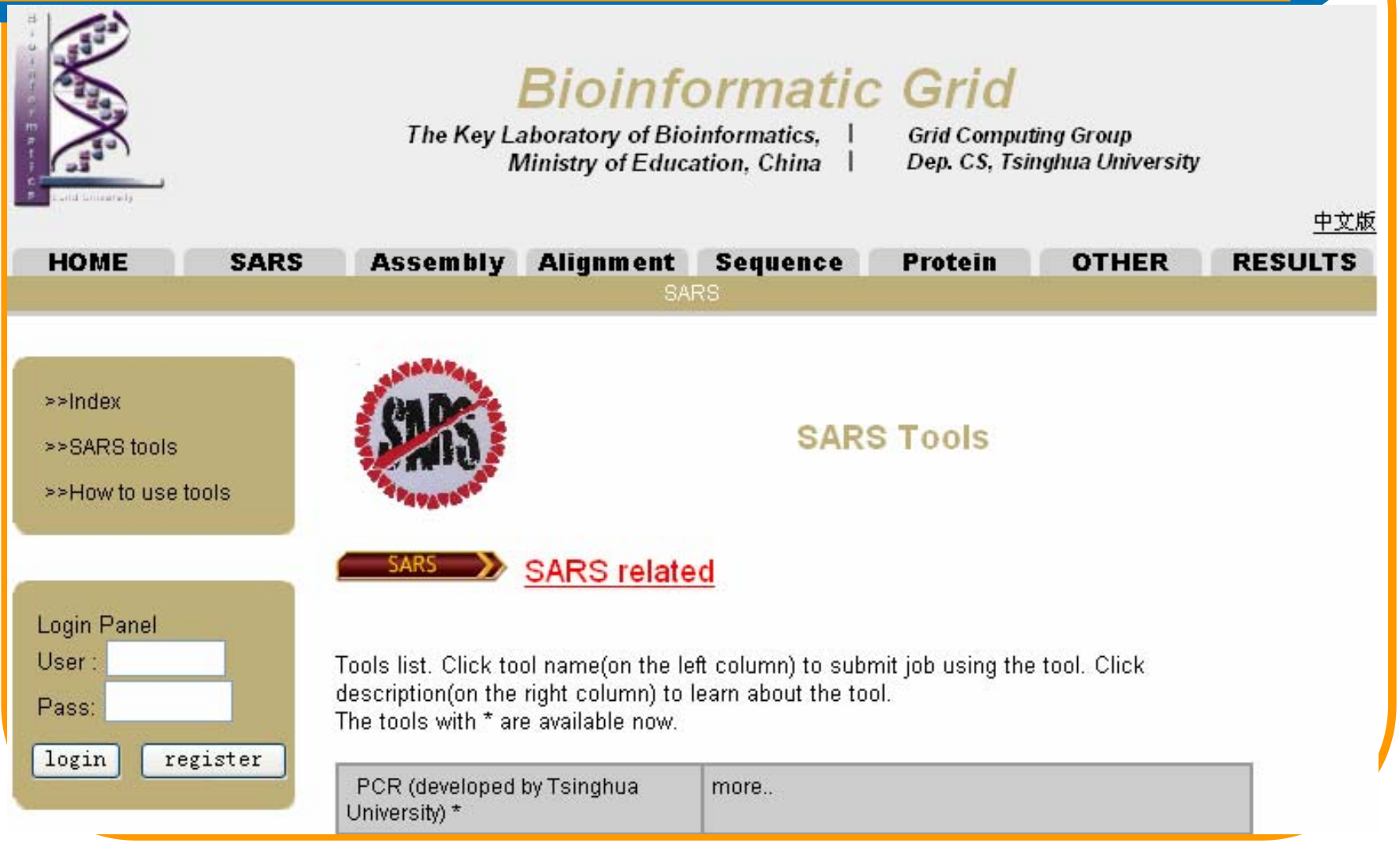
清華大學
Tsinghua University

Sequence Matching with Grid Computing

- Simple usage
  - Submit job via web page
  - Wait for the final computation results
- Automatically update data base and keep synchronization
  - Domestic databases update with USA or Europe databases via CERNET every 0:00am to keep synchronization
  - Domestic databases update each other every 4:00am to keep synchronization
- User can get the unique result without comparing and deleting duplicated and abundant results
- High efficiency
  - Multiple servers perform the sequence matching for different parts of sequence simultaneously
  - Sequence can be partitioned automatically
  - No redundant matching operations

清華大學
Tsinghua University

Portal of BioGrid



Qinghua University

1. 1 Teraflop/s HP Itanium2 cluster
2. 200 gigaflop/s Intel Xe cluster.
3. SUN 10000 Server
4. SGI Origin 2000

PeiKing University

1. SUN Fire 4800
2. 2 CPU XEON Server
3. SUN Fire V880

XI'an Jiao Tong University

1. IBM RS6000 Cluster
2. XEON Cluster

北京大

Summary:

More than 10 super computers from 7 top universities;
Computing power: 2T;
Storage capability: 5T;
Nearly 60 computing tools;
50000 users everyday;

National universit
Defense Technol
Yinhe SuperCom

Shandong University University: LangChao Cluster

versity

1. MPP Linux Cluster
2. SMP Cluster

清华大学

Tsinghua University

生物信

Shared Computing tools (1)

SARS Tools SARS related	PCR PClustal W） Clustal W Protein docking (Gramm) Protein Structural Analysis. (Interproscan, cog, pdb)
Assembly tools	Euler Phrap Phred Cross_match Cap3 Tigr
Alignment tools	ATGC PClustal W Blast Fasta Clustal W Mumer

Shared Computing tools (2)

Sequence Analysis tools	Genscan Glimmer Glimmer M TransTerm RepeatMasker RepeatFinder Cpgplot StackPack sirnaPro EMBOSS Toolkits BLAST, ClustalW
Protein Analysis tools	Interproscan
3D Structure Prediction	SAM/HMMer Modeller / Prospect AMBER/ GROMACS Procheck CE /CE-MC

Shared Computing tools (3)

Gene Analysis Tools	GeneKey EDSAc
Others	Phlip Paml DCGene
DataBases	Embl Swissprot Pdb Pfam siDB Cellulase RESID PubMed SPIES ……

清華大學
Tsinghua University

Bioinformatics Applications

- Protein target selection for rice genome
- Multi-sequence alignment for ganoderma family
- Gene joint for white mice
- Cardiovascular disease research
- ……

清華大學
Tsinghua University

Next Step

- Extend BioGrid to more and more research institute;
- Replant more computing tools and super computing power into BioGrid;
- Build more Bioinformatics related database mirrors and update in time;
- Make it more Compliant with existing bioinformatics research habits
  - Define workflow easily;
  - Interactive;

清華大學
Tsinghua University

Thanks!