The Evolution of Computing and Data Grids for Science and Engineering

William E. Johnston

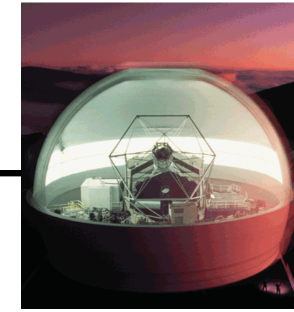
http://dsd.lbl.gov/~wej/

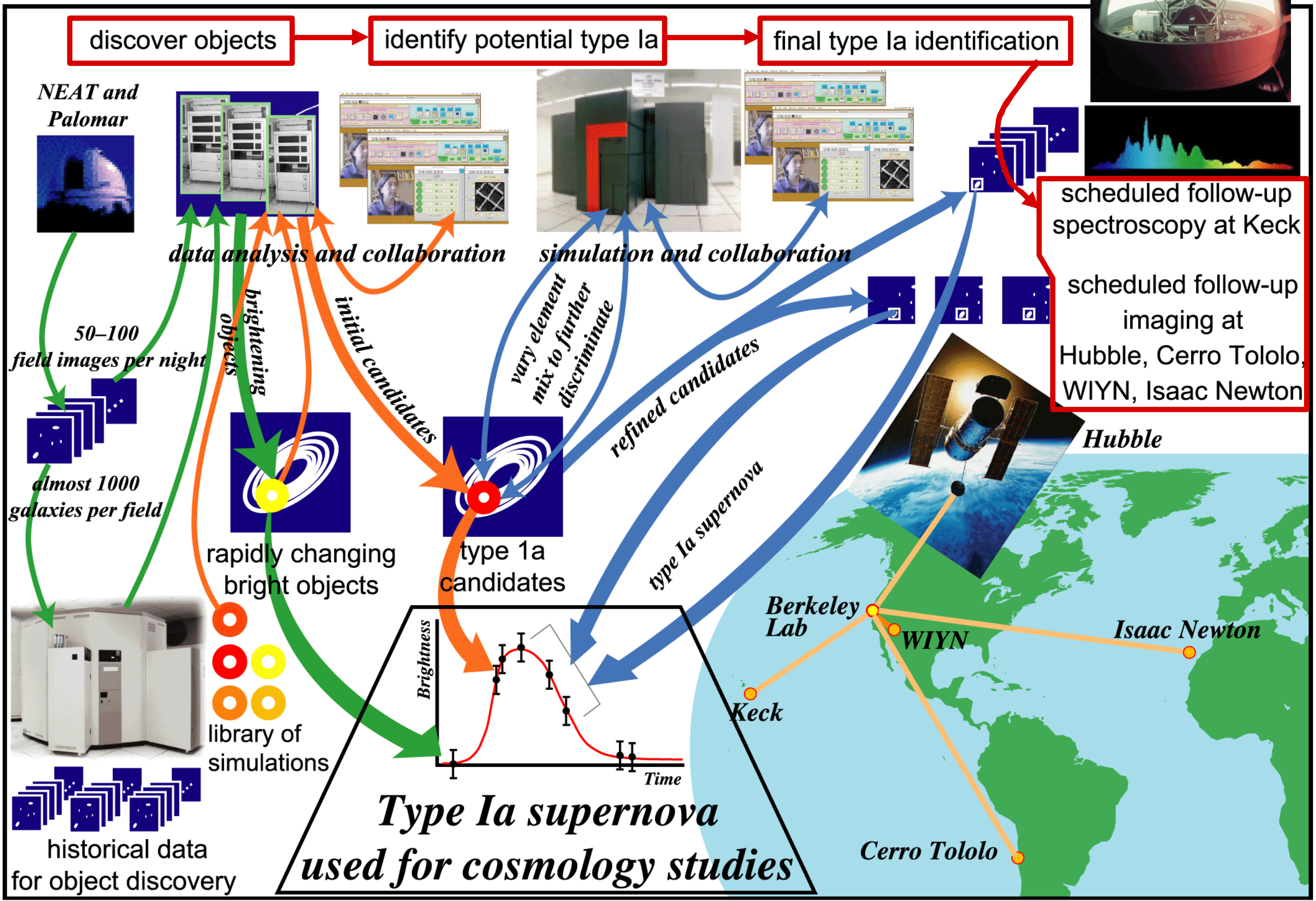
Computing Sciences and ESnet,
Lawrence Berkeley National Laboratory,
U. S. Dept. of Energy

Evolution

- Applications seen as candidates for Grid environments have evolved from
  - complicated, but straightforward workflows, to
  - data management intensive, to
  - multidisciplinary simulations
- This has gone hand-in-hand with the evolution of Grid services

Complicated Workflow – Many Sites

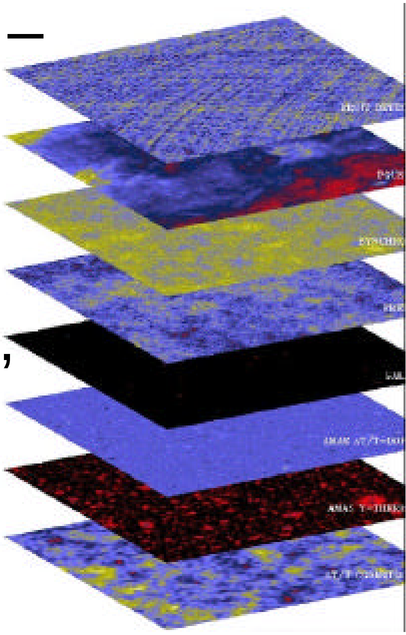




Supernova Cosmology Project, Perlmutter, et al. (http://www.supernova.lbl.gov)

Data Management Intensive

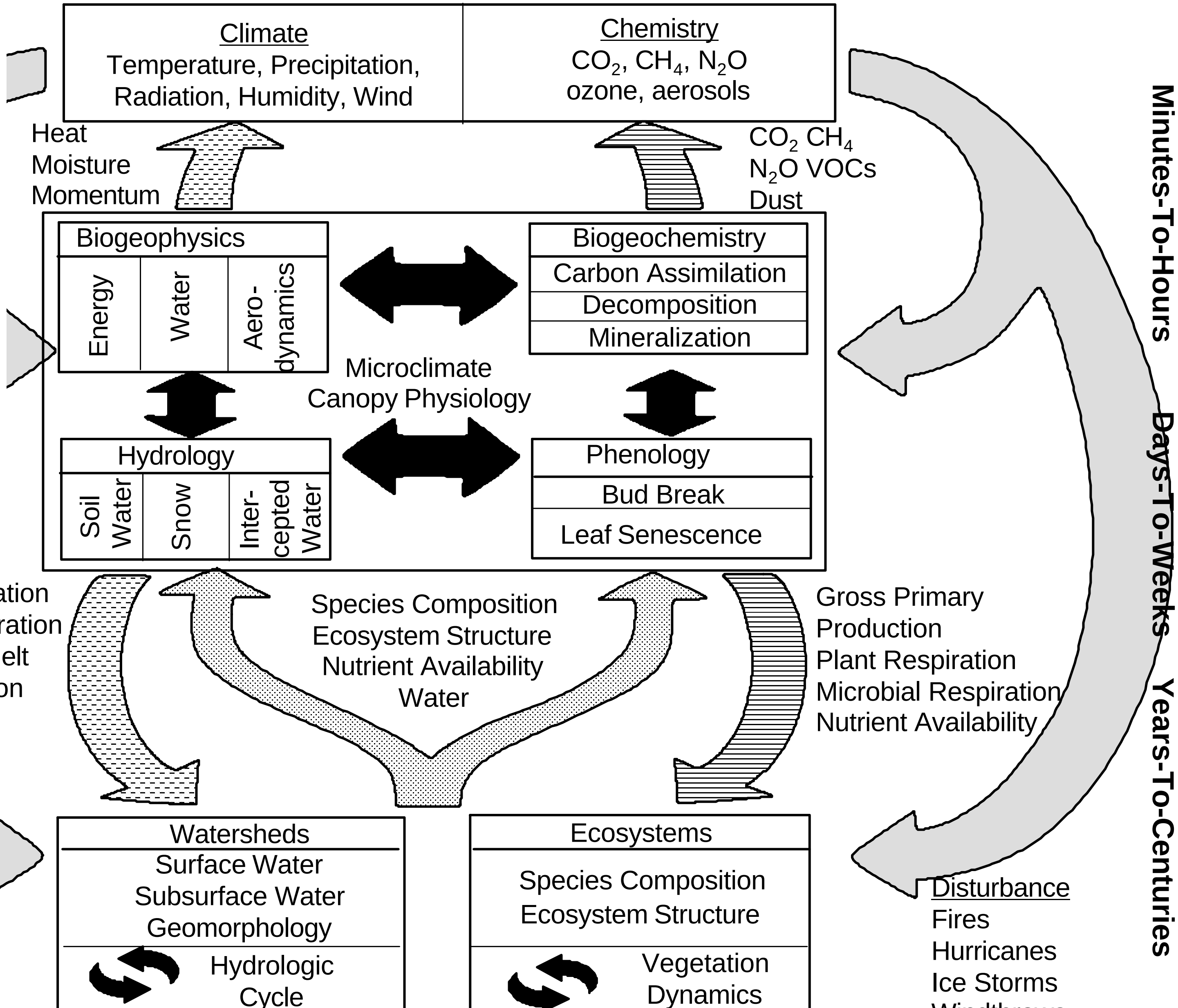
- The ability to federate astrophysics survey data is enormously important
- Studying the Cosmic Microwave Background – a key tool in studying the cosmology of the universe – requires combined observations from many instruments in order to isolate the extremely weak signals of the CMB
- The datasets that represent the material “between” us and the CMB are collected from different instruments and are stored and curated at many different institutions
- This is immensely difficult without approaches like National Virtual Observatory in order to provide a uniform interface for all of the different data formats and locations



(Julian Borrill, NERSC, LBNL)

Multidisciplinary Simulation

A "complete" approach to climate modeling involves the many interacting processes and data of terrestrial biogeoscience that are modeled by different groups at different locations



(Courtesy Gordon Bonan, NCAR: *Ecological Climatology: Concepts and Applications*. Cambridge University Press, Cambridge, 2002.)

➢Progress in Grids

- The focus of grids has clearly evolved since the late 1990s
- The early goals of uniform access to computing resources and a uniform security model was a reasonable place to start, but did not provide enough added value to justify users investing a lot of time learning new infrastructure
- However, as the large-scale data management capabilities emerged, and the integration with Web Service started, interest picked up
- The increasing interest was, however, somewhat driven by a shift in the target application communities

NASA’s Information Power Grid

A Vision for Large Scale Distributed Computing and Data Management

William E. Johnston, Project Manager

Arsi Vaziri, Deputy Project Manager

Tom Hinke, Deployment Task Manager

Leigh Ann Tanner, Implementation Manager

The NAS Division
NASA Ames Research Center

William J. Feiereisen, Division Chief

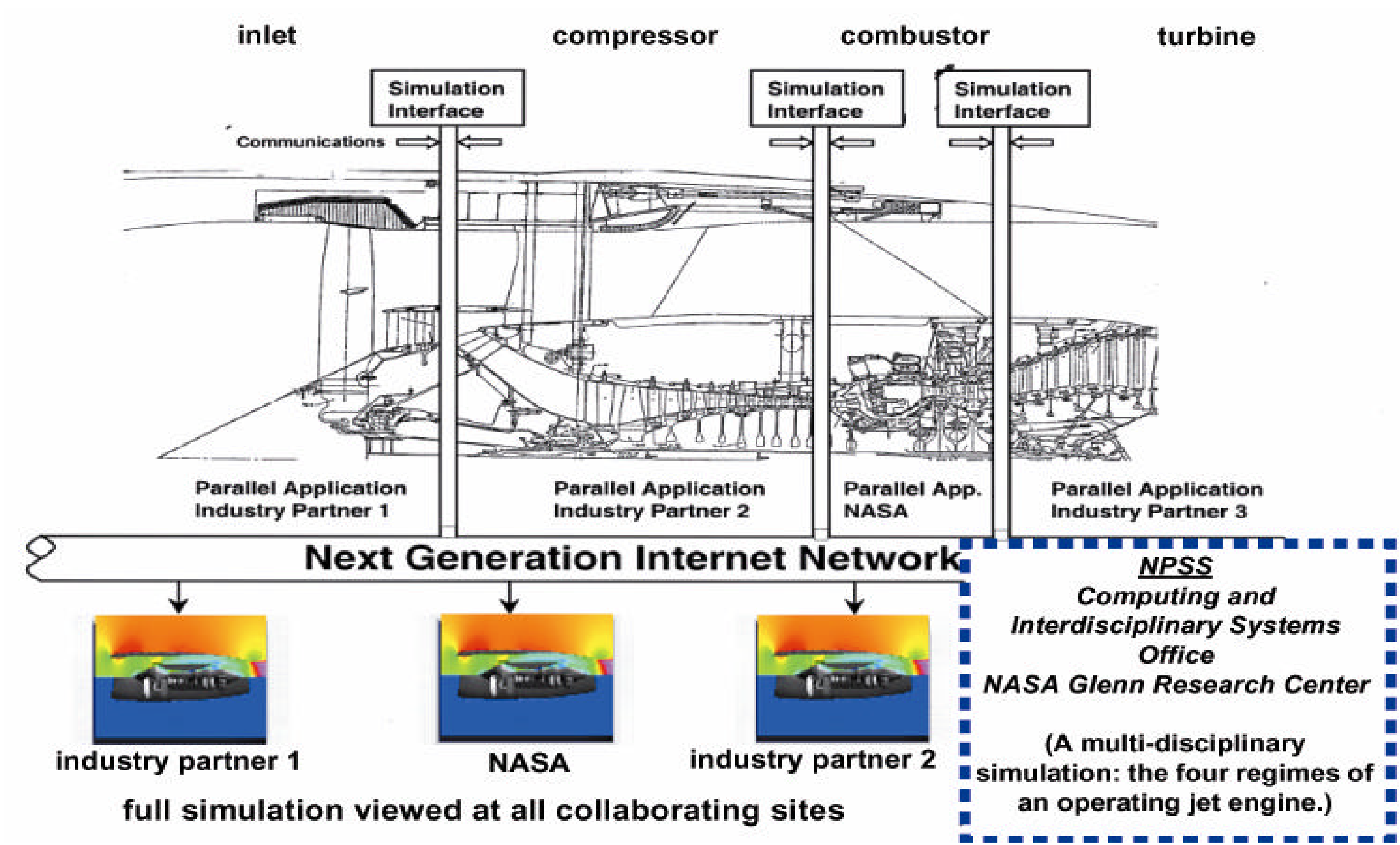
William Thigpen, Engineering Branch Chief,

http://www.ipg.nasa.gov

➢Simulation Workflow

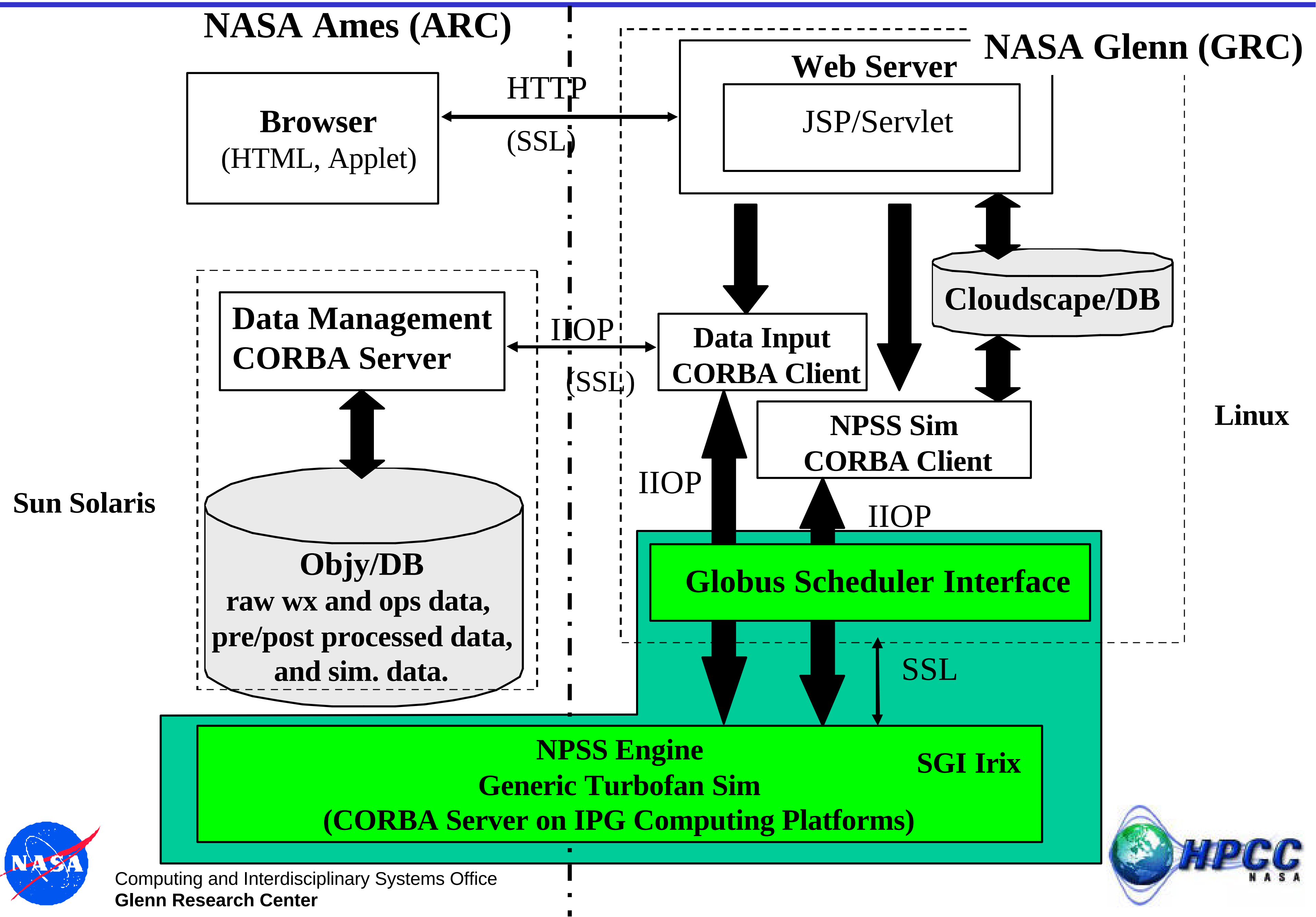
- Multi-component simulations involve executing multiple, coupled, medium to large scale simulations on multiple computing resources
- Grids provide co-scheduling and data stream management to support large scale pipelined applications

Simulation Workflow



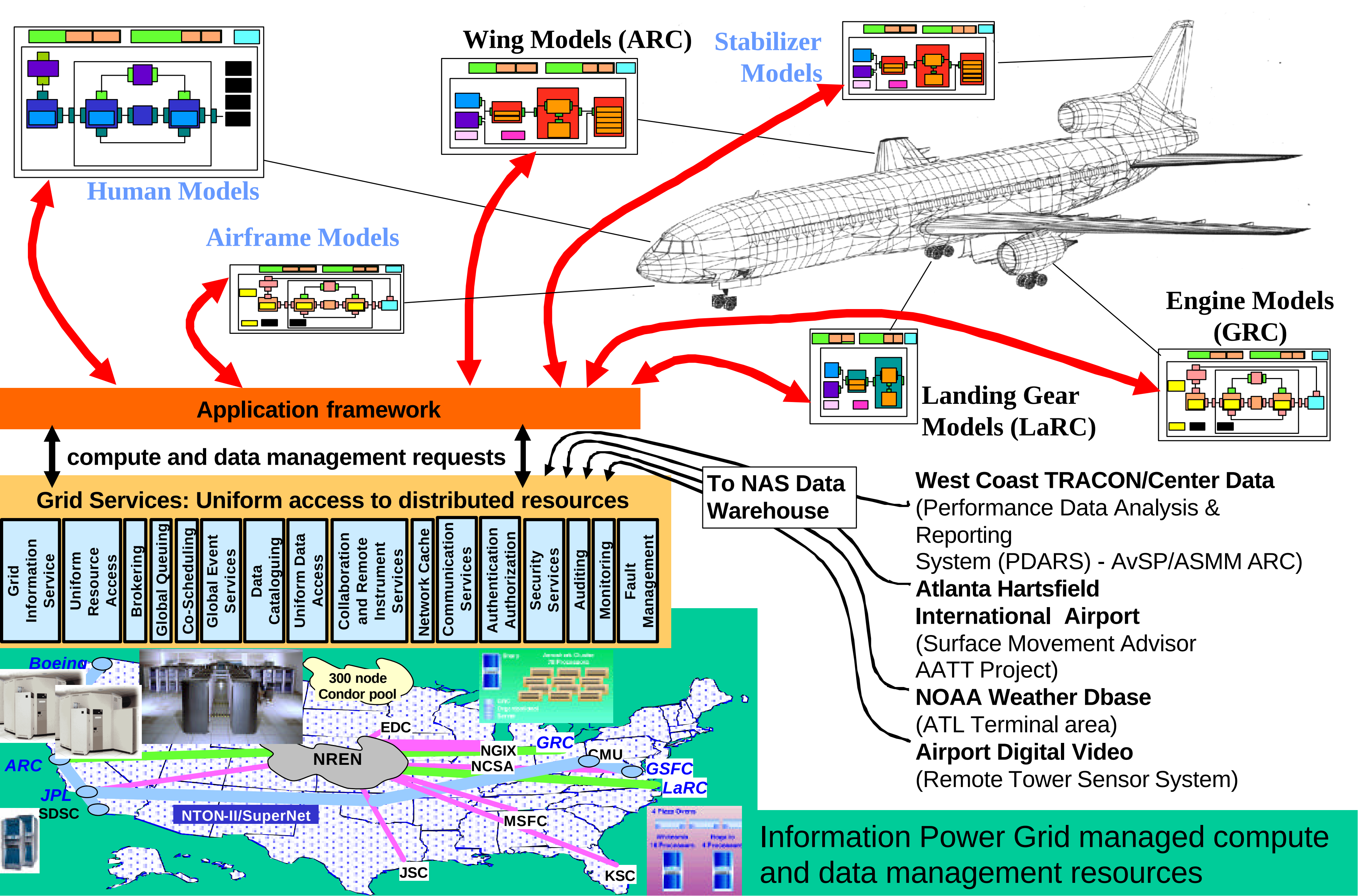
Component simulations are combined to get sub-system simulations

NPSS Data Sharing and Resource Access Architecture



2001

Vision for Grids Supporting Aviation Safety Simulations



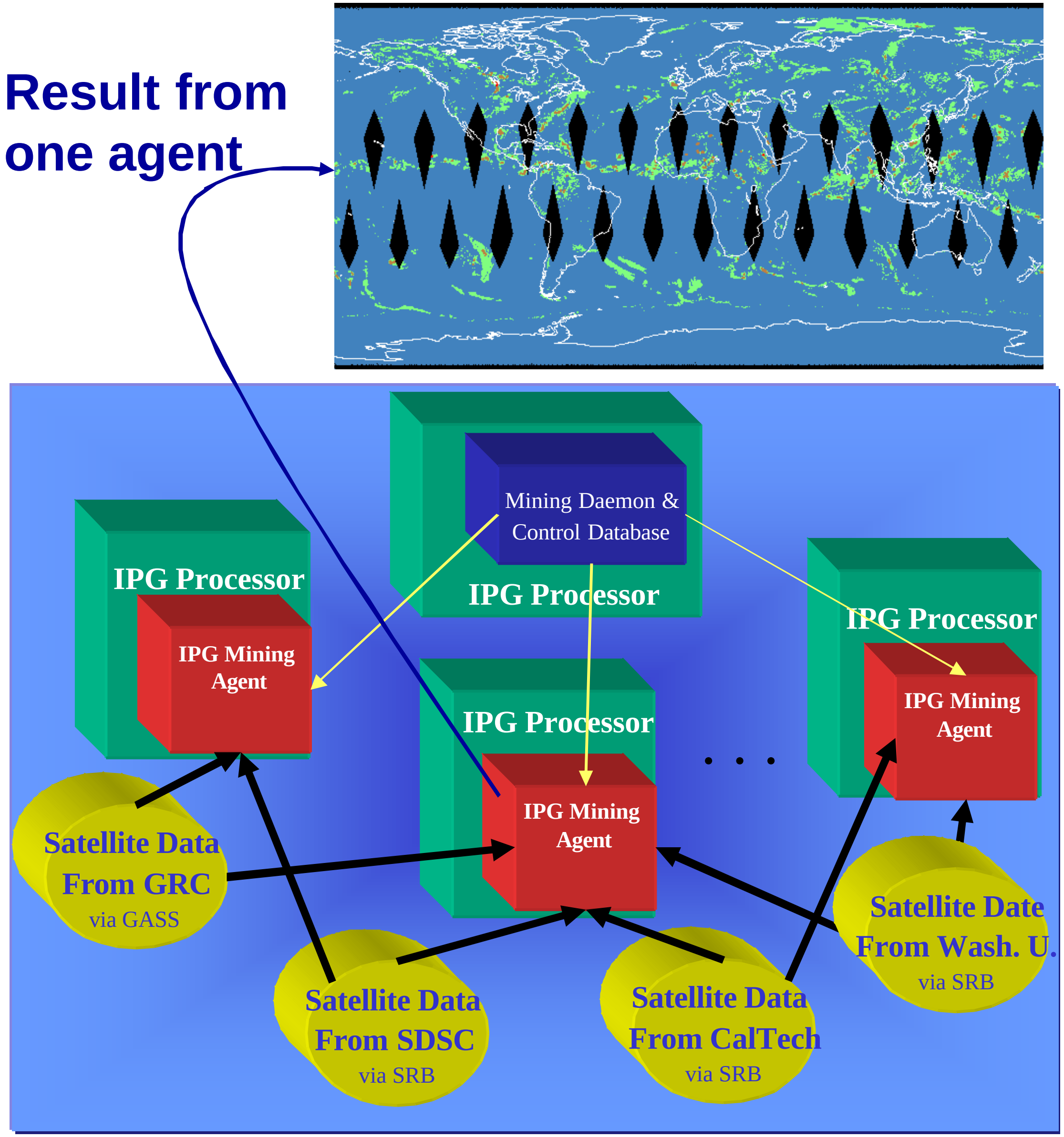
➢Data Management Intensive Applications

- Grids provide the tools and middleware for discovering and access data archives that are maintained by discipline experts at many different organizations.

1999

High Speed Distributed Data Access: IPG Milestone Completed 3/2000

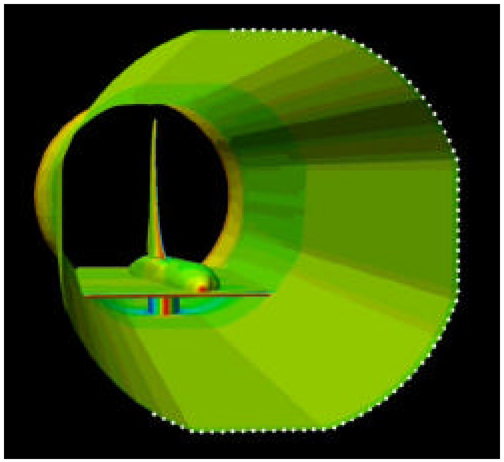
- Agent based parallel data mining
- 512 node SGI Origin at Ames uses IPG data access tools to simultaneously mine hydrology data from four sites
- Grid provides
  - uniform data access
  - uniform access to computing to start the agent servers



2000

Grids as Meta-Computers: IPG Milestone Completed 12/2000

high-lift subsonic wind tunnel model



The research branch of NAS is investigating algorithms that are suitable for a Grid computing “meta-platform.” One candidate is overset grid codes that can tolerate timestep mis-matches on the intra-object boundaries. A version of the OVERFLOW, Navier-Stokes, CFD simulation code is being modified for this approach. It has been demonstrated operating across systems at ARC, GRC, and LaRC, solving for flow about large test objects mounted in a wind tunnel.

Ames
Moffett Field, CA

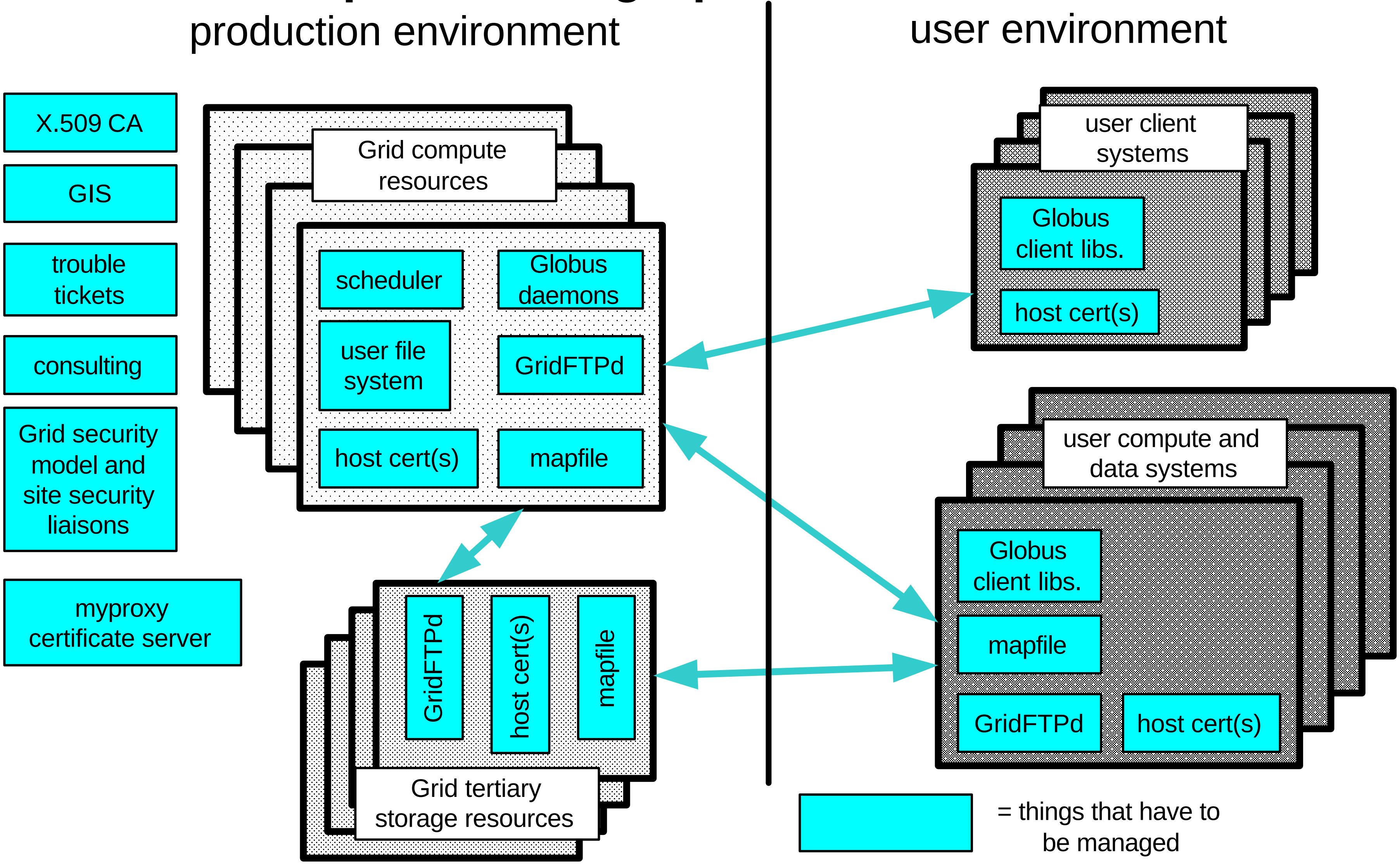
Lomax
512 node SGI Origin 2000

Application POC: Mohammad J. Djomehri

Some IPG Grid Successes

- Standardized access to multi-institutional resources
- A common security approach and infrastructure
- Generic persistent services (Globus job management and scheduling) that are *used to run application frameworks on an as-needed basis*
  - CORBA (this case)
  - CONDOR job manager ("Glide-in")
  - Agent systems / servers (data mining example)
- This allows users great flexibility in building their applications in the framework of their choice. *They do not have to rely on that framework being provided as a persistent service on all of the computing systems where they need to run* – they can instantiate their own environment using persistent Grid services.

Steps to Setting Up a Multi-Site Grid

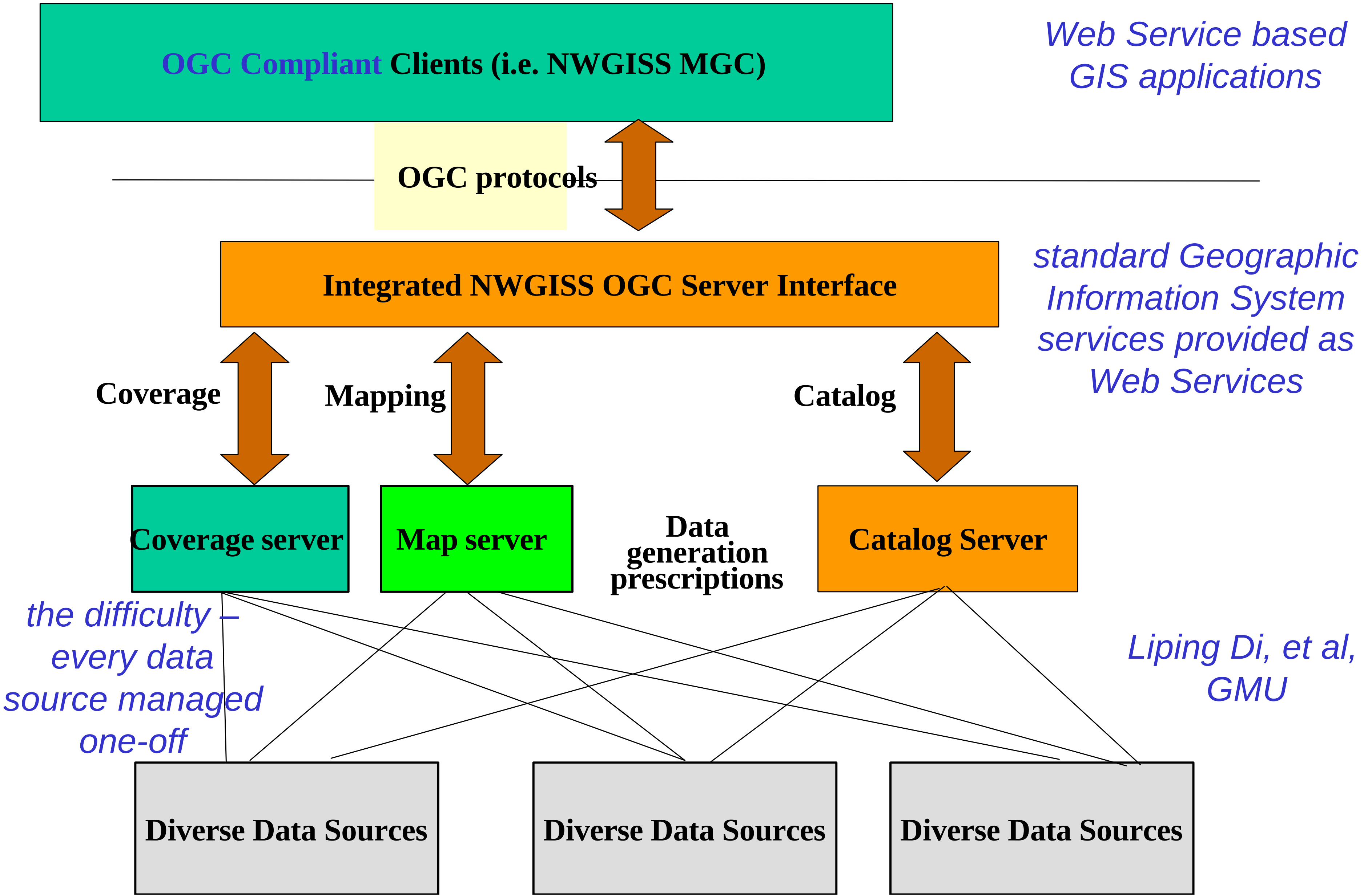


Establish Your Grid Service Model

➢Building More Complex Data and Simulation Systems: Combining Web Services and Grids

- The Web Services approach provides
  - Language independence
  - Module composition
  - Self- describing interfaces
  - A natural integration with Web browsers
  - Commercial tools

Existing Prototypical Earth Sciences, Web Services Based, Information Technology System

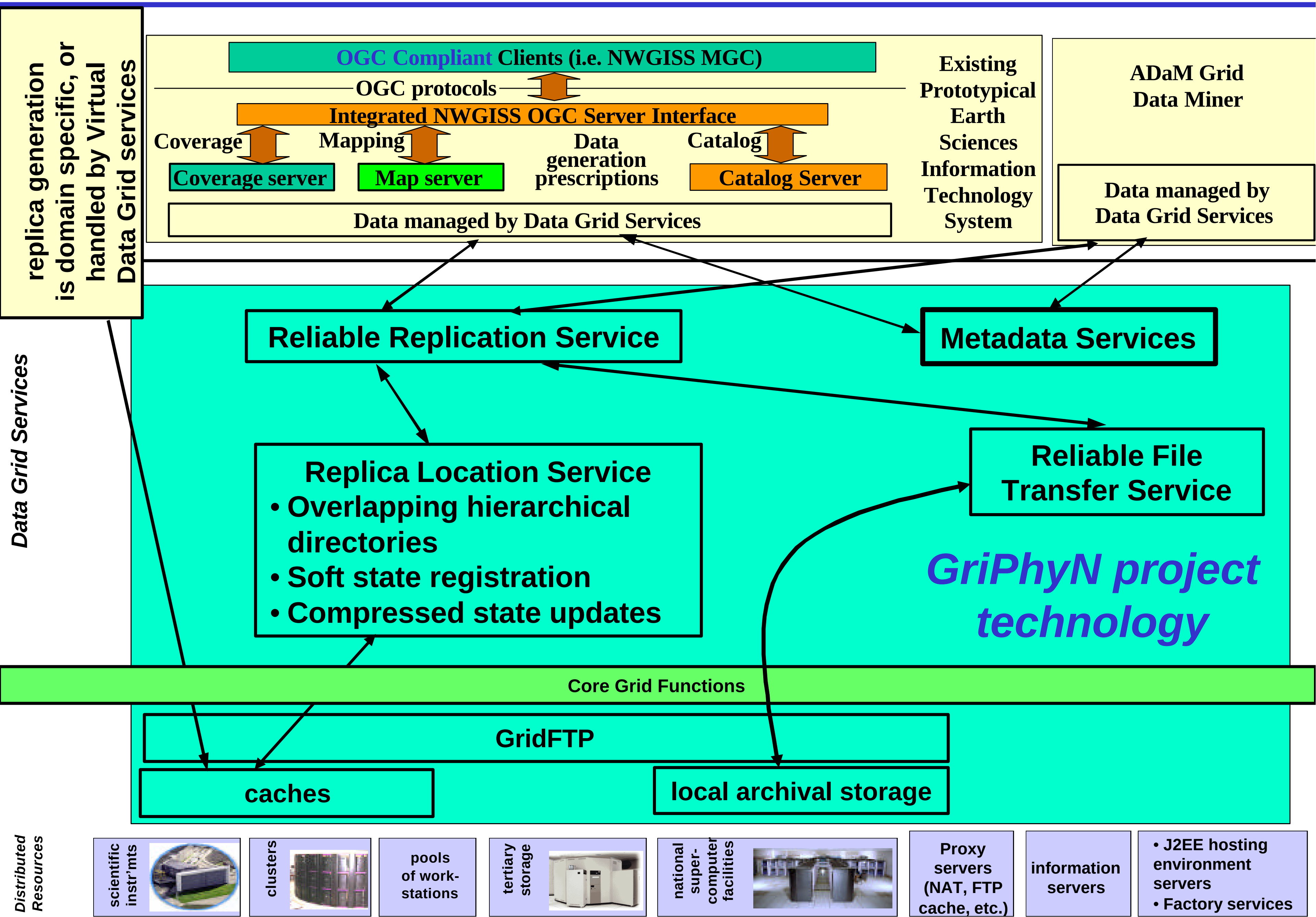


2003

Using Grid Services to Manage Data

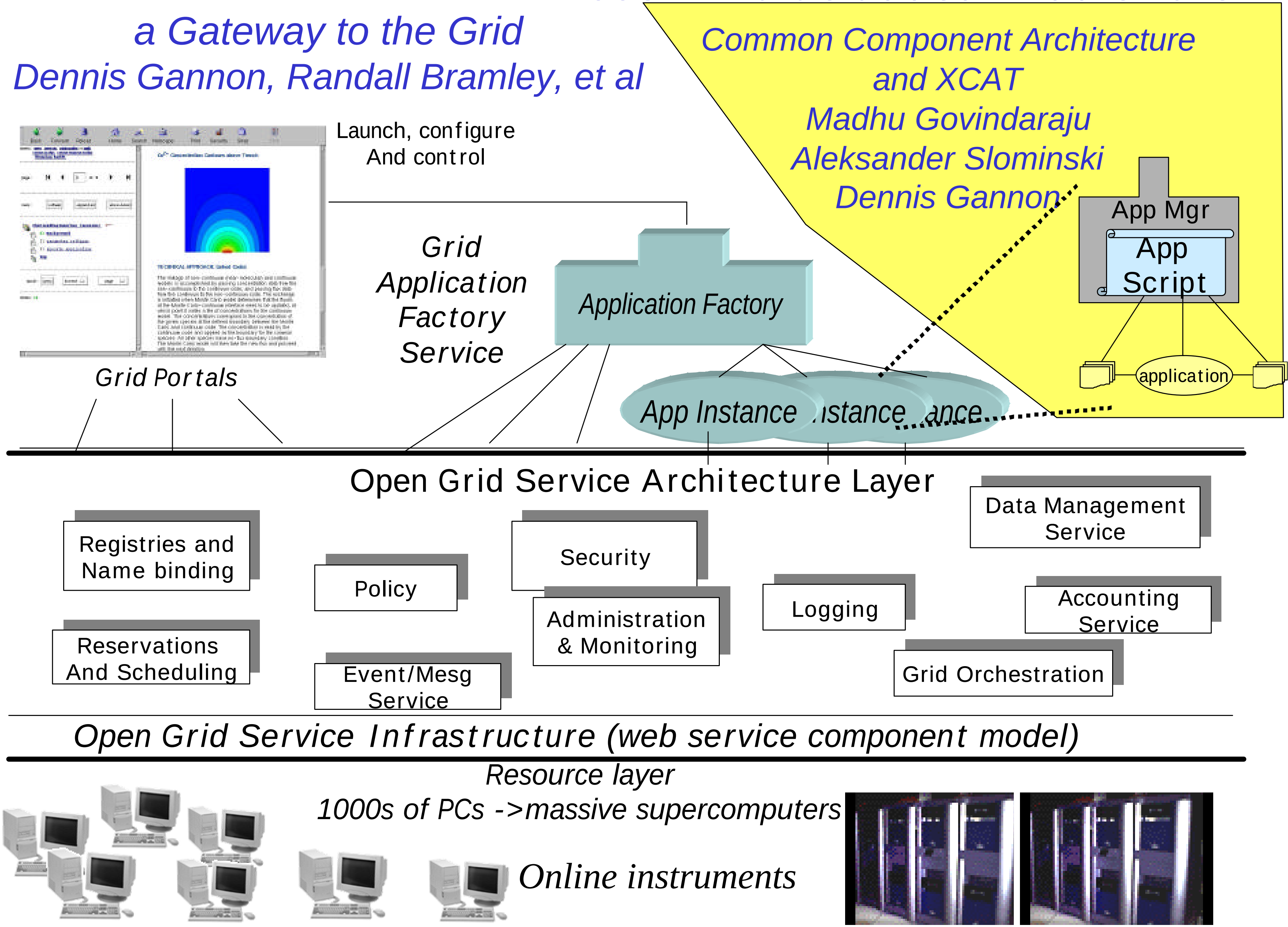
- NASA AIST Grant: Integration of OGC and Grid Technologies for Earth Science Modeling and Applications – Liping Di, GMU; W. Johnston, Ames; D. Williams, LLNL
- Rich and capable <u>metadata description</u> approach
  - All of the Grid Data Services turn on the ability of metadata to represent all of the essential data characteristics for all relevant data
- Capable and scalable <u>naming transparency</u> mechanism to support federating distributed data archives
  - Persistent logical dataset names regardless of physical storage
- Reliable and scalable <u>data location transparency</u> through replica management
  - A logical file may represent identical data at several different locations in order to allow for optimizing access by many different users – that is, the service returns the physical file name that is best suited (usually fastest access) for a user

Naming and Location Transparency



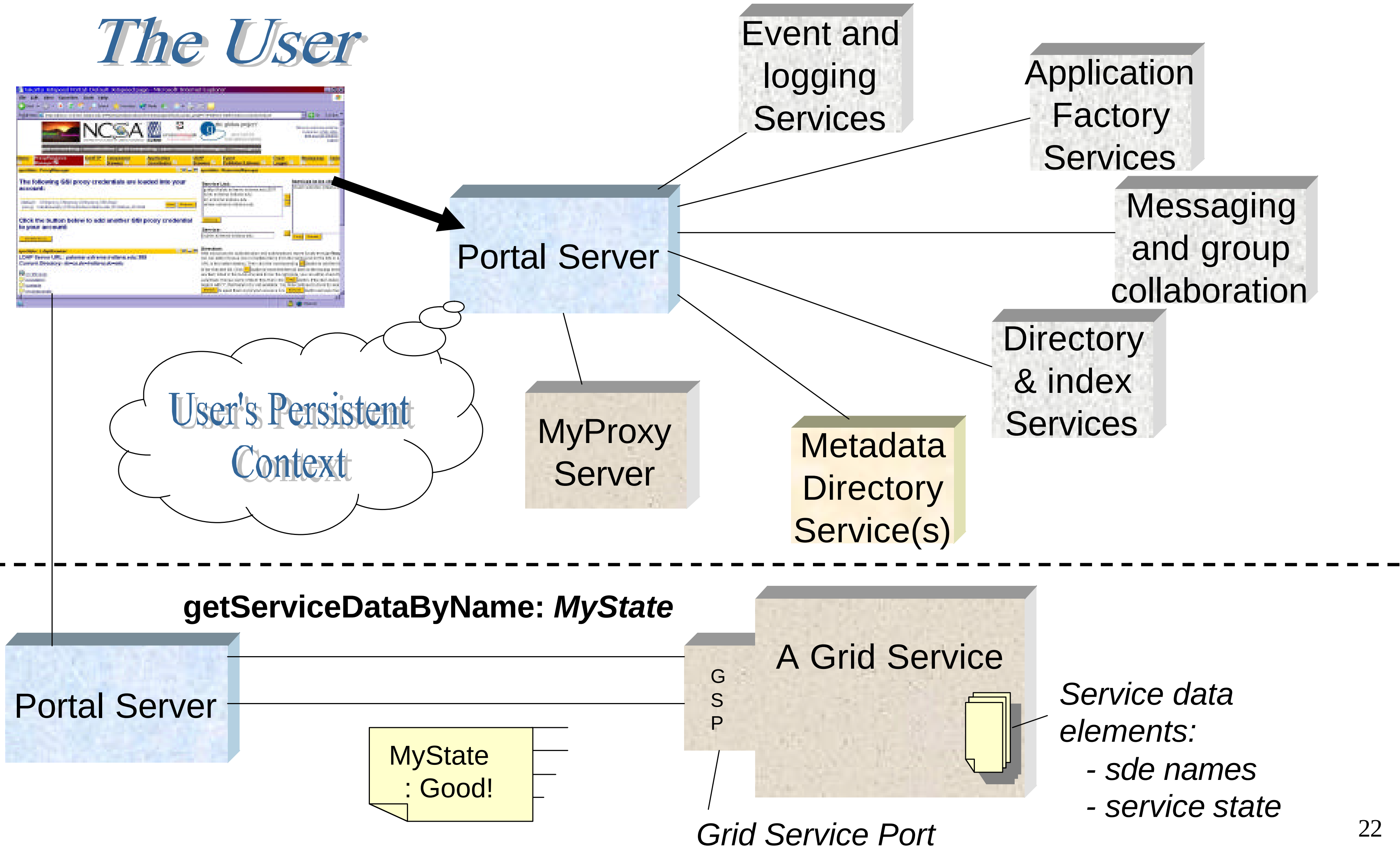
Science Portals:
a Gateway to the Grid
Dennis Gannon, Randall Bramley, et al

2003-4: Value added middleware



The Big Picture *(with apologies to Dennis, et al)*

- The Grid is defined by a collection of distributed Services
  - The portal is the user's access point to orchestrating these resources



➢Where to in the Future? The Semantic Grid

- Even when we have well integrated Web+Grid services we still do not provide enough structured information and tools to let us ask "what if" questions
  - have services that assemble the required components in a consistent way to answer such a question <u>without the user having to be an expert in all of the disciplines needed</u> to build the multidisciplinary system

Beyond Web Services and Grids

- A commercial example "what if" question:
  - What does my itinerary look like if I wish to go SFO to Paris, CDG, and then to Bucharest.
  - In Bucharest I want a 3 or 4 star hotel that is within 3 km of the Palace of the Parliament, and the hotel cost may not exceed the U. S. Dept. of State, Foreign Per Diem Rates.

Beyond Web Services and Grids

- To answer such a question – a relatively easy task, but tedious, for a human – the system must <u>“understand” the relationships</u> between maps and locations, between per diem charts and published hotel rates, and it must be able to apply constraints (< 3 km, 3 or 4 star, cost < $ per diem rates, etc.)
- This is the realm of “Semantic Services”

Semantic Services

- In *“Semantic Services” for Grid Based, Large-Scale Science** I argue that there is a progression of capabilities that need to be built up in order to realize the benefit of a componentized science simulation environment
  - we need various discipline specific semantic models for operations and data to provide automatic query structuring

IEEE Intelligent Systems, Special Issue on e-Science

Acknowledgements

- This work was funded by the U.S. Dept. of Energy, Office of Science, Office of Advanced Scientific Computing Research, Mathematical, Information, and Computational Sciences Division under contract DE-AC03-76SF00098 with the University of California and by NASA's Aero-Space Enterprise, Information Technology (IT) program