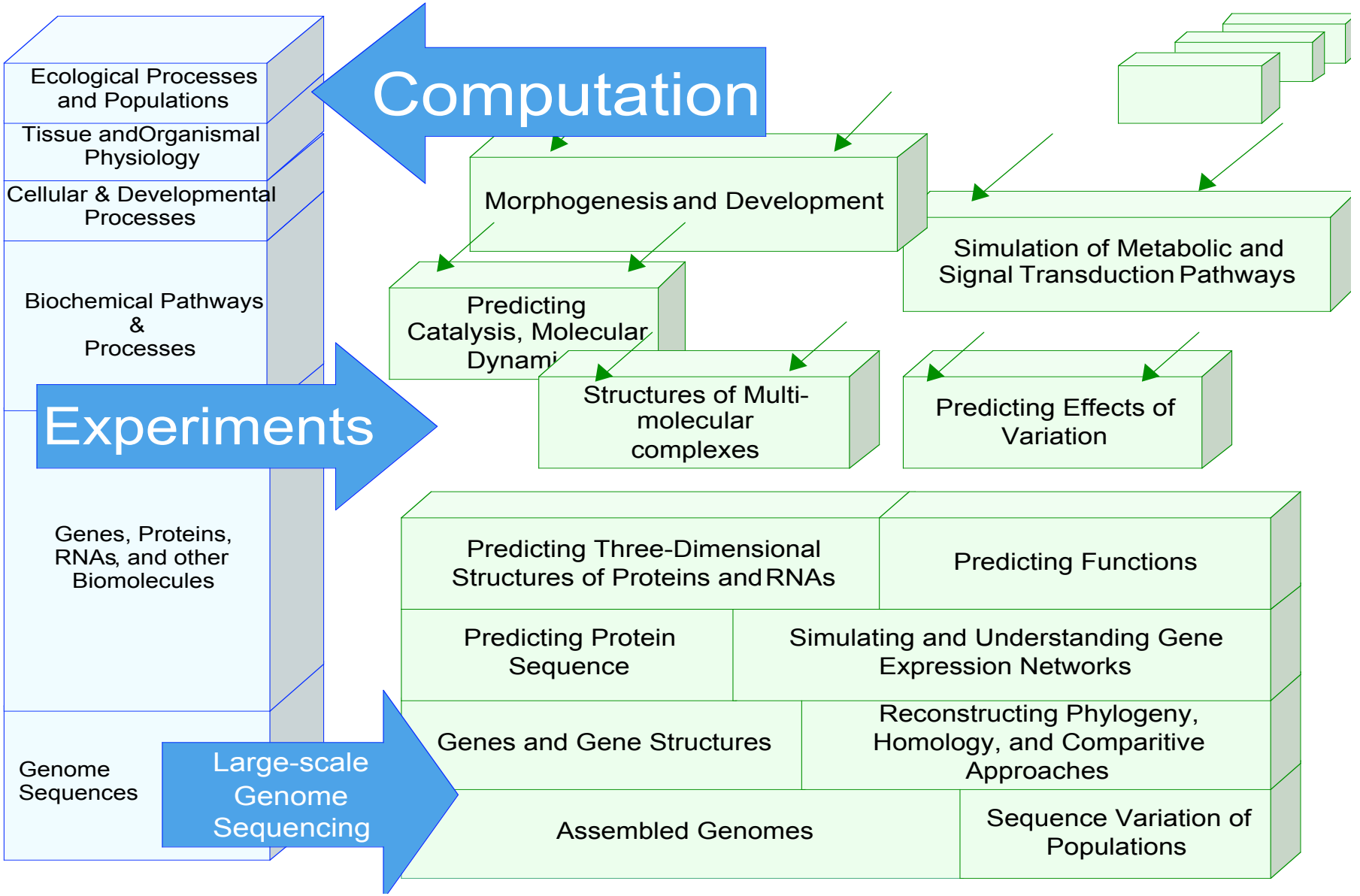
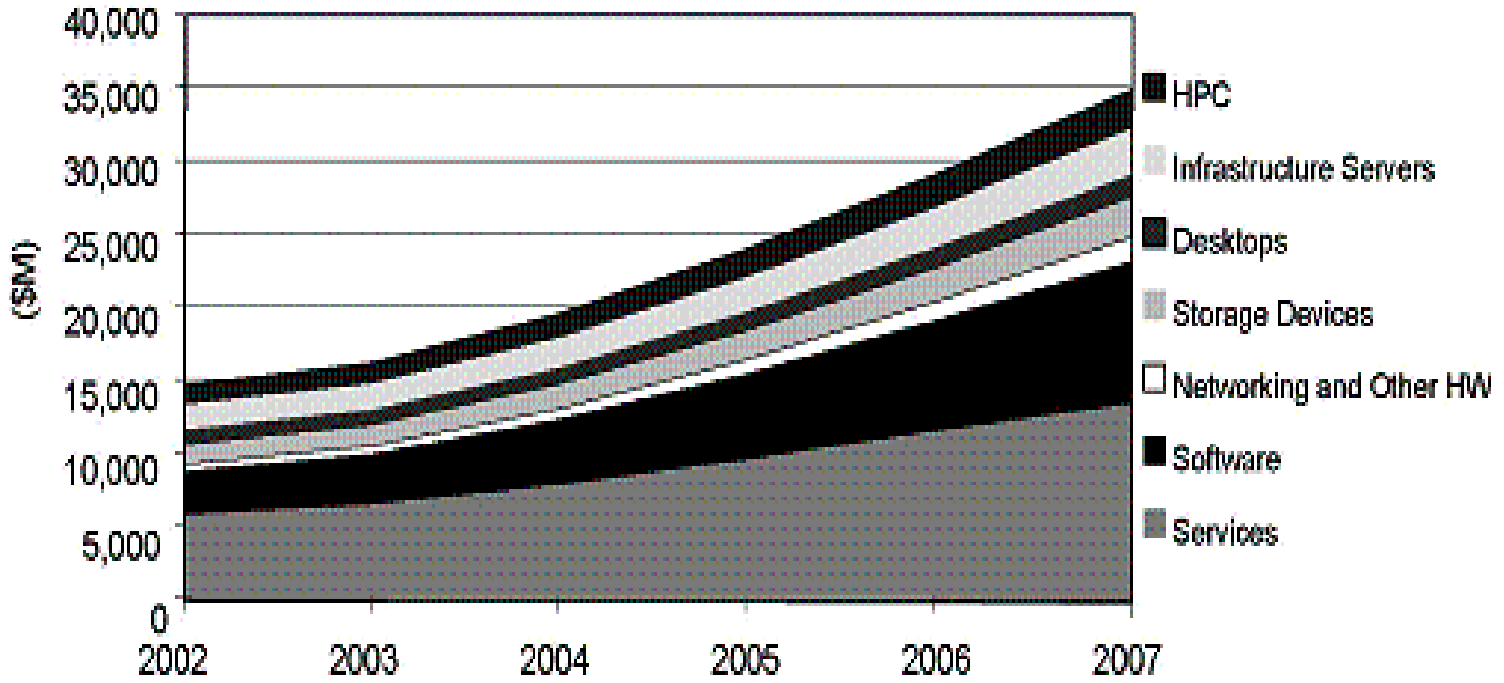
Update on Progress Towards an Open Life Science Reference Architecture

Rick Stevens
Argonne National Laboratory
University of Chicago
stevens@mcs.anl.gov

Genomics is Powering the New Biology, but Computing is in the Drivers Seat



IDC Worldwide Bio-IT Market



The New Biology

- Genomics
- Functional Genomics
- Proteomics
- Structural Biology
- Gene Expression
- Metabolomics
- Advanced Imaging

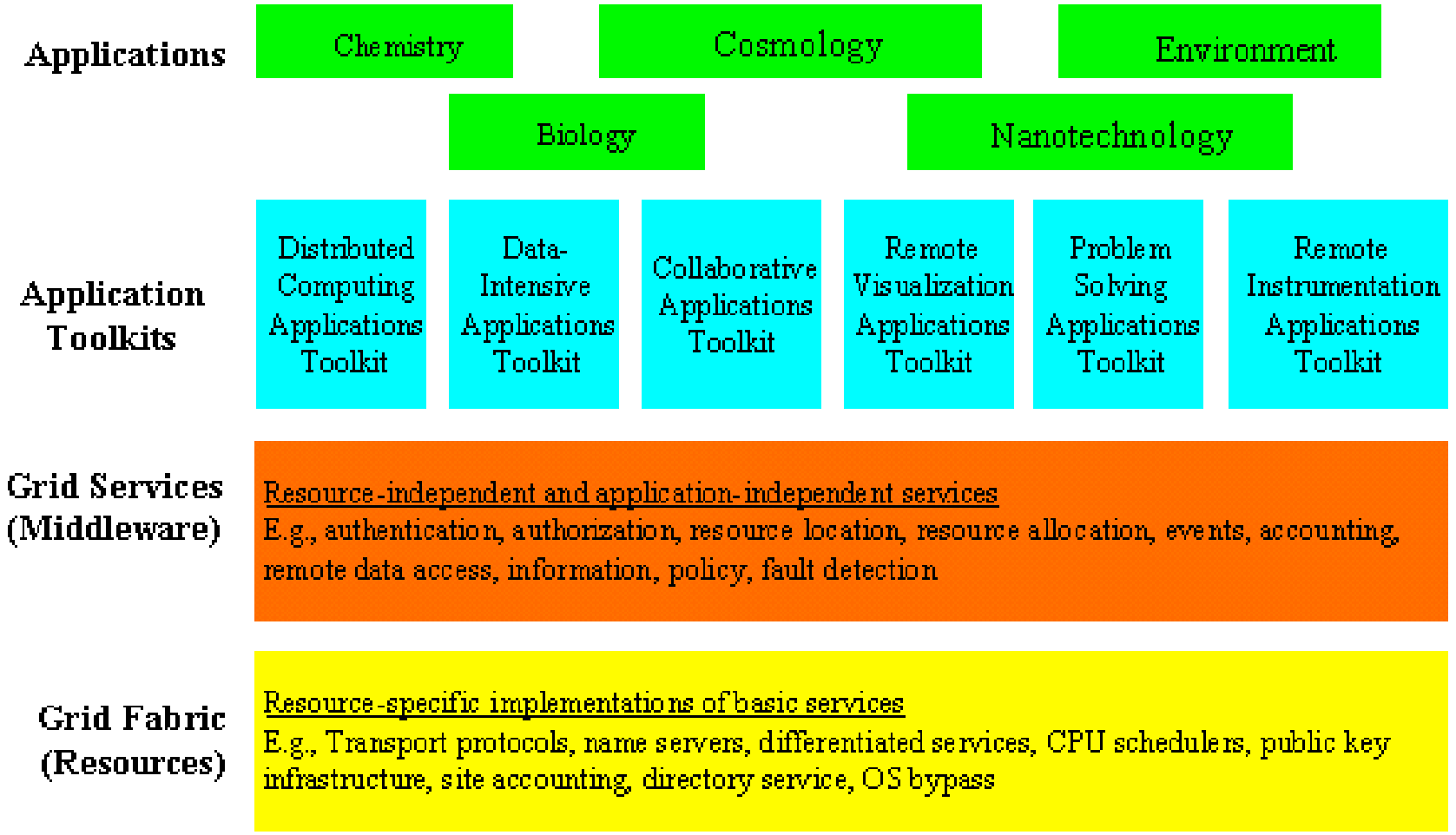
- High-throughput methods
  - Low cost
  - Robotics
- Bioinformatics driven
- Quantitative
- Enables a systems view
- Basis for integrative understanding
  - Global state
  - Time dependent

Open Life Science Grid Infrastructure

Open Grid Services Infrastructure

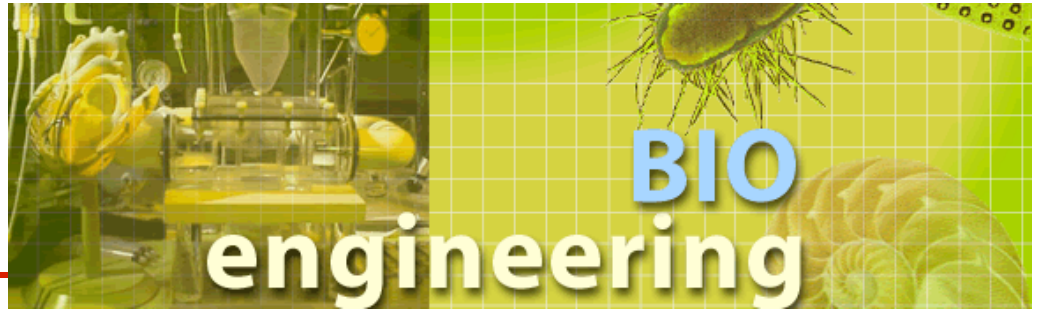
mcs

The Grid Software Stack



mcs

Future Headlines



- First synthetic model prokaryotic organism
- Characterization of human microbial ecology
- Global index to life on earth
- Characterization of microbial life
- Theory of cell evolution and organization
- Theory of evolution of intelligence
- First synthetic eukaryotic organism
- Confirmation of extra-solar earthlike planets
- Synthetic self-reproducing biomemetic nanosystem

mcs

Determining Requirements for the version 1 of the Open BioGrid (OBG-1)

- Model for Community Involvement
  - MPEG-7 process
  - MPI Forum
  - More focused effort than the current GGF LSG processes
- Developing a call for proposals
  - Technologies (mostly existing tools and capabilities)
  - Architectures (following from the Bluetooth concept of usage scenarios)
  - Interfaces and APIs (leverage GGF, W3C, I3C)
- Requirements Collection (for the next say the next six months)
  - Input for an eventual RFP
  - Scope the components of a (set of) “Standard” (s)
  - Related to existing Standards (GGF, W3C, I3C, etc.)

mcs

Some High-Level Requirements Driven by Life Science Demographics

- Platform for highly distributed data sharing and curation
  - Directly supports distributed networks of LS researchers
- Peer-to-peer updates and resource management
  - Decentralized administration
  - Provides for the concept of super-Peers
  - Enables self-organizing groups
- Services rich environment
  - Workflows (standard protocols.. links to “current protocols”)
  - Data update and access services (wrappers around major data providers)
  - Tools (informatics) suites and code push services
  - Ontologies and conceptual discovery frameworks
  - Computing services
    - Generic computing services (cycles on platforms.. Reseller grid type services)
    - Specific computing services (named, typed services, e.g. BLASTP, etc.)
- Bio applications neutral
  - Genomics, Proteomics, SysBio, Imaging, Discovery, etc.

mcs

An International Systems Biology Grid

- A Data, Experiment and Simulation Grid Linking:
  - People [biologists, computer scientists, mathematicians, etc.]
  - Experimental systems [arrays, detectors, MS, MRI, EM, etc.]
  - Databases [data centers, curators, analysis servers]
  - Simulation Resources [supercomputers, visualization, desktops]
  - Discovery Resources [search servers perhaps optimized]
  - Education and Teaching Resources [classrooms, labs, etc.]
- More fine grain than many current Grid projects
  - More laboratory integration [need small laboratory software interfaces]
  - Most of the participants will be experimentalists [workflow, visualization]
  - More diversity of data sources and databases [integration, federation]
  - More portals to simulation environments [ASP models]

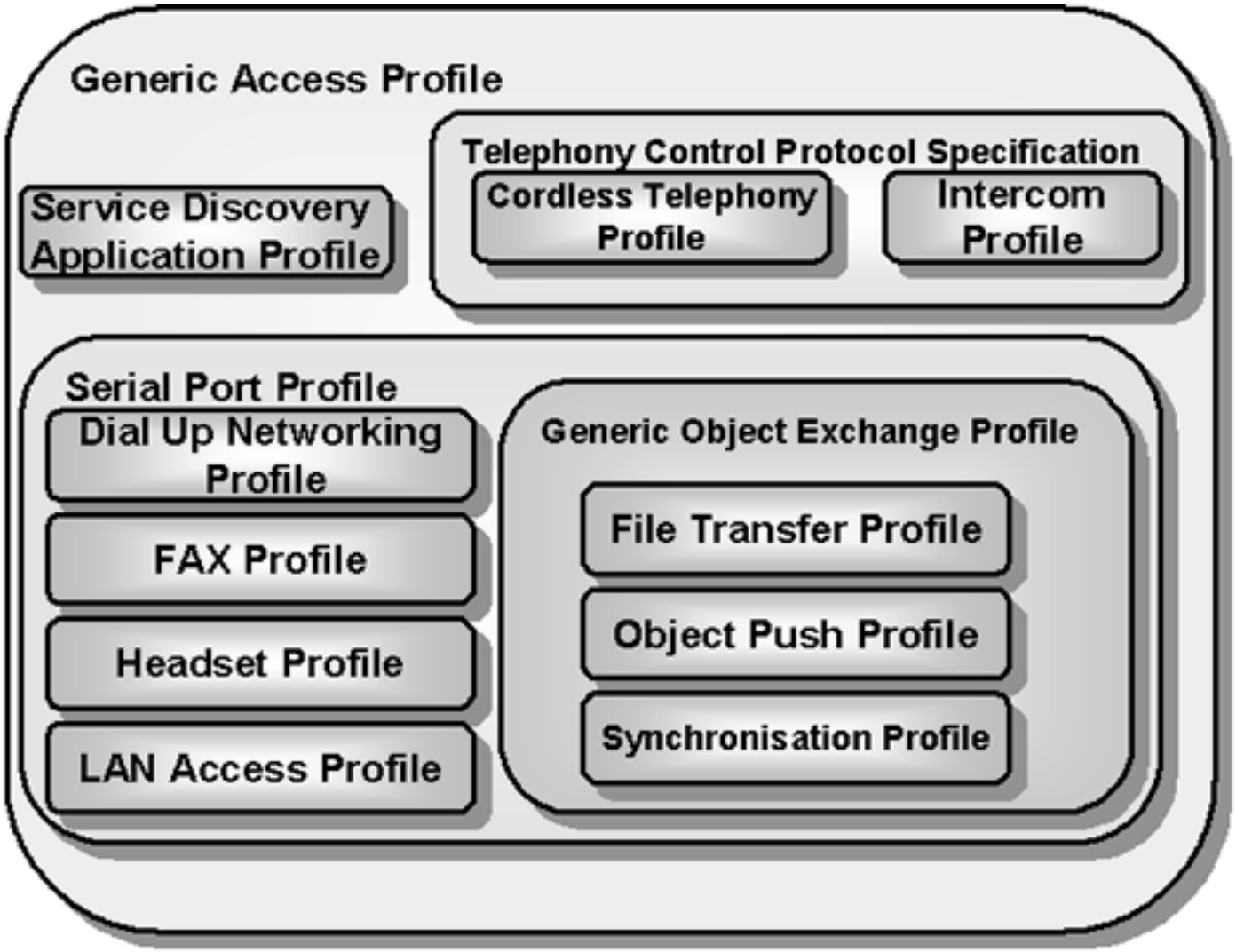
mcs

Bluetooth "Profiles"

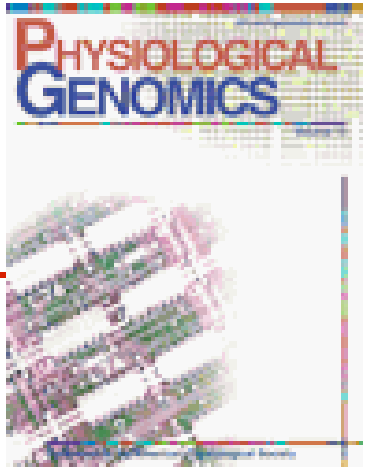
- A profile is just a description of how to use a specification to implement a given end-user function. The International Standards Organization (ISO) first came up with the idea of profiles. Profiles help interoperability in four key ways:
  - Implementation options are reduced, so applications share the same features.
  - Parameters are defined, so applications operate in similar ways.
  - Standard mechanisms for combining different standards are defined.
  - User interface guidelines are defined, giving uniformity across devices.
- The profiles describe minimum implementations

mcs

Bluetooth “profiles”

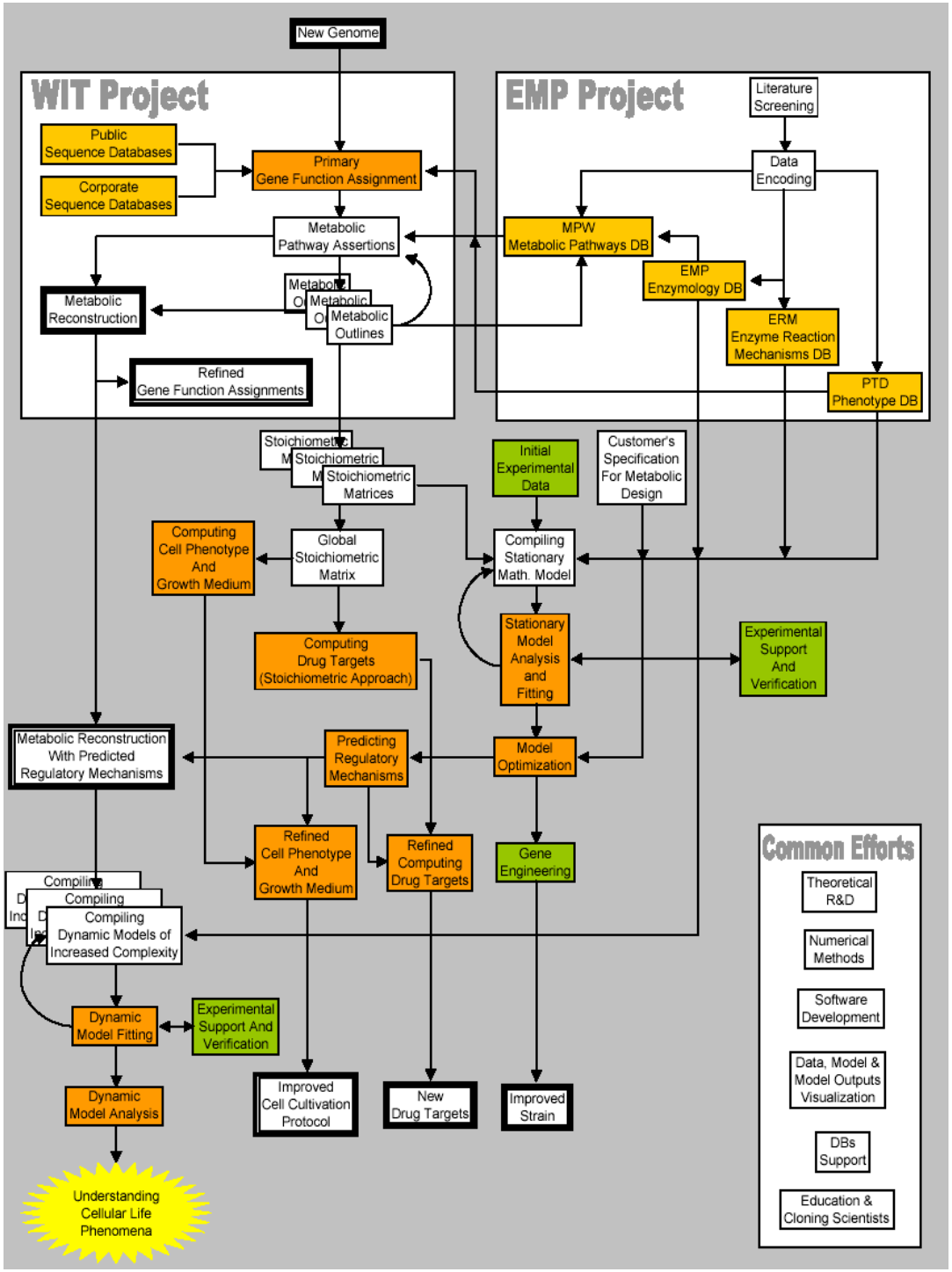


Open BioGrid Example Scenario



- A P2P Distributed Curation Environment
  - Core database(s)
    - Extensible core schemas
    - Object modeland external data representation support
    - Language independence (PERL, Python, Java)
  - High-performance services interfaces
    - Bulk data xfer
    - HPC conduit for Local and Grid computing
  - Web based interfaces
    - Human (portals)
    - WSDL
  - Peer-to-Peer synchronization/updates
    - Open Sub/Pub model
    - Data and Code

mcs

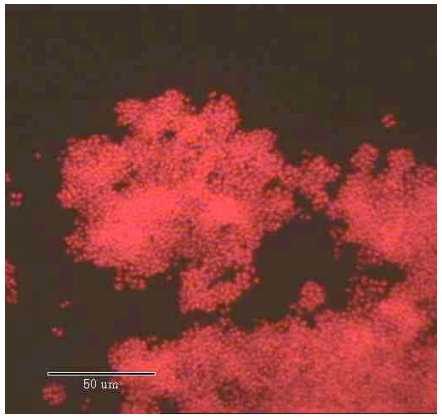


A Systems Biology Approach to Understanding Cellular Life Integrating:

- Theory
- Numerical methods
- Software development
- Data models and Visualization
- DB support
- Education

Mathematical Toolkits Focused on Biological Systems

- "A Mathematica for molecular, cellular and systems biology"
  - Core data models and structures [see db]
  - Optimized functions [see core libraries]
  - Scripting environment [e.g. Python, PERL, ruby, etc.]
  - Database accessors and built-in schemas
  - Simulation interfaces
  - Parallel and accelerated kernels
  - Visualization interfaces [info-vis and sci-vis]
  - Collaborative workflow and group use interfaces



mcs

Proposed Scope of an Initial Effort

- Open SW platform for biological data integration
  - Supporting distributed (ad hoc) team curation with versioning
  - Supporting rapid update cycles and annotations
  - Extensible data classes
- "Conduits" for synchronization (implemented via 'profiles')
  - Major community databases
  - Peer-to-peer services (instances of the standard infrastructure)
- An Open Architecture
  - Open interface for components
  - DB independent kernel (Oracle, SQL, DB2, Postgres, etc.)
  - Language independent design (PERL, Python, Java, C/C++)
  - Extensible APIs for local applications support
  - Grid/Web services (OGSA/OGSI, WSDL, etc.)
  - Flexible data sharing
    - Publish/subscription model of data sharing with IP

mcs

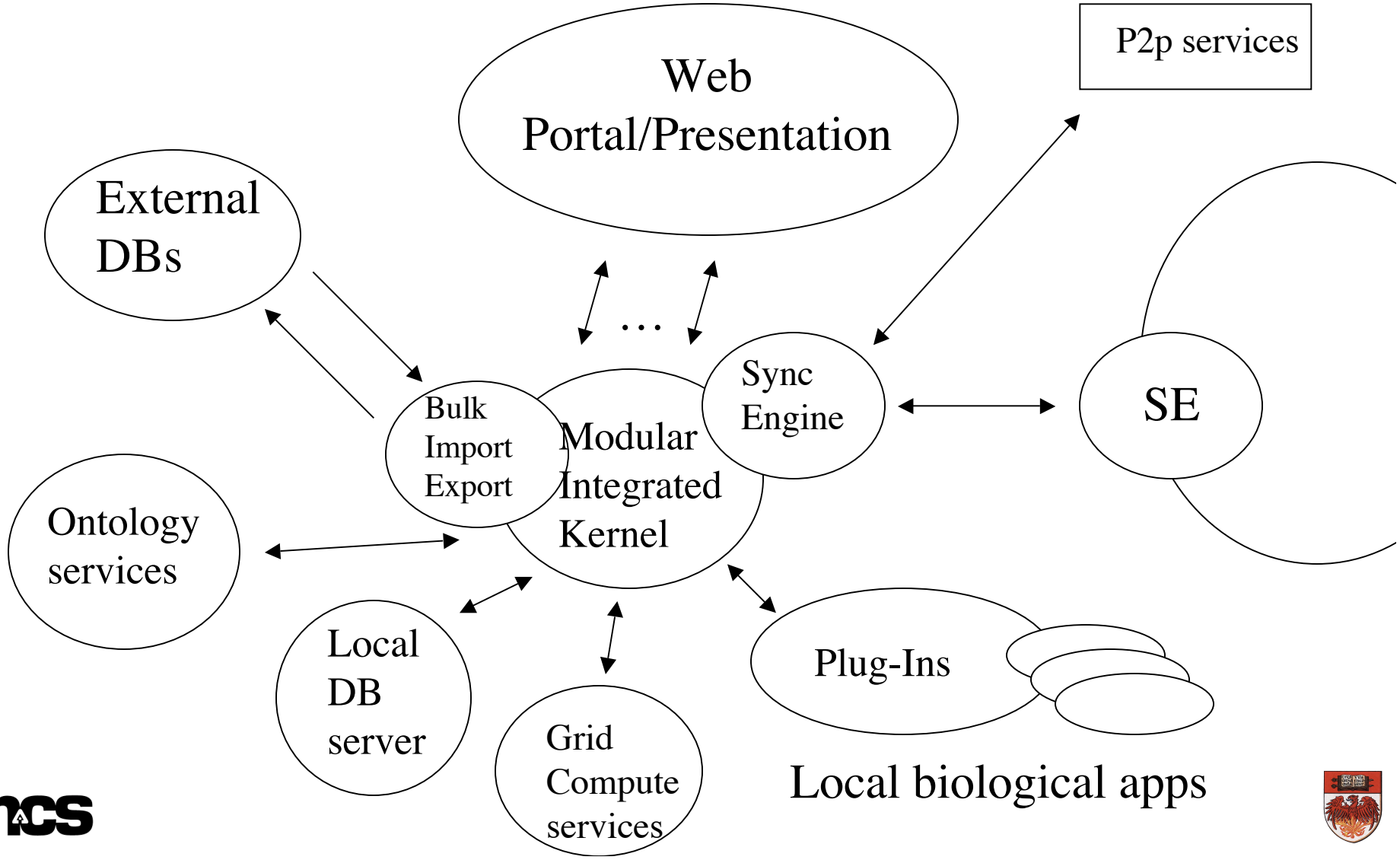
Scope II

- Supports multiple views and protection of proprietary data
  - Private data can be integrated with public data (policy engine)
  - Public data from many sources
- Interfaces
  - Web (WSDL) based Transactions
  - High-throughput data paths, bulk transfers
  - External Computing/Simulation and DB connections
  - Import/export APIs
- Scalability
  - Peer Target is 1000 instances in 36 months
  - Super Peer Target is 50 instances in 36 months
- Security
  - Based on PKI and GT3 technologies
- Portability
  - Reference implementations for Mac, Windows, Linux

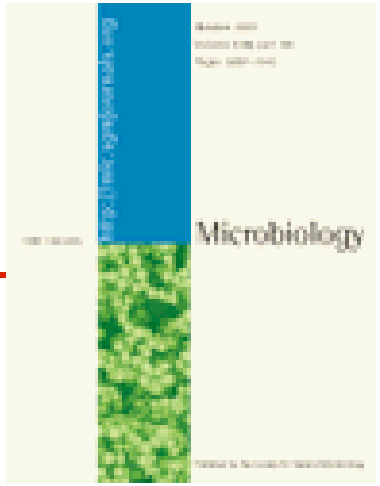
mcs

Peer-to-Peer Open Life Sciences Grid

the prototype SEED



More Details on the P2P ‘profile’



- Kernel server
  - Services registry
  - Computation on the DB
  - External representation of objects
  - Security
  - Versioning
  - Transaction support
  - Update (local) support
  - Schema extensions
- Import/Export engine
  - Portable formats
  - Interfaces to external sources/sinks
- Synchronization engine
  - Publish and subscription services
  - Update channels

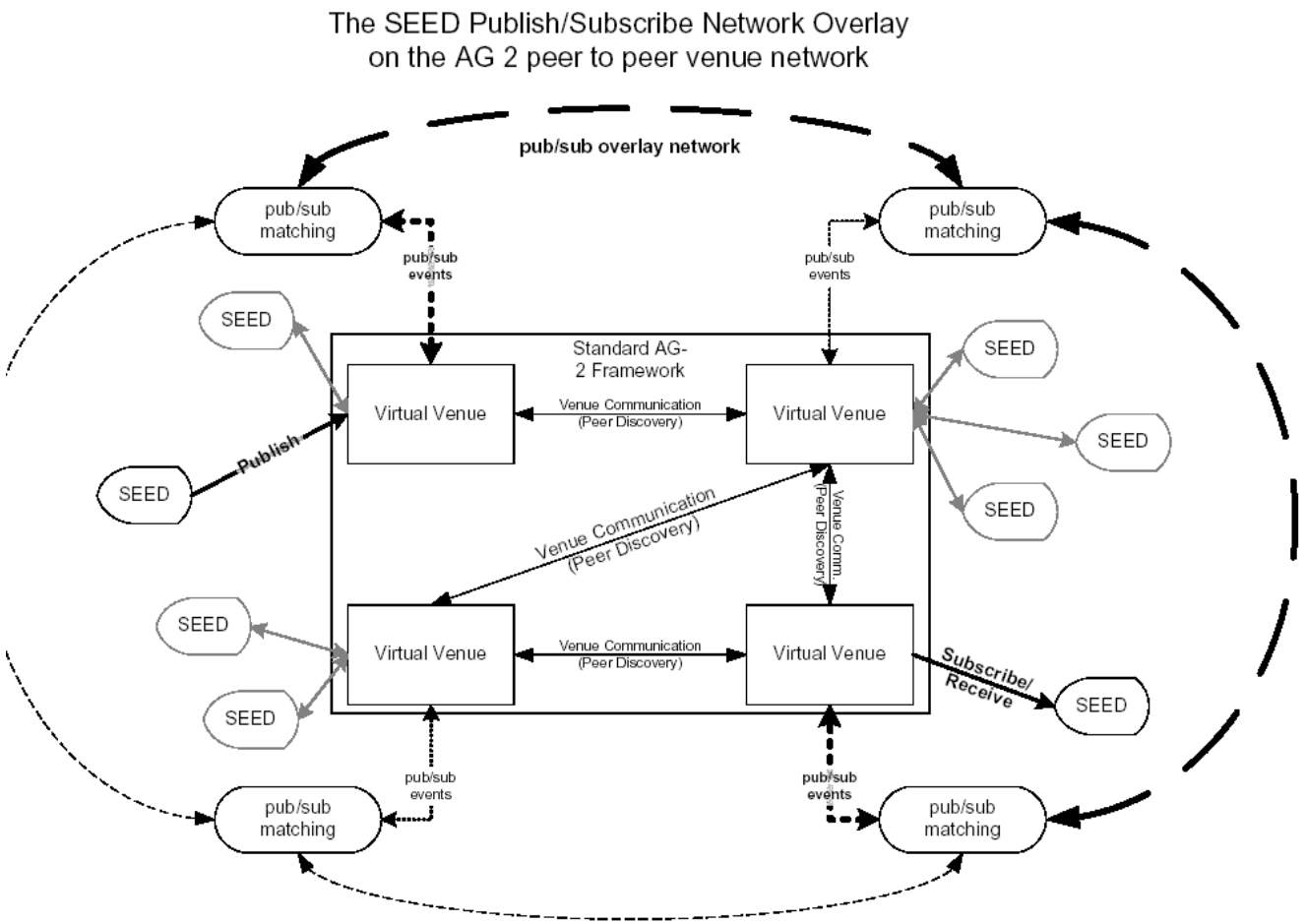
mcs

Scalability Goals of the P2P Design

- 100,000s of genomes (organisms)
  - Including support of many close variants
- Millions of genes and gene products (proteins, etc)
  - GBs-TBs of annotation per gene or pathway
- Thousands of Deployed Instances
  - Thousands of cooperating sites
  - Gigabit class networks
- Update Channels (pub/sub)
  - Thousands (some private, some open)
- Loose synchrony
  - Hourly/Daily/Weekly/Monthly updates

mcs

Example Peer-to-Peer Network Overlay



Proposed Process to Move Forward

- Inventory of stakeholders
  - Email directory of interested parties (goal is broad participation)
- Issue an initial informal RFI (request for information)
  - Requirements for reference an architecture and use scenarios
  - 3-4 meetings resulting in a RFP document
- RFP announcement
  - 90 days (proposals tech/arch/interface)
  - Evaluation of proposals criteria/reviewers
- Draft standard – open architecture – LSG
  - 3-4 meetings digest-negotiation/compromise
- Standard is developed in Chapters – in a "standards book"
  - Each area of the standard has an defined authorship
  - Reviews are by committee of the whole
- Reference Implementation(s)
  - Interoperability and use of profiles
  - Ideally two or more reference implementations should be developed
- Publication – open source
  - Document should be open
  - Reference implementation(s) can/should track the proposals

mcs

Principal Partners and Stakeholders

- Biology and Biomedical Communities
  - Genomics, proteomics, medical imaging, neuroscience and sysbio
- Computer Science Community
  - University and Laboratories
- Industry
  - User community (pharma, bt, discovery, etc.)
  - Technology providers (IBM, Oracle, Sun, HP, etc.)
- Agencies (NIH, NSF, DOE, etc.)
- Standards Organizations
  - GGF, W3C, I3C, ISO(?)
- Professional Societies
  - ISCB, ASM, APS, ACM, IEEE etc.

mcs

Open Issues

- Determining scope of "The Standard"
  - Build on existing technology
  - Not just LS x OGSA
- Core team
  - 4 to 10 people are required for this to have critical mass
  - MPI Forum was successful with about 10 key authors and 40-60 participants
- Fast track process with a meeting every 6 weeks of 2-3 days
  - I propose to host the first meeting in December at Argonne/U Chicago
- Buy-in from stakeholders
  - Interest appears high, particularly in funding agencies and in researcher labs for demonstration of scalability and potential ubiquity of infrastructure
- Sponsorship
  - Under development
- License issues
  - Wide support for open source reference implementation and BSD style licensing
- Time Frame for completion
  - 12 to 18 months is a reasonable target for version 1.0

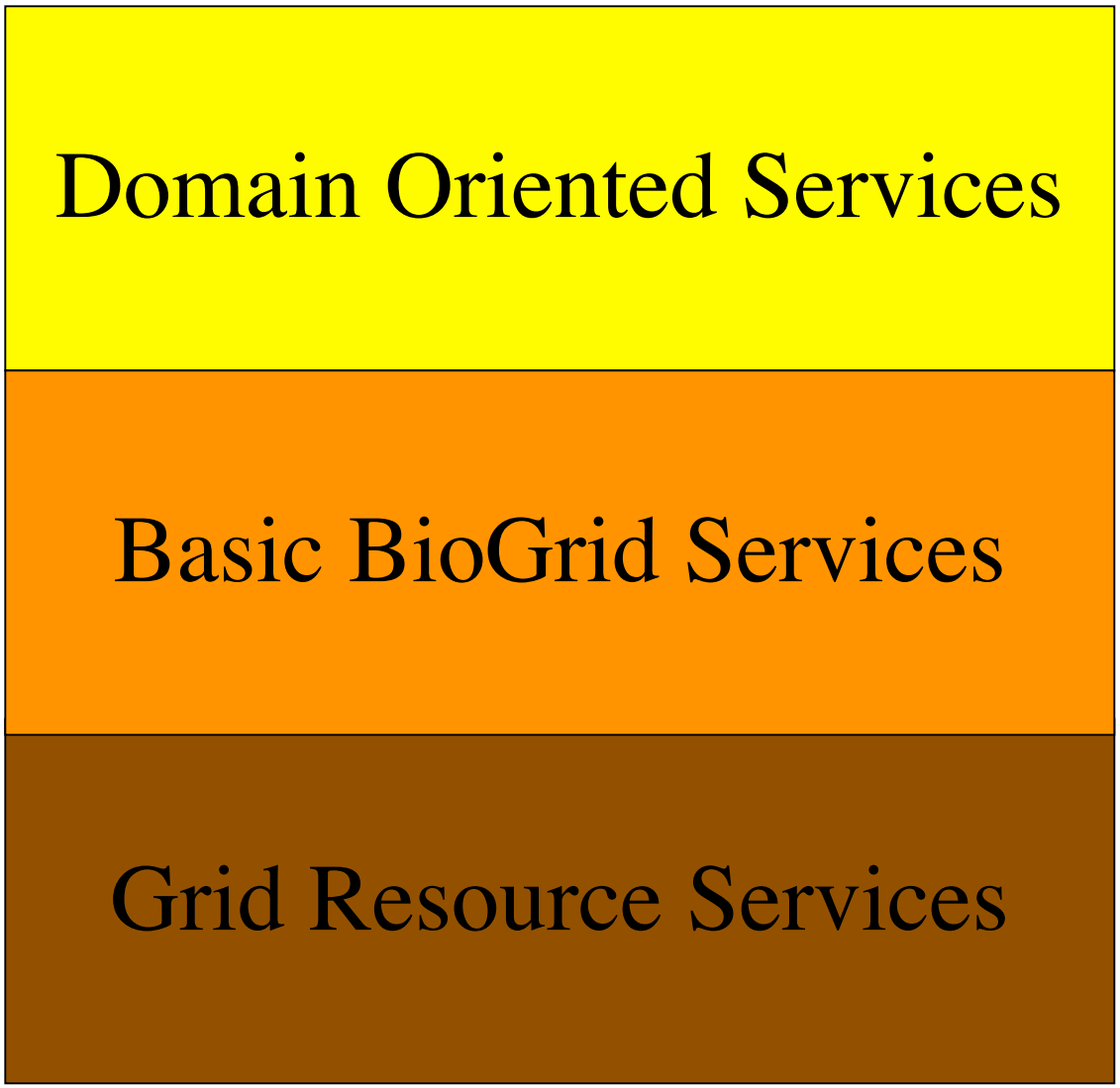
mcs

The Stack

- OLSG Services
  - Discovery, Directory and Data Brokering services
  - IP Access Policy Engine
  - Code and Data Synchronization Update Services
  - Namespace/Ontology Services
  - Sub/Pub “Channel” Subscription Services
  - Computing Services (generic and typed)
  - Web Interfaces
- Grid services
  - Peer-to-Peer services
  - Security
  - Transport
  - Etc.

mcs

An Example BioGrid Services Model



- Drug Discovery
- Microbial Engineering
- Molecular Ecology
- Oncology Research

- Integrated Databases
- Sequence Analysis
- Protein Interactions
- Cell Simulation

- Compute Services
- Workflow Services
- Data Service
- Collaboration Services

Next Steps

- Determine if LSG would like to sign on to this process
- If so, then we need to:
  - Plan a kick-off meeting for developing the scope of a standard
  - Recruit some leadership from the community
  - Recruit some support/sponsorship for the effort
  - Process to develop some white papers for the first meeting

mcs