Genomes to Grids

Bio Data Distribution for Grid Computing

Biologists have discovered many millions of genes and genome features, now part of the bio-data "library" distributed on computers around the world. Grid computing methods for finding and using interesting genome knowledge from this mountain of data are discussed - their promise and practical concerns for building usable bioinformatics grids.

See http://iubio.bio.indiana.edu/biogrid/

Don Gilbert, gilbertd@bio.indiana.edu October 2002

Bio-grids - what might they be ?

- ***transparent use of available workstations ; commodity grid resources (commercial, academic)***
- ***find biodata [directories], computing resources easily and automatically***
- ***personal/project resources and peer-peer sharing***
- ***less reliance, less cost for centralized services or building local IT centers***

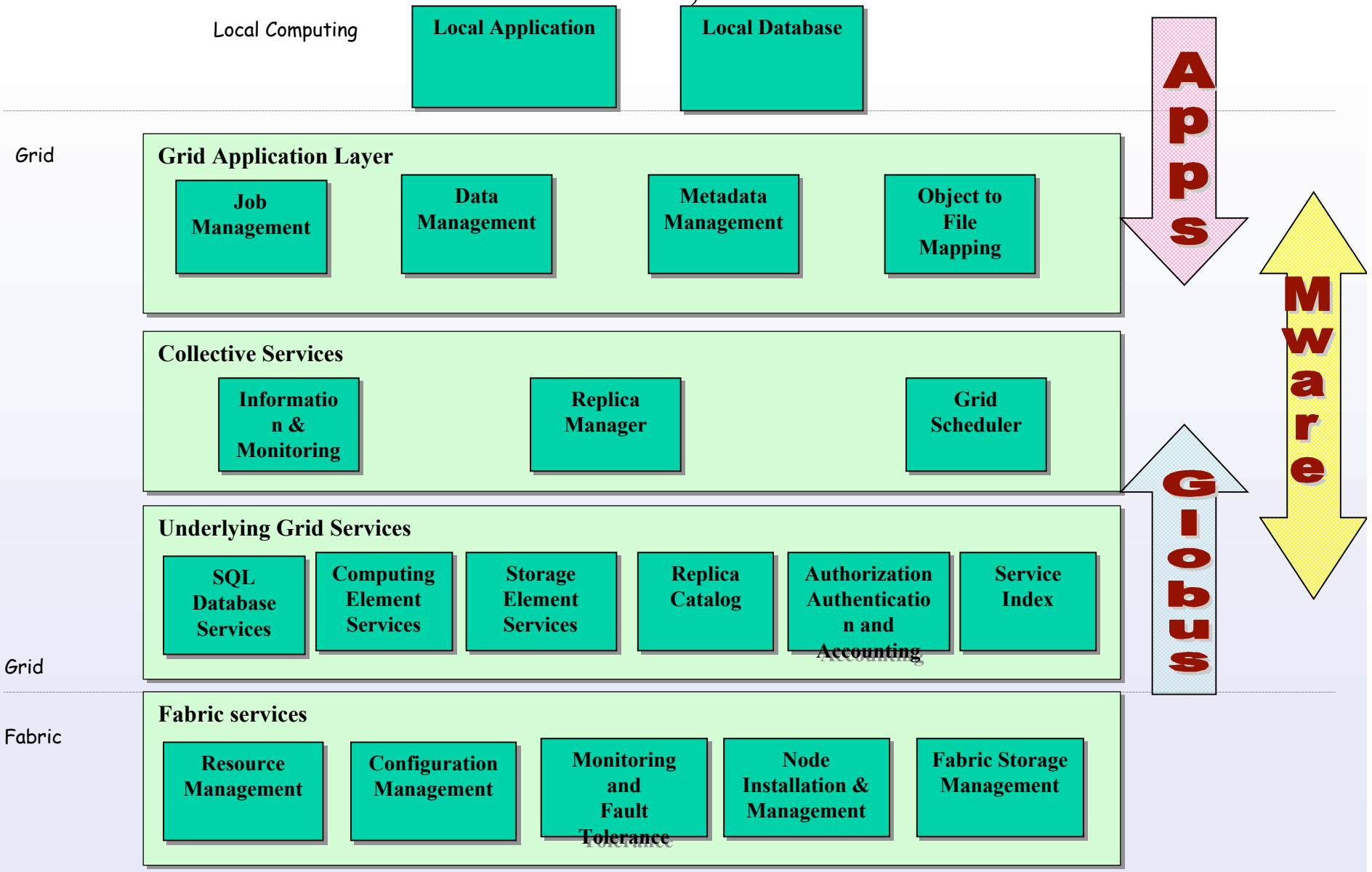
Science Grid Components

Infrastructure for distributed analyses

- analyses distributed among 1000s of commodity computers
- high-volume data distribution
- data and resource directories (catalogs)
- security, authenticated use
- peer-to-peer sharing and collaborations
- Need efficient discovery and distribution of high volume data
- *Data grid infrastructure still needs work*

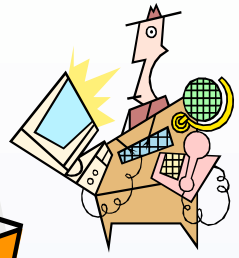
European DataGrid Architecture

from Bob Jones, CERN



EU DataGrid Interfaces

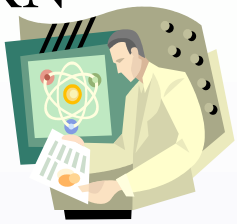
from Bob Jones, CERN



Application Developers

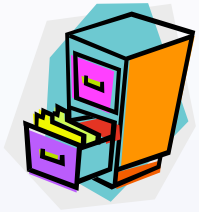
Local Application

Local Database



Scientists

Certificate Authorities



System Managers

File Systems

User Accounts

Grid Application Layer

Job Management | Data Management | Metadata Management | Object to File Mapping

Collective Services

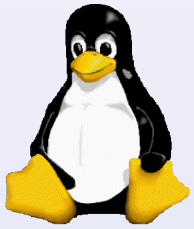
Information & Monitoring | Replica Manager | Grid Scheduler

Underlying Grid Services

SQL Database Services | Computing Element Services | Storage Element Services | Replica Catalog | Authorization Authentication and Accounting | Service Index

Fabric services

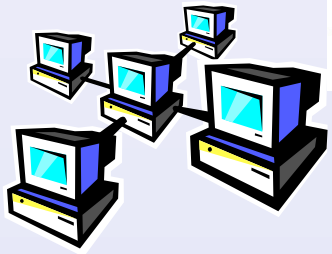
Resource Management | Configuration Management | Monitoring and Fault Tolerance | Node Installation & Management | Fabric Storage Management



Operating Systems

Mass Storage Systems
HPSS, Castor

Storage Elements



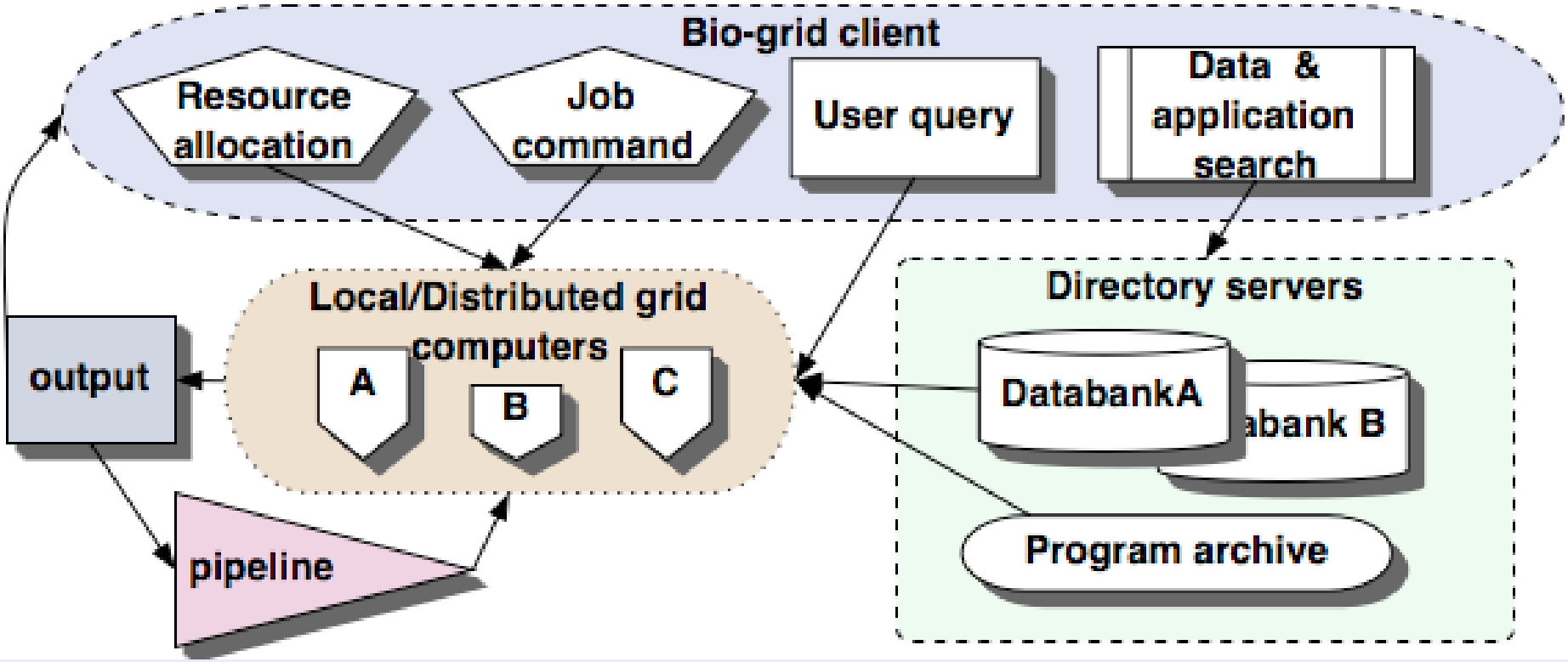
Computing Elements

Batch Systems
PBS, LSF

Current Grid Choices, Directions

- ***Commercial grid offerings***
  - Avaki, Blackstone Computing, Entropia, Platform Computing, and United Devices have Grid computing packages for Life Sciences.
  - Sun Microsystems (GridEngine), IBM, others are involved
- ***Science grids***
  - Globus package (globus.org); NSF Middleware Initiative (NMI) ; Condor
  - Physics grid ; iVDGL.org ; others
- ***Biology grids***
  - **EU DataGrid** - all sciences, esp. physics, biology included ; eu-datagrid.org
  - **myGrid (UK)-** focus on bioinformatics, combines Grid, Web Services, and Semantic Web efforts, http://www.mygrid.org.uk/
  - **US Biogrids -** where are they**?**

BioGrid Schematic



- Grid-aware client software
- Data and software resource directories
- Grid of processing computers

Moving Data on the Grid

1. @virtualdata= biodirectory "find protein coding sequences for species X,Y,Z"
2. @realdata= biodirectory "get locators for @virtualdata split 100 ways"
3. for i (1.. 100) {
   copydata(realdata[i],gridcpu[i]);
   runapp(gridcpu[i]) }

BioData

- ***BioData: size, contents, dispersion, uses***
- ***Genome data***
  - **very important, highly complex, harder to find, long lived**
  - **Literature (abstracted and curated), Sequence and feature analyses, maps, controlled vocabulary/ontologies, people, biologics, contacts, etc.**
- ***BioData access***
  - **Need to find and use "best data"**
  - **New data kinds and sources - bio-information is very fluid**
  - **Need current data ; update monthly, weekly, daily**
  - **Distributed widely in world among 1000s of national, regional centers & labs**

Bio Databanks, EBI, Sept. 2002

Databank	Contents	Entries
EMBL	DNA Sequences	18,800,000
SWALL	Protein sequences	900,000
InterPro+	Protein motifs	1,000,000
HGBASE	SNP database	1,500,000
	Metabolic Pathways	250,000
MEDLINE	Literature	11,350,000
Total		33,800,000

Genome Data Objects

Contents	Entries
Literature References	140,000
Gene variants	112,000
Genome features	50,000
Genes	40,000
Transgene constructs	37,000
Chromosome aberrations	16,000
Fly Stocks	15,000
Drosophila Researchers	6,600
Total	416,600

Drosophila genome, FlyBase, Sept. 2002

Contents	Entries
Named genes	188,000
Genome features	864,000

8 eukaryote genomes, euGenes, July 2002

Directories of Genome Data

- **For genome data, *"broad and shallow"* directories federate the *"narrow and deep"* data-bases**
- ***BioData access tools***
  - ***SRS - Sequence Retrieval System; Entrez ; AceDB***
  - ***RDBMS ; Ensembl ; IBM DiscoveryLink; BioSQL; BioDAS ;***
- ***Directory services - Data tools + LDAP , Web Services***
  - **LDAP:** mature, efficient for high volumes, allows federated queries over distributed directories, and works well for SRS databanks and genome annotations
  - **Web Services:** new, simple & complex for XML messages over Web ; has wide industry support , but its many standards are in flux

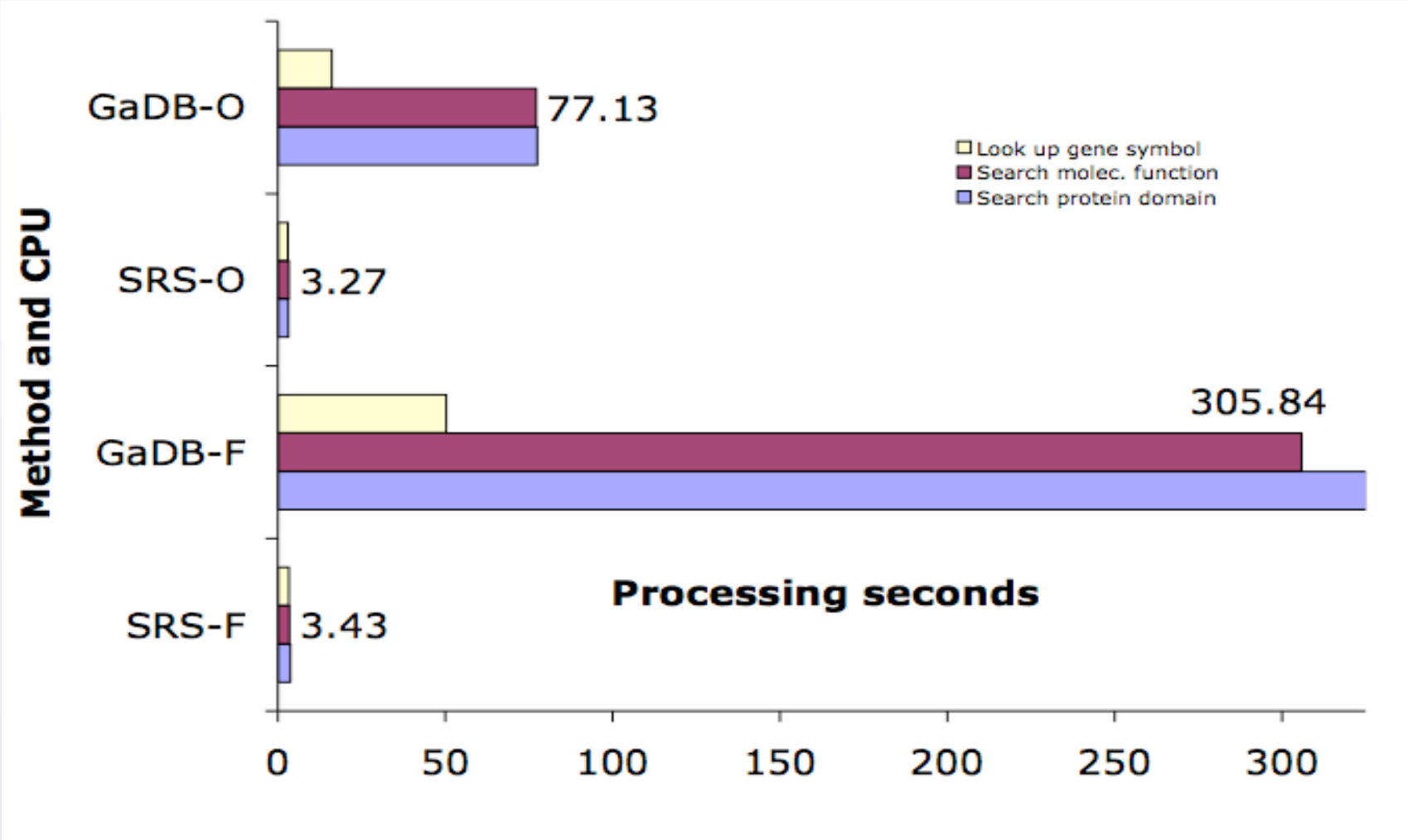
Data Access: myGrid & SRS

From myGrid presentation, Spring 2002
http://www.mygrid.org.uk/

- Much of the functionality that myGrid seeks to implement exists in LION's SRS
  - SRS is highly evolved & very powerful
- myGrid is about developing a scalable, distributed framework built on open standards & services
  - It's about linking pieces, not the pieces themselves
- SRS should be available as a Web service.

Data Access: SRS & RDB

Drosophila Genome Annotations
SRS or Gadfly DB relational database
Web search time (*shorter is better*; two computers - O,F)



Bio-directories: Design

1. Develop schema describing directory objects and attributes. Essential fields include ID/accession, data class / category, update time. Start with minimal directory descriptions.
2. Create directories of data records, with existing backend software such as SRS, RDMBS, Entrez, others.
3. Replicate directories among data centers; use for determining primary data to be fetched or mirrored.
4. Common exchange formats, schema, directory query syntax are necessary, implementation details are at the choice of a data center.

Bio-directories: Technology

- ***Technology for finding bio-data***
  - ***Current: Web pages ; Web Indexers (Google); FTP servers ;***
  - ***Sometimes: CORBA ; Java RMI***
  - ***Usable: Lightweight Directories - LDAP***
  - ***Developing: Web Services (XML on Web: SOAP, WSDL, UDDI, ...)***
- ***Related: BioDAS ; Biomoby ; Life Sciences ID (LSID) ;***

LDAP or Web Services ?

- LDAP for bio-data directories
  - Available now with many features needed
- Web/XML
  - Web/SOAP/WSDL/UDDI: SOAP for communication of directory requests, WSDL for an interface to the directory repository, UDDI to locate the service *(some assembly required…)*
  - DSML: a direct conversion of LDAP to XML, for Web/XML interoperability to LDAP (e.g., http://www.dsmltools.org/); supported by industry (Msoft, Sun, others)

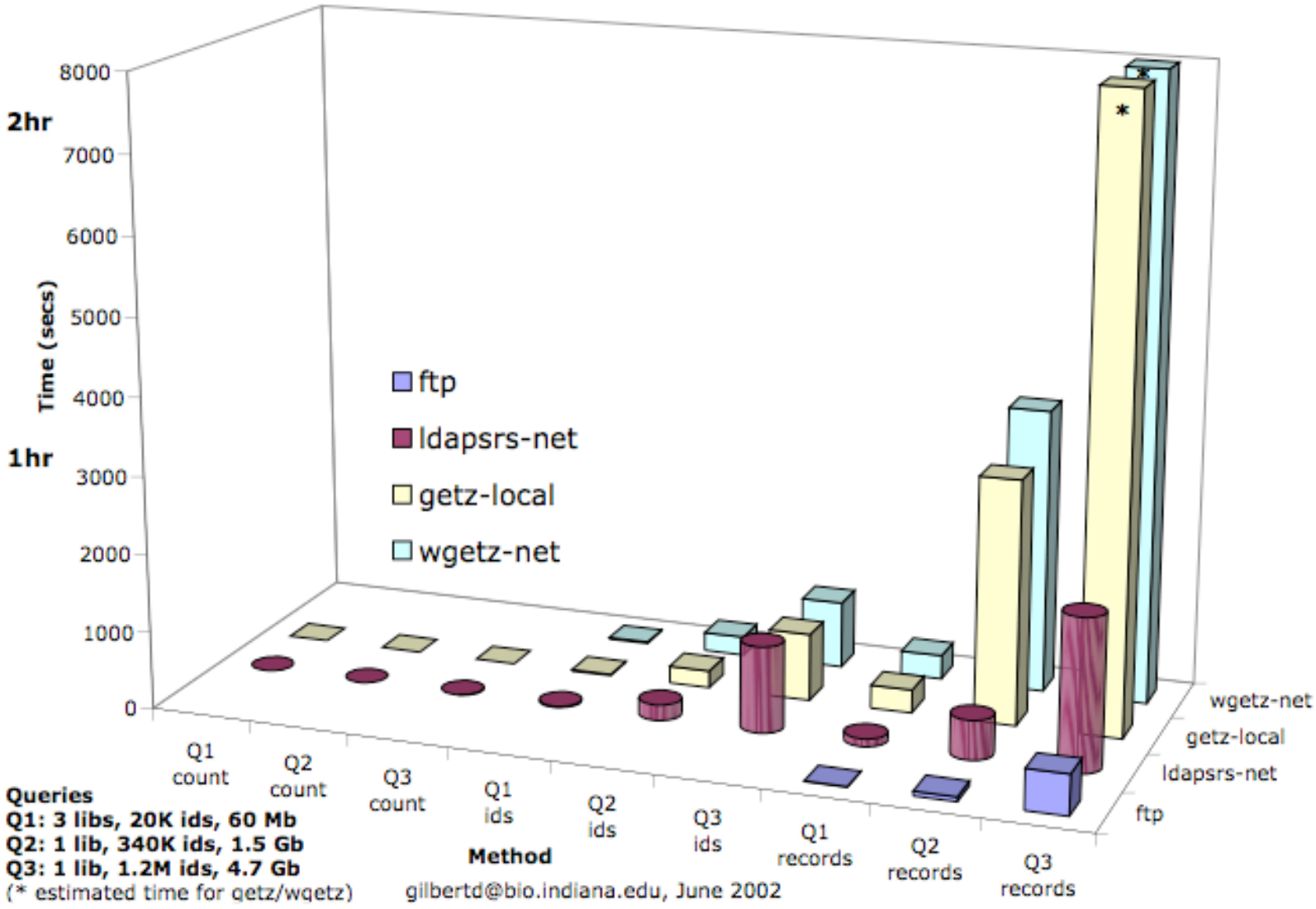
Light-weight Directory Features (LDAP)

- Flexible, hierarchical directory of objects with identifiers to community definitions. Objects are simple or complex. Each has attributes (fields) composed of strings, numbers, binaries and complex structures
- Use many backend systems (RDBMS, SRS); can be added to search/retrieval systems relatively easily
- Globally distributed searches of many directories
- Schema are documented: objects and attributes have unique identifiers and definitions (e.g. IETF RFC documents)
- Schema search/retrieval for directory 'discovery'
- Computable search, browse, retrieval; referrals to other servers and remote objects; extension mechanisms for new object types
- Replication of directories; mechanisms for peer group updates
- Security mechanisms for data transport, access and updates

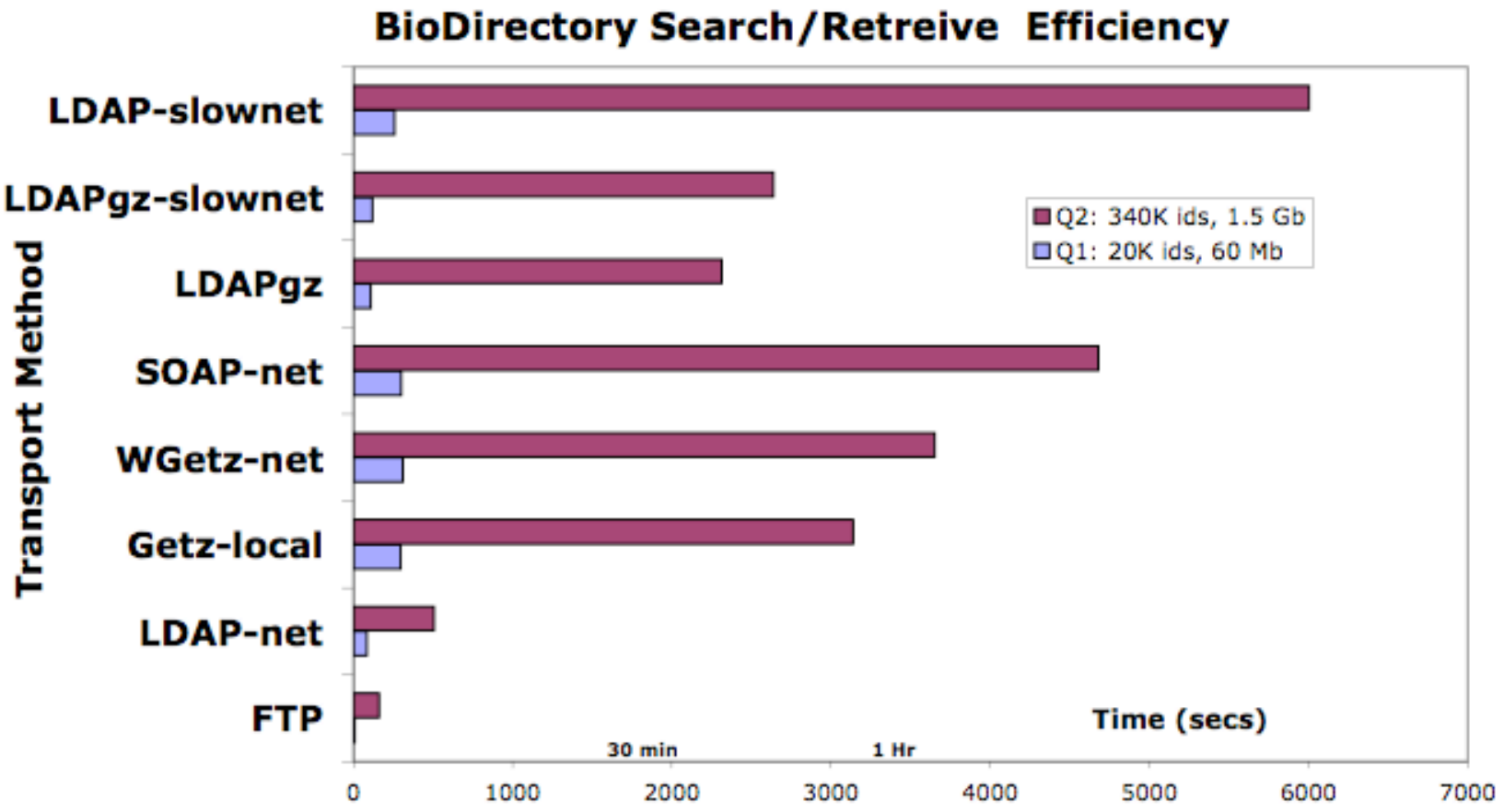
SRS6 - LDAP & WebXML gateways

- Sequence Retrieval System (SRS) knows millions of bio-objects; good start for bio-directories
- OpenLDAP server combined with SRS6
- WebService SOAP server/client with SRS6
- SRS-LDAP-SOAP software is available at http://iubio.bio.indiana.edu/biogrid/directories/
- Compare LDAP, SOAP, Wgetz, FTP for Grid uses
- LDAP is 5x faster than SOAP, Wgetz

BioDirectory tests for Grid



BioDirectory Methods Efficiency



FTP > 5x LDAP > 5x Web	FTP	LDAP-net	Getz-local	WGetz-net	SOAP-net	LDAPgz	LDAPgz-slownet	LDAP-slownet
Q2: 340K ids, 1.5 Gb	162	506	3138	3653	4680	2314	2639	6002
Q1: 20K ids, 60 Mb	8	84	297	309	300	107	119	258

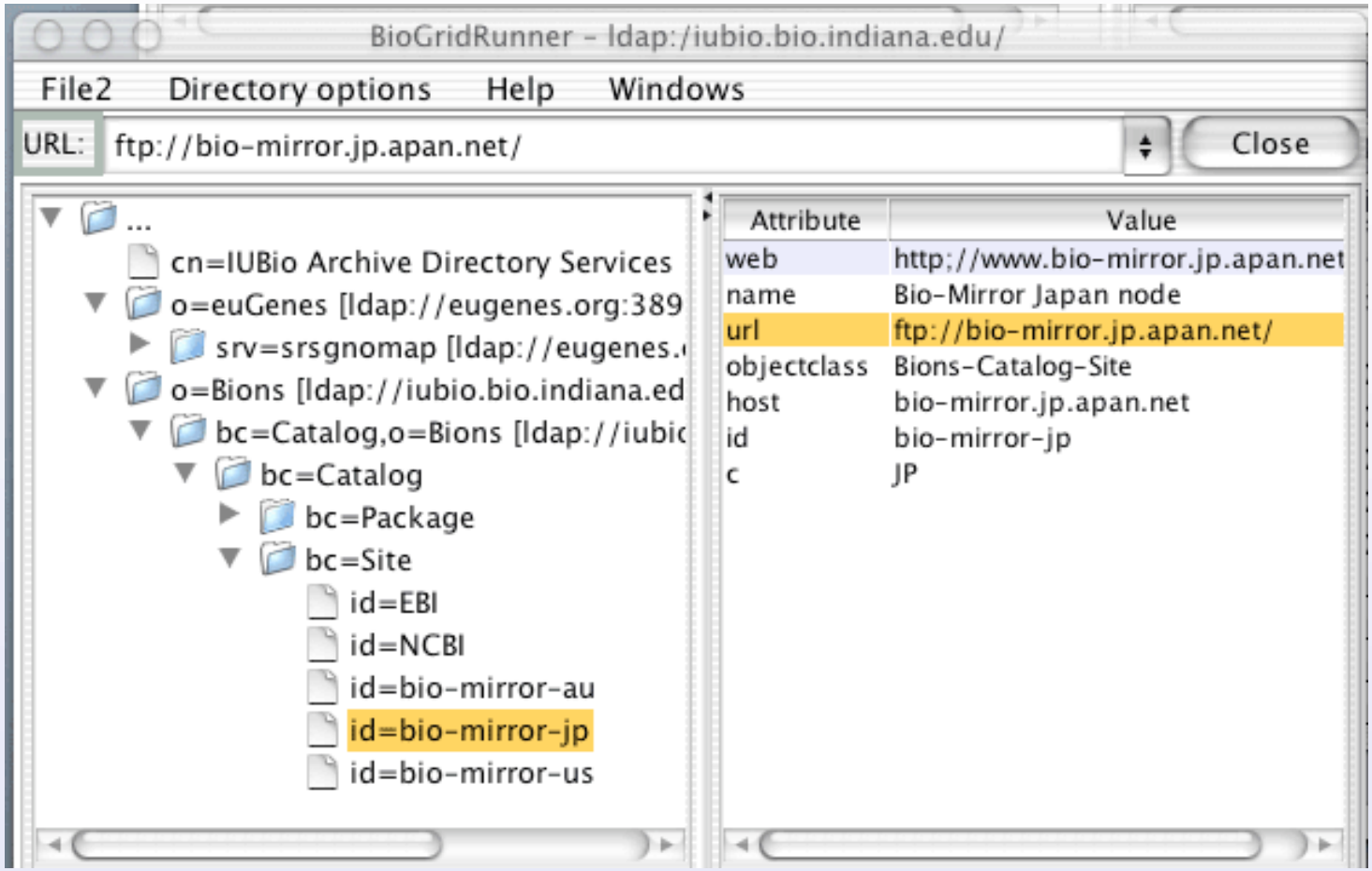
Q1/Q2 - Query biosequence directories ; **LDAPgz** - 10 fold record compression ; **FTP** - no query selection

Q1 = {swissprot trembl refseq}-des:kinase , 20K records; **Q2** = genbank-org:drosophila , 340K records

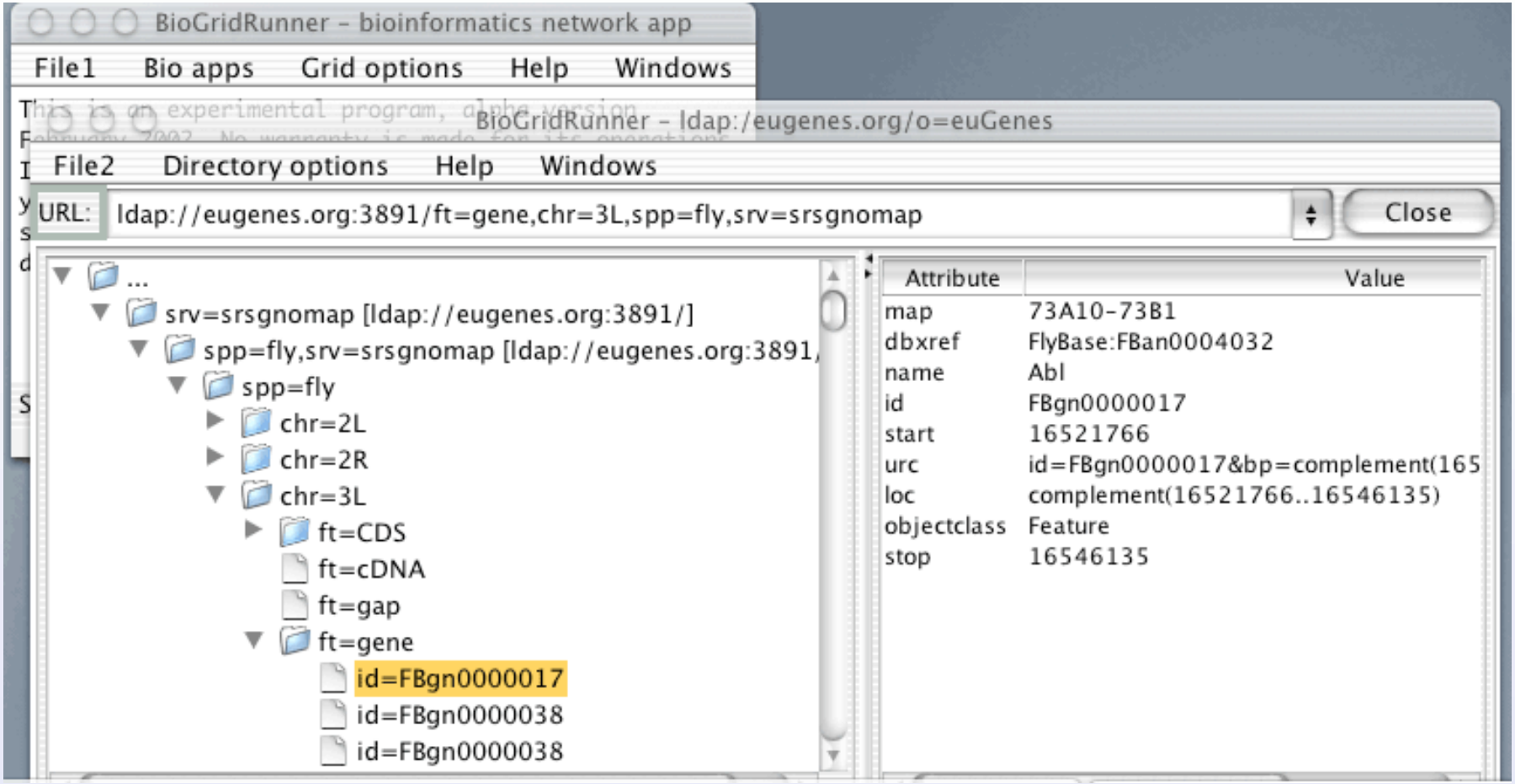
gilbertd@bio.indiana.edu, Oct 2002

Using Bio Directories

- Simple client software
- Automated use
- People use
- Discovery
- Search by many criteria
- Retrieve bulk subsets



Genome Feature Directory



- Genome objects fit in directories, search & retrieve as annotated sequence (like BioDAS)
- Same directory methods work for objects, databases, software and other resources

Basic Directory Client Programs

```
// Java LDAP client to Bio-directory
void SrsLdapClient() {
  String url   = "ldap://iubio.bio.indiana.edu:3895/srv=srs";
  String query = "(&(objectClass=BioseqRecord)(lib=genbank)(org=drosophila))";
  Properties env= new Properties();
  env.put( Context.PROVIDER_URL, url);
  env.put( Context.INITIAL_CONTEXT_FACTORY, "com.sun.jndi.ldap.LdapCtxFactory");
  DirContext dir = (DirContext) new InitialDirContext(env);
  NamingEnumeration results = dir.search( "", query, new SearchControls());
  while (results.hasMore()) System.out.println(results.next());
}
// Java SOAP client to Bio-directory
public interface ISrsSoap {int search(String query); String getResult(int index);}
void SrsSoapClient()  {
  String wsdl = "file://srsglue.wsdl"; //Web Services description
  String query= "(&(objectClass=BioseqRecord)(lib=genbank)(org=drosophila))";
  ISrsSoap dir=(ISrsSoap)Registry.bind( wsdl, ISrsSoap.class );
  int nResults= dir.search( query );
  for (int i= 0; i<nResults; i++) System.out.println(dir.getResult(i));
}
// Perl LDAP client to Bio-Directory
sub SrsLdapClient {
  my $url  = new URI::URL("ldap://iubio.bio.indiana.edu:3895/srv=srs");
  my $query= "(&(objectClass=BioseqRecord)(lib=genbank)(org=drosophila))";
  my $ldap = new Net::LDAP($url->host, port => $url->port) or die "$@";
  $ldap->bind; my $results = $ldap->search("filter" => $query);
  foreach my $entry ($results->all_entries) { $entry->dump; }
}
```

Wrap up

- **Future of Bio-data on Grids**
  - **Computationally find and use dispersed, complex data**
- **Best methods for Bio-data to Grids**
  - **High volume and complex data**
  - **Efficient selection and transport to grid computers**
  - **LDAP works well ; Web-XML is usable**
- **Community needs and uses**
  - **Shared data descriptions, schema, ontologies (Semantic web)**
  - **Simple, practical, flexible grid methods ; use existing dbs**
  - **Use common & developing standards**

See http://iubio.bio.indiana.edu/biogrid/

Don Gilbert, gilbertd@bio.indiana.edu October 2002