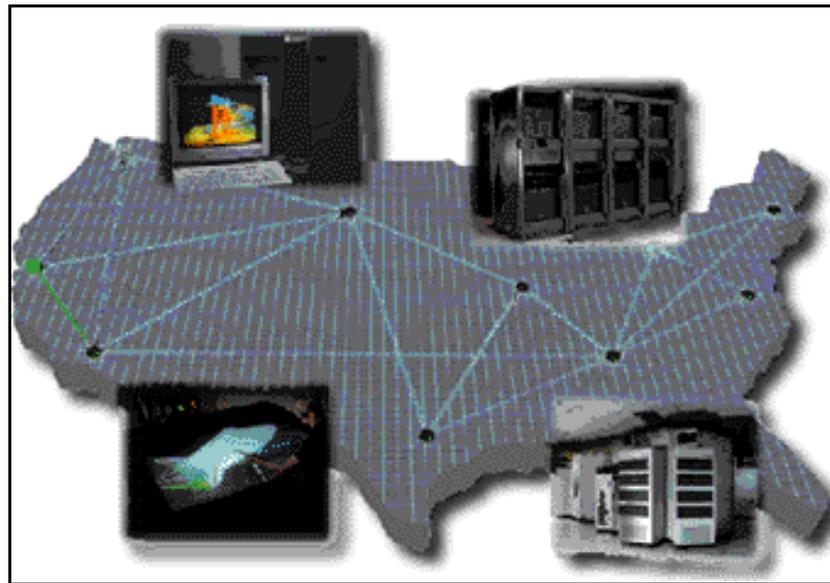


# Information Power Grid: Distributed High-Performance Computing, Large-Scale Data Management, and Collaboration Environments for Science and Engineering

*William E. Johnston, Dennis Gannon, William Nitzberg, William Van Dalsem*



**Numerical Aerospace Simulation Facility at NASA Ames Research Center**

*William J. Feiereisen, Division Chief*

**<http://nas.nasa.gov/~wej/home/IPG>**

- ◆ **Prototype applications have demonstrated both the potential and the “reality” of high-speed, service based, wide area distributed systems.**
- ◆ **However, a comprehensive and consistent set of the services are needed to *routinely* build, operate, and manage *transient and widely distributed* applications.**

**Together with an operational infrastructure, this environment is called a “grid” [1].**

- ◆ **The Information Power Grid project is developing and evolving these technologies into a *prototype production* computational and data grid, providing the infrastructure for widely distributed systems.**

# **Requirements**

**General capability and services requirements come from experience with the way science and engineering R&D uses computer related resources. Specific IPG requirements come from analyzing NASA Aerospace Engineering Systems and Earth Sciences, Data Assimilation Office applications. Together these give a fairly comprehensive set of requirements.**

## **Grids will provide:**

- Application capabilities**
- Distributed resource access**

## **Grids must provide:**

- Scalability**
- Usability**

**These are the points that will be explored in this talk.**

# **Grids Provide Four Application Related Capabilities**

- ◆ **Distributed Computation**
- ◆ **Data Intensive Computing**
- ◆ **Problem Solving Environments**
- ◆ **Computer mediated human collaboration**

**These capabilities are related in that they:**

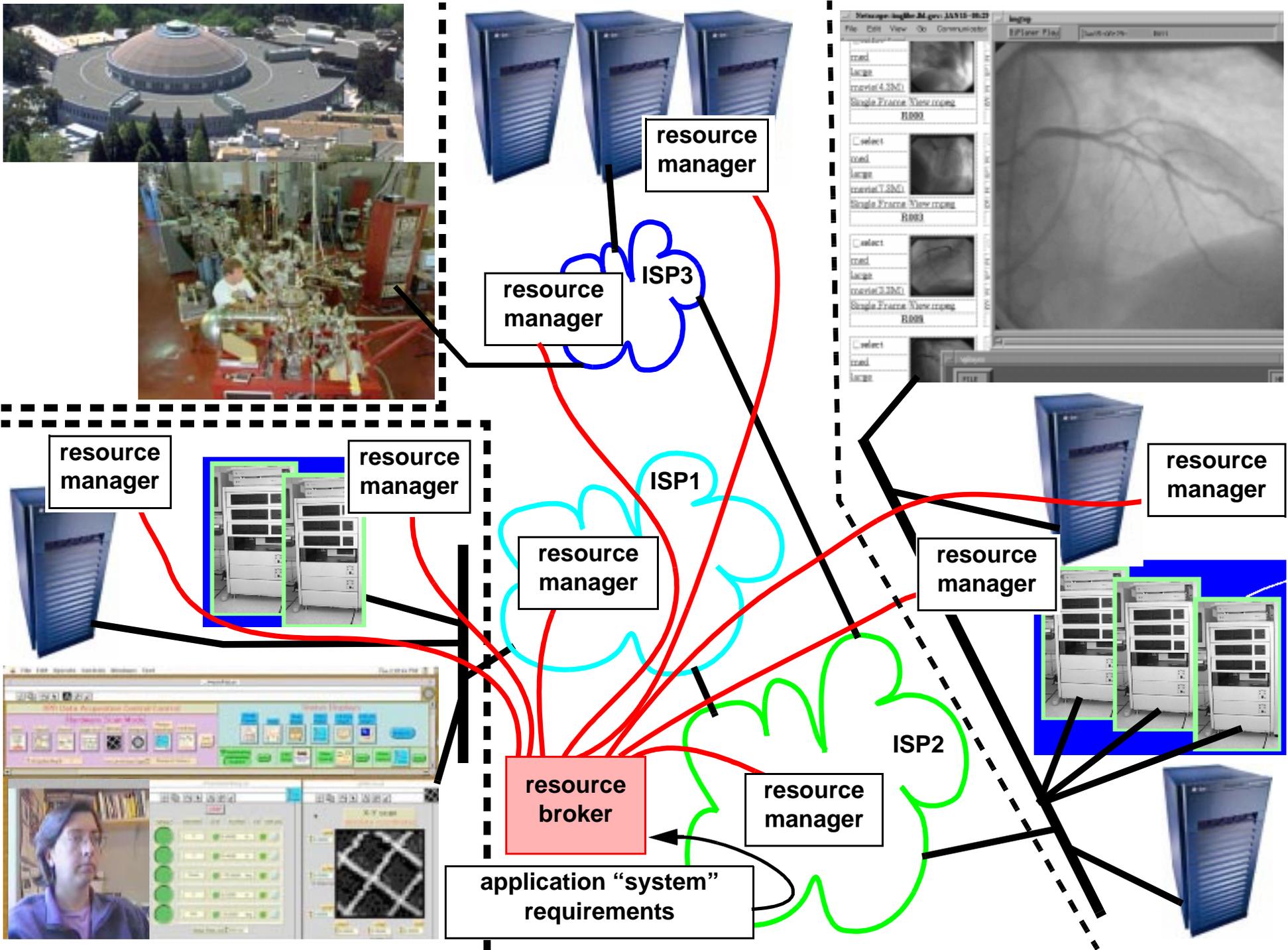
- **are all needed to support grid based science and engineering**
- **have common requirements for the supporting infrastructure**

**This talk focuses on data intensive computing.**

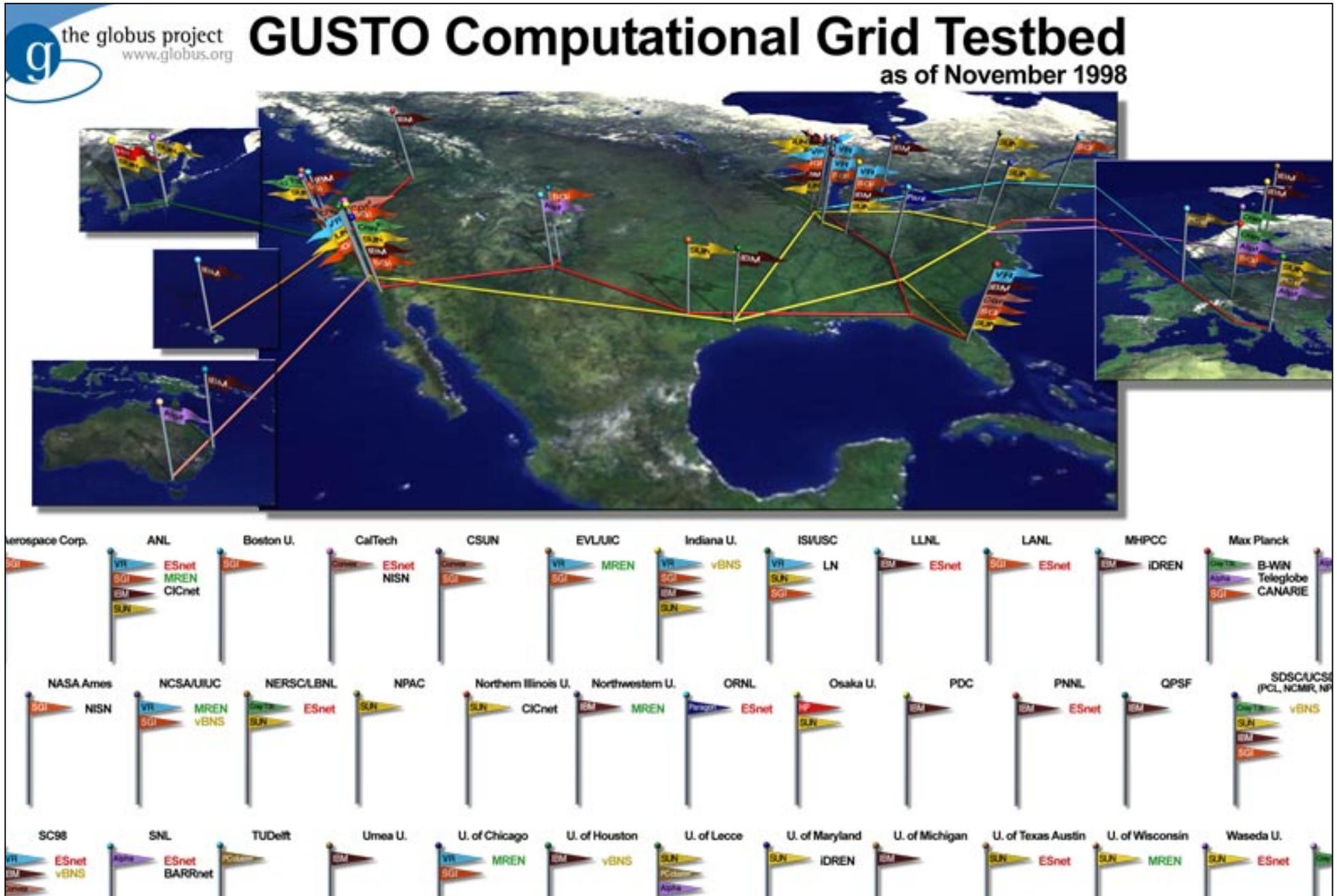
# **Grids are Intended to Support “Large-Scale” Applications**

**“Scale” refers to several dimensions:**

- large computational and storage capacity through aggregation of resources**
- complexity of resources (independent of capacity)**
  - data intensive computing tends to require a complex mix of resources, with or without high capacity**
  - grids are intended to provide transparent access to these resources**



- geographic and organizational scope (see [2])



- **scale also refers to the abstractions needed by different grid user communities**

<b>audience</b>	<b>needed services/interfaces</b>
<ul style="list-style-type: none"> <li>» Lay public, schools</li> <li>» Community emergency services</li> <li>» Military</li> </ul>	Web browser / kiosk
Application domain scientists and engineers	Problem Solving Environments / application frameworks
Application domain computational scientists and tool developers	Middleware supporting: <ul style="list-style-type: none"> <li>• distributed computation</li> <li>• aggregated/federated access to catalogued data</li> <li>• computer mediated collaboration</li> </ul>
Distributed system developers	Job management, access control, generalized communications services, resource discovery and brokering
Middleware / grid common service developers	Local resource managers (queuing, network QoS, scheduled tape marshaling), security services, network services, resource information bases

# Examples

Each example\* was designed, developed, and debugged “end-to-end, top-to-bottom” - this is essential for high performance applications in wide area distributed environments:

- *end-to-end*: the full scope of the application from data generation and management through computation to user interface (all typically remote from each other) is addressed
- *top-to-bottom*: means that every aspect of the distributed system from data storage/generation and CPU elements, all the way down through the network fabric, is monitored, evaluated, and refined

***This comprehensive “system approach” is essential for making widely distributed applications work, and grid services are intended to facilitate this approach.***

---

\*Several of these examples are not NASA applications, however they all represent - through IPG - various ARC partnerships.

## **Example: Real-Time Digital Libraries for On-Line, High Data-Rate Instruments [3]**

**Goal was to provide cardiac care physicians with immediate access to major patient studies, rather than having to wait for weeks.**

**Demonstrated technologies:**

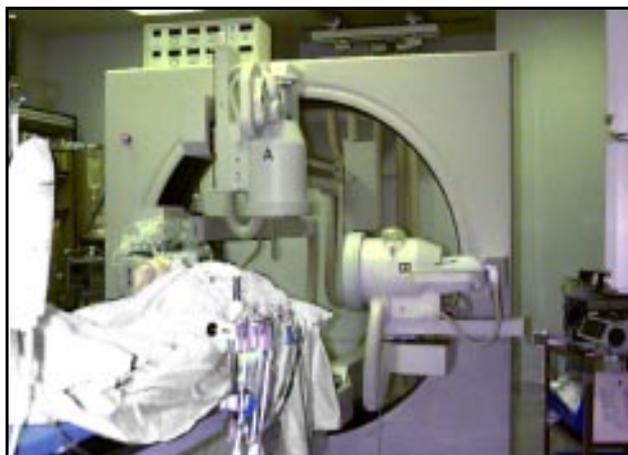
- **High data-rates and large data sets in wide area**
- **Optical WDM metropolitan area network (now part of NGI)**
- **On-line, real-time, high data-rate instrument**
- **Remote data analysis**
- **Remote automatic data cataloguing and archiving**
- **Remote data users**
- **Widely distributed, “application” level cache**
- **A “data flow” architecture for high data-rate on-line instrumentation systems**

WALDO real-time digital library system and DPSS distributed cache [4] for data cataloguing and storage



Compute servers for data analysis and transformation

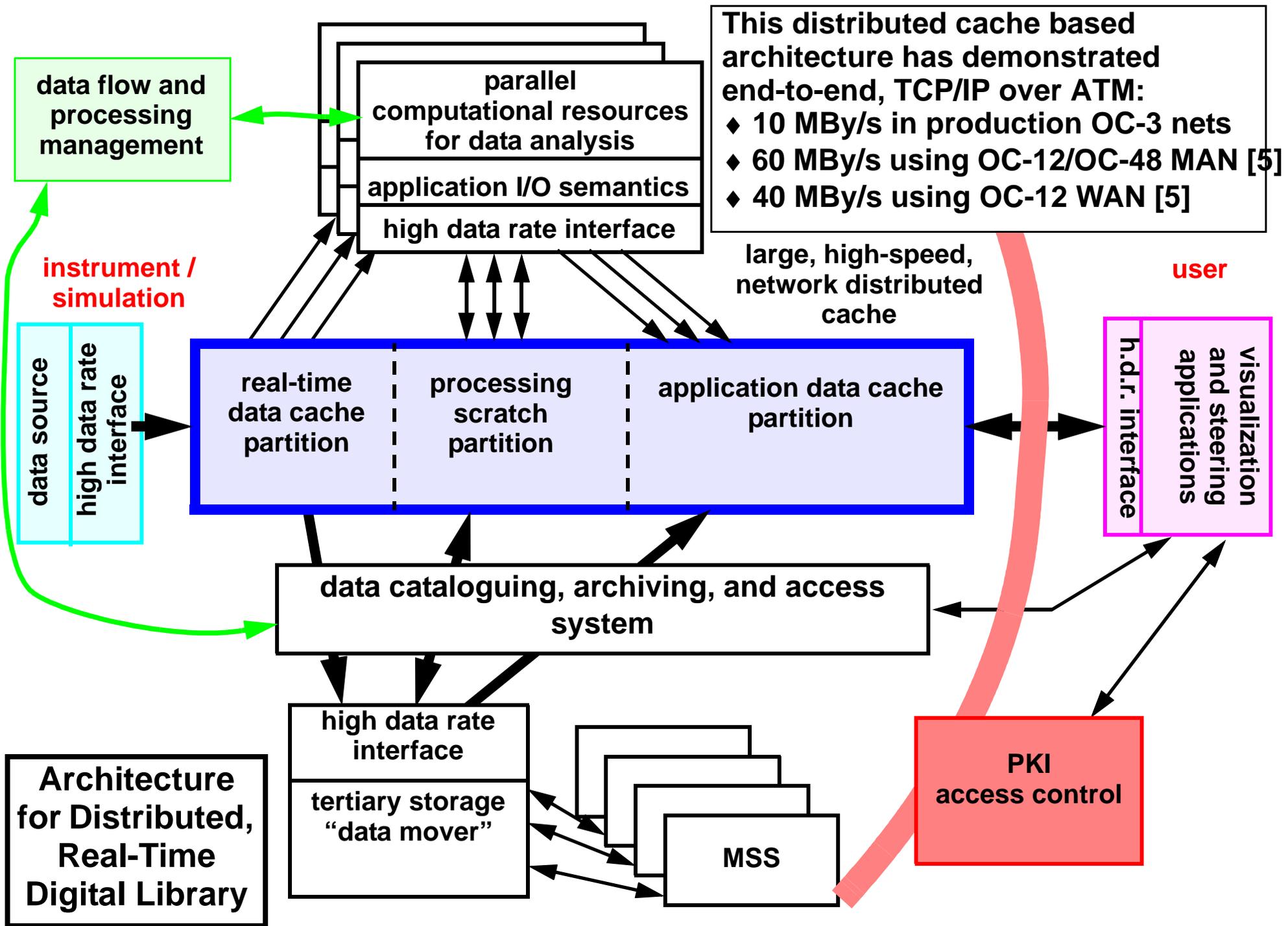
Kaiser San Francisco Hospital Cardiac Catheterization Lab (X-ray video imaging system,  $\approx 130$  mbit/s, 50% duty cycle 8-10 hr/day)



The PSE: Automatically generated user interfaces providing indexed access to the large data objects (the X-ray video) and to various derived data.



Lawrence Berkeley National Laboratory and Kaiser Permanente Health Care On-line Health Care Imaging Experiment in the San Francisco Bay Area

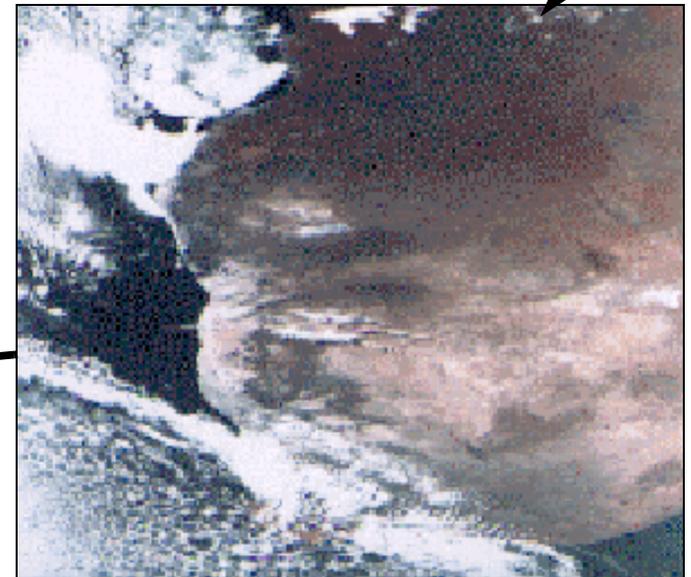


# **Example: High Data-Rate Distributed Data Management and Federated Access for Satellite and Aerial Imagery, Digital Terrain Data, and Atmospheric Data [6]**

- ◆ On-line, real-time access to multiple environmental data sets that are (and always will be) maintained by domain experts at their own sites.
- ◆ On demand, real-time interactive exploration of an operational environment (military, community emergency services)
- ◆ Aggregation of multiple, widely distributed, multi-discipline data sets
- ◆ DARPA MAGIC testbed consortium (see [www.magic.net](http://www.magic.net)) developed distributed services, data and visualization from EROS Data Center, NCAR, NAVO, SRI (collab. with NASA NREN)
- ◆ MAGIC wide-area, gigabit network testbed is now part of NGI

Landscape represented by  
tiled images and terrain at  
EROS Data Center

11	12	13	14	15	16	17
21	22	23	24	25	26	27
31	32	33	34	35	36	37
41	42	43	44	45	46	47
51	52	53	54	55	56	57
61	62	63	64	65	66	67
71	72	73	74	75	76	77



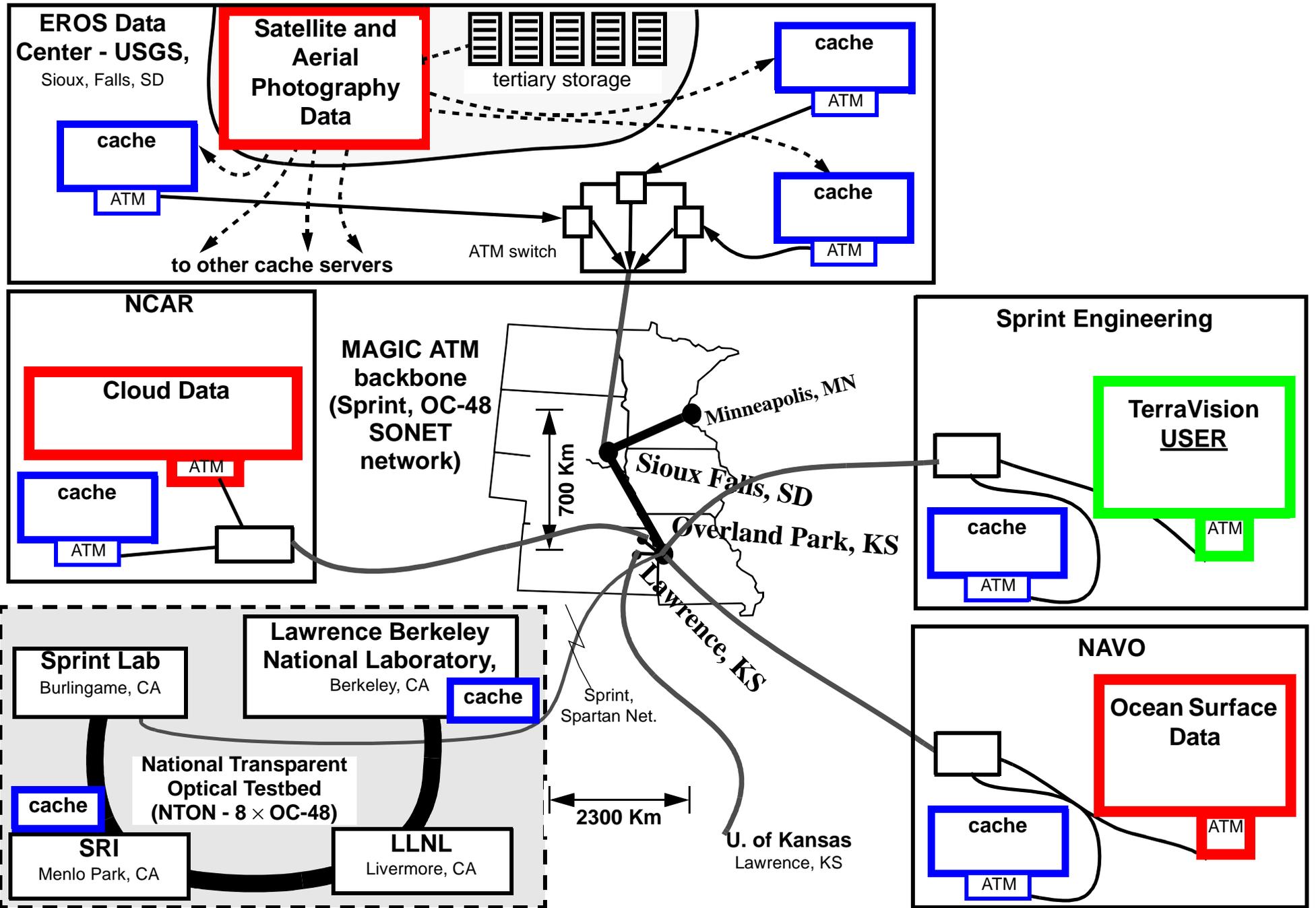
Path of travel

TerraVision produces a  
accurate visualization of  
the landscape

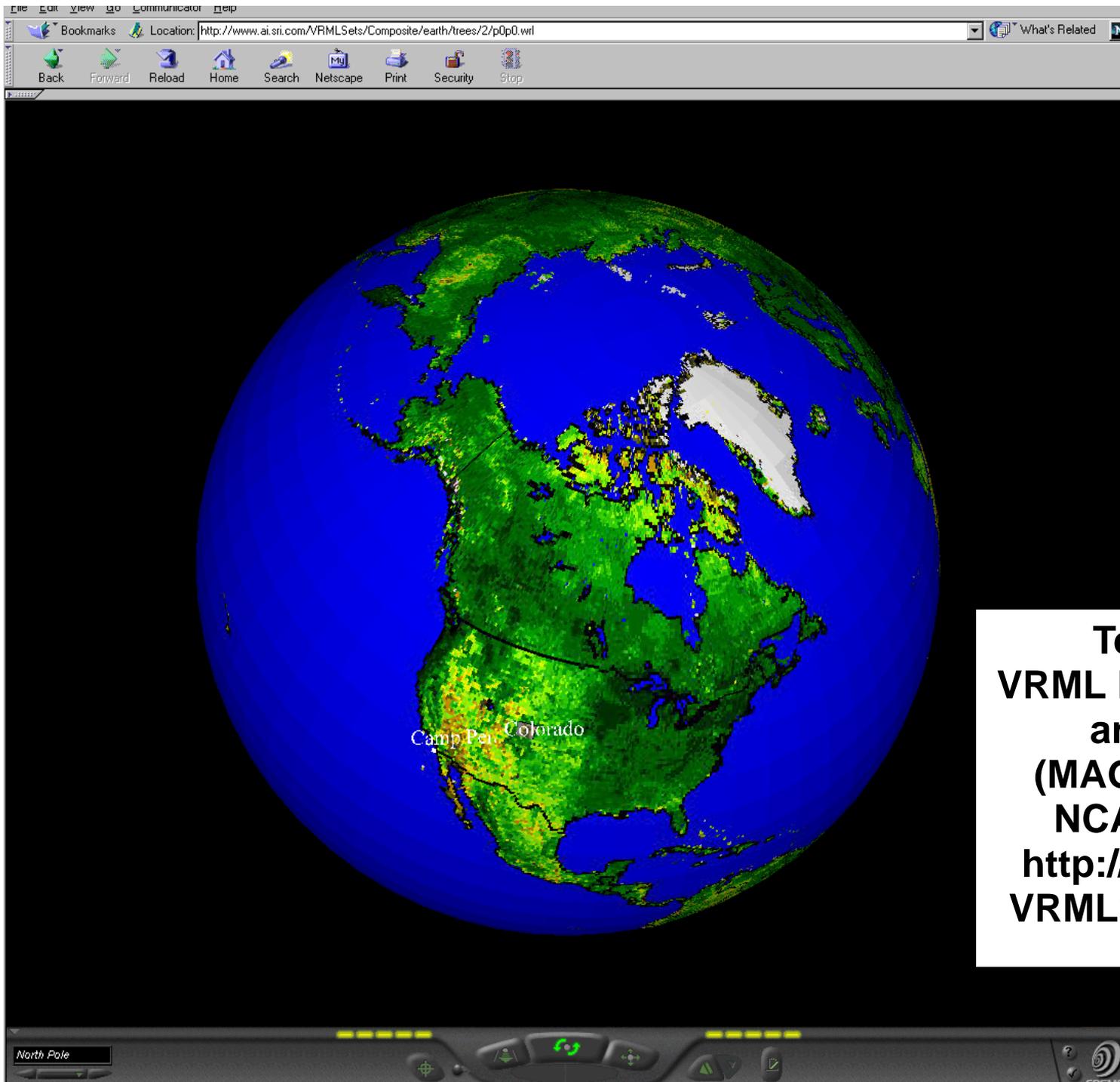
Human user  
navigates  
(controls path  
of travel)



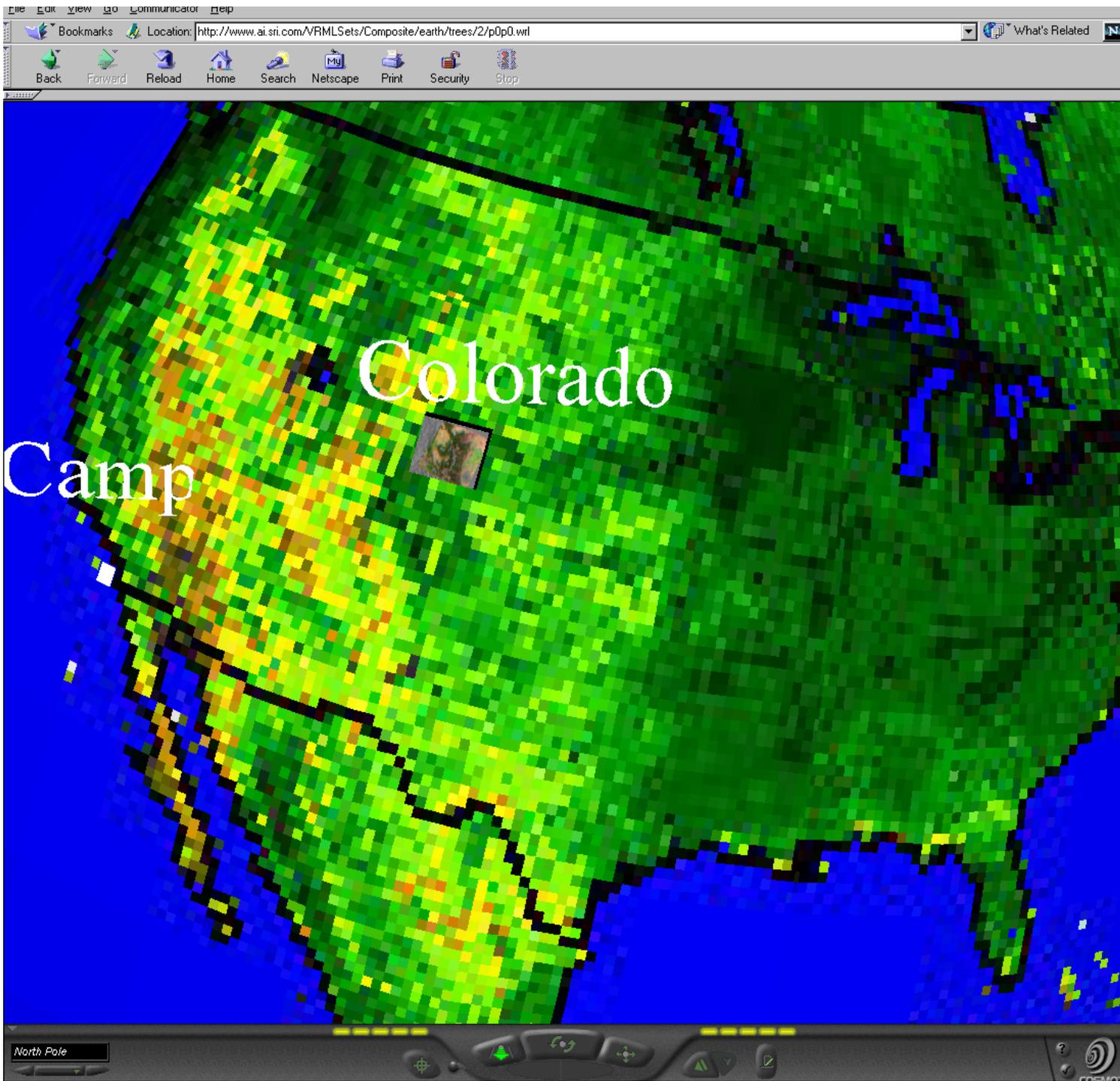
## TerraVision Provides Real-time Visualization of Aggregated Data



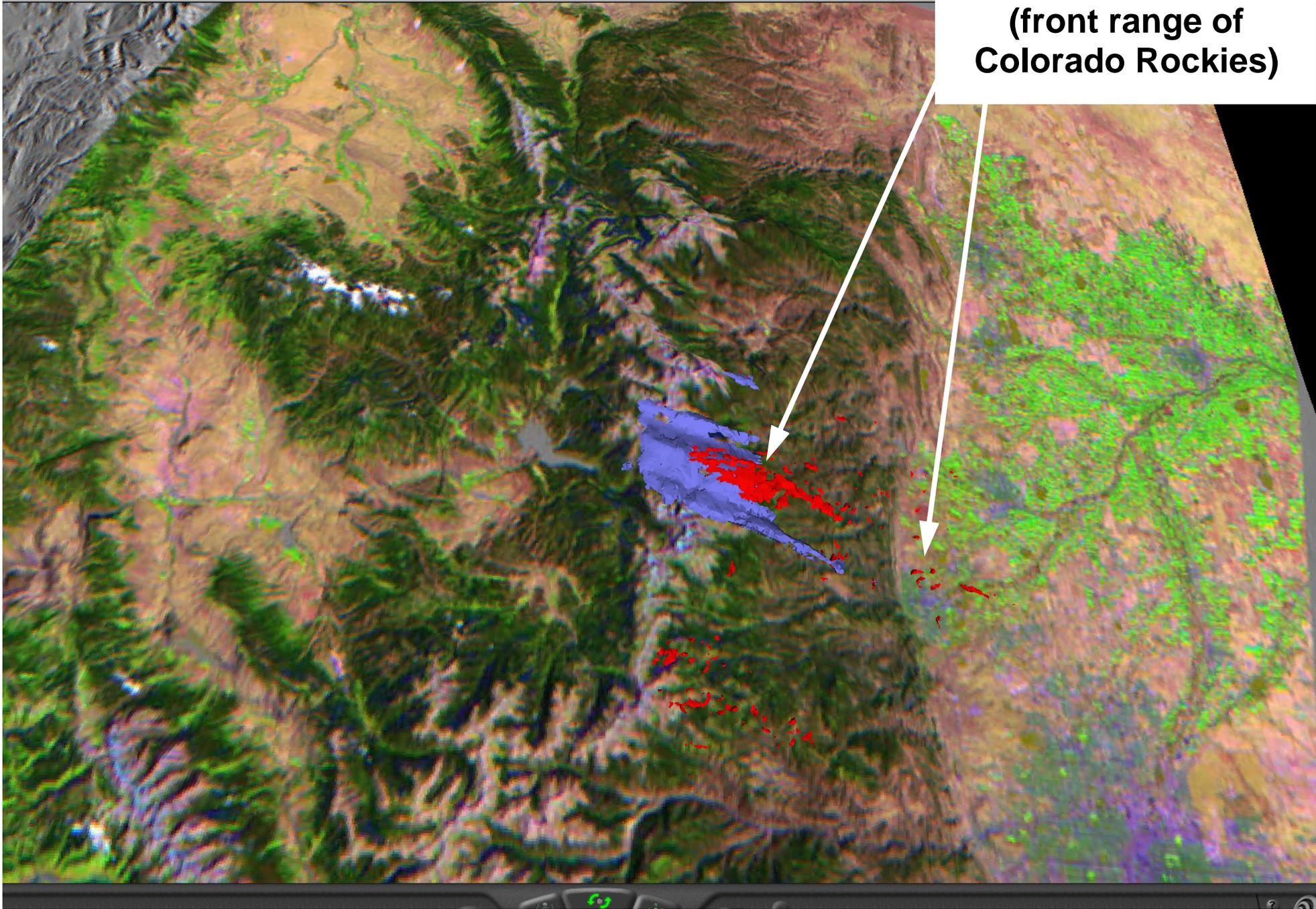
**The MAGIC Testbed Distributed Application Environment**



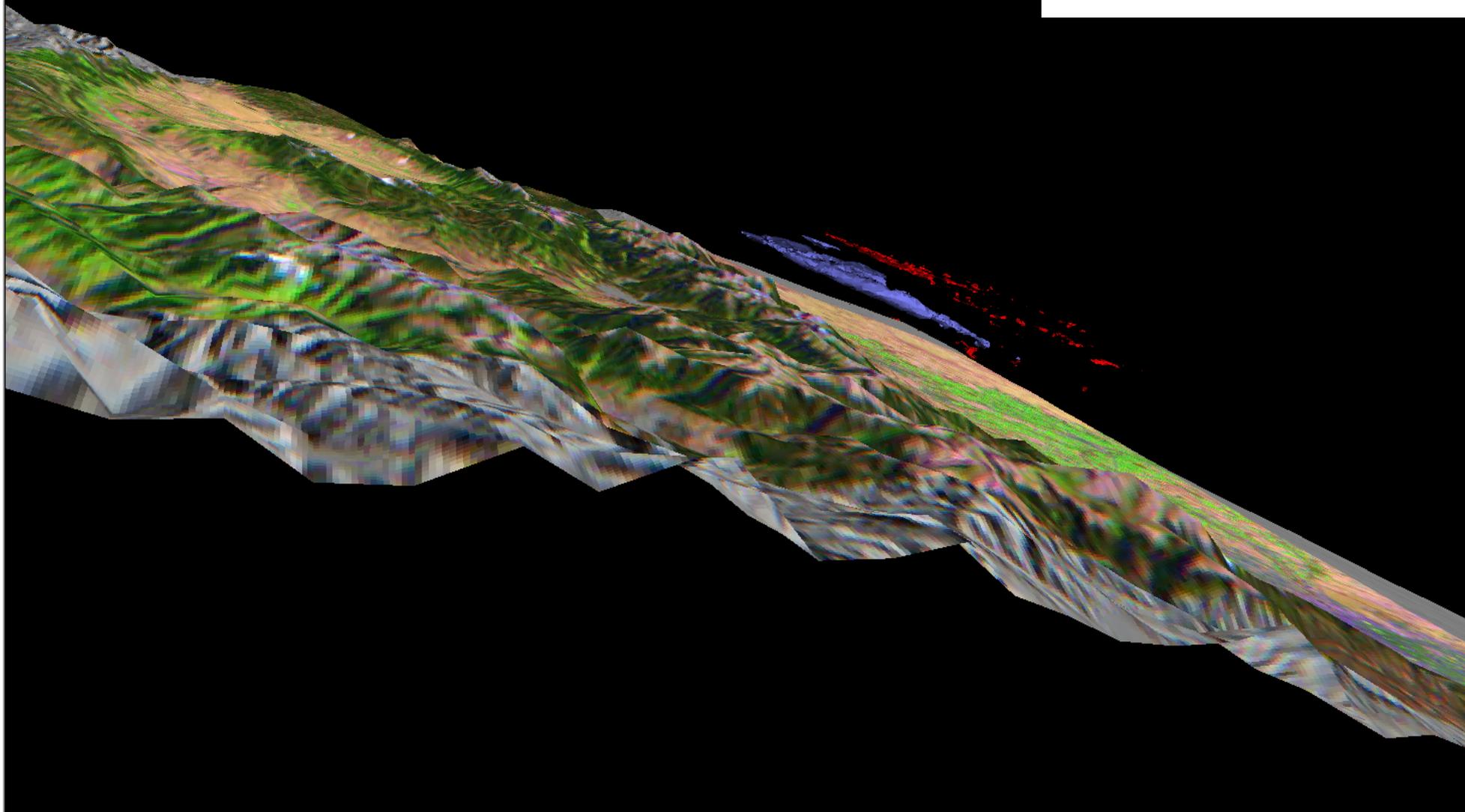
**TerraVision-2:  
VRML based data fusion  
and browsing.  
(MAGIC consortium,  
NCAR, and NAVO:  
[http://www.ai.sri.com/  
VRMLSets/Composite/](http://www.ai.sri.com/VRMLSets/Composite/))**

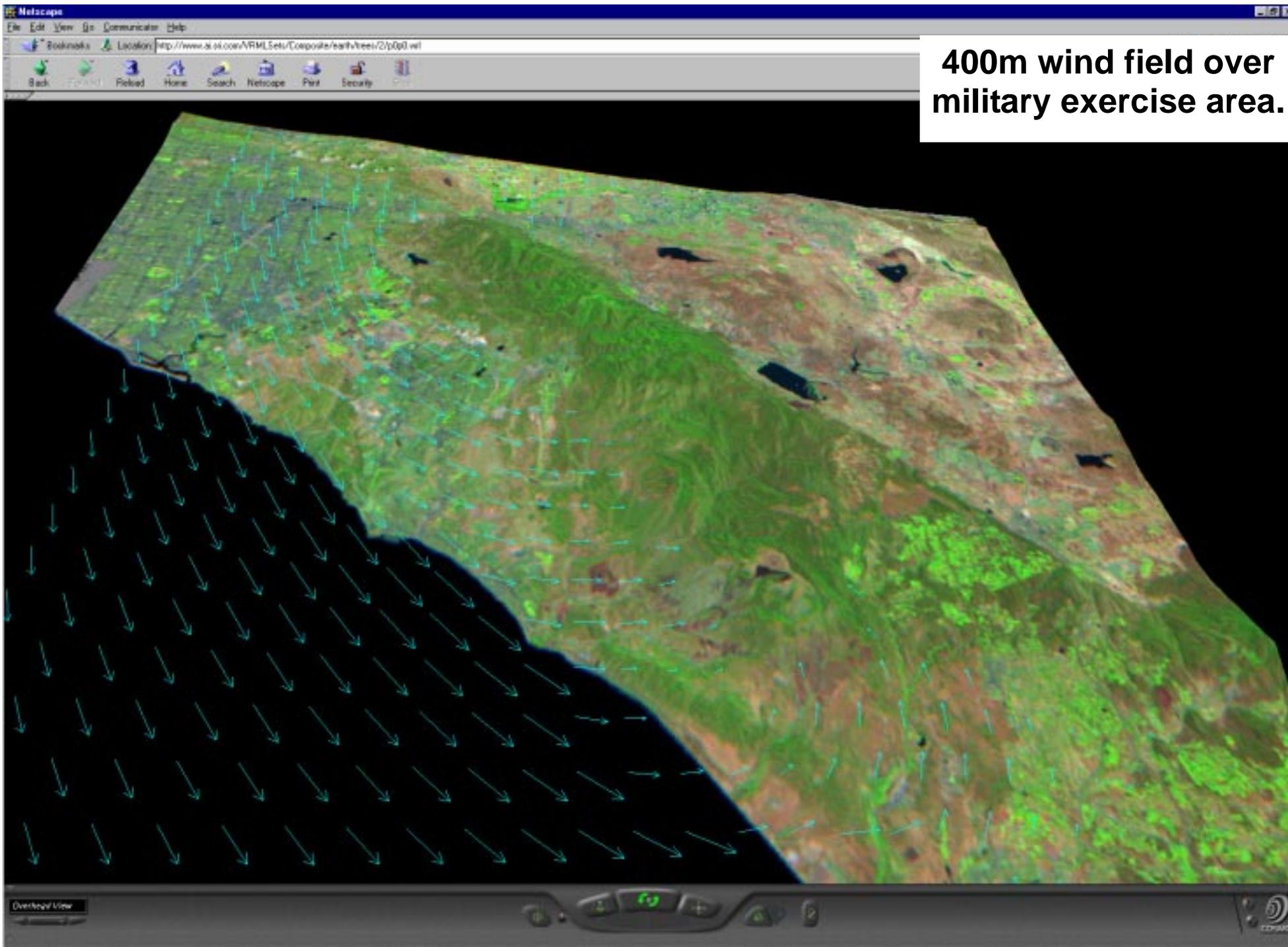


**clear air turbulence  
(front range of  
Colorado Rockies)**



**clear air turbulence  
(front range of  
Colorado Rockies)**





**400m wind field over military exercise area.**

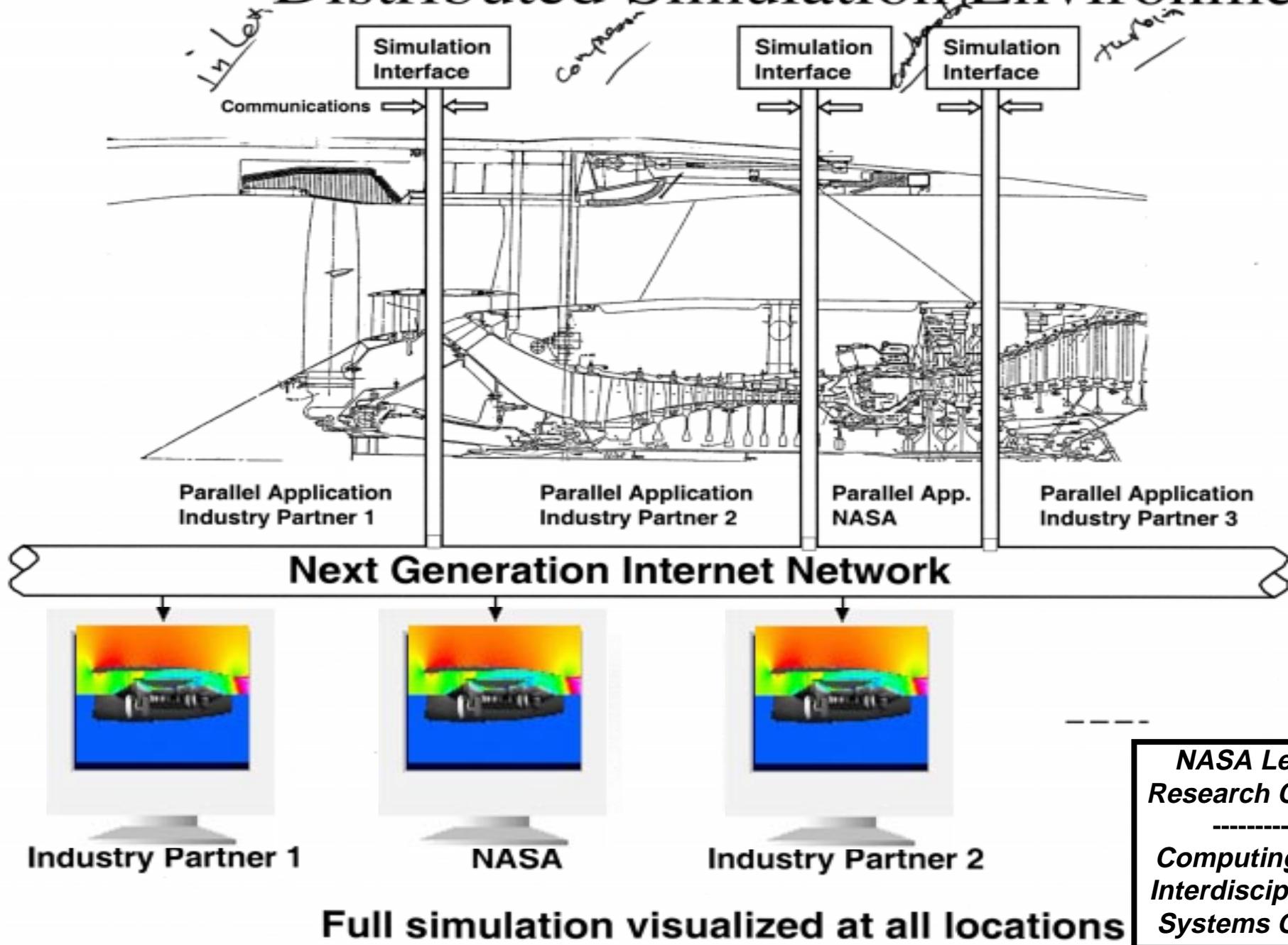
# **Distributed Simulation**

**Some problems that are too large for a single computing platform that can be distributed “naturally” in a grid environment:**

- ◆ **Not true for single address space / tightly coupled algorithms (What most people call “teraflop computing” will happen on the grid when teraflop computers are connected to the grid.)**
- ◆ **True for inherently parallel, loosely coupled problems (event analysis, parameter studies, etc.)**
- ◆ **Sometimes true for coupled simulations**

**Large-scale, “independently” computing simulation components with “reasonable” communications needs can be combined to provide “whole” system simulations by using the services of the grid.**

# Distributed Simulation Environment



# **Data Intensive Computing Characteristics**

- ◆ **Ingestion**
  - **manage instrument data streams and/or simulation output**
- ◆ **Cataloguing**
  - **describe data (generate metadata)**
  - **describe data formats**
  - **assign use conditions**
  - **publish**
- ◆ **Multi-Archive management**
- ◆ **Access**
  - **dataset “access protocols” (e.g. http, ftp, nfs, ...)**
  - **uniform I/O mechanisms (e.g. read, write, seek for all access)**
  - **discover data syntax (structure)**
- ◆ **Analysis / transformation (computation)**
- ◆ **Transience**
  - **many grid application systems (resource configurations) will be built on-demand and used for limited periods**

# **Data Intensive Computing: Required Services**

- ◆ **Location management**
  - local, remote, cache (where?)
- ◆ **Fault tolerance**
- ◆ **Access control**
- ◆ **Resource brokering**
- ◆ **Autonomous management**

## **Vision for the Grid**

**Uniform, location independent, and transient access to the**

- computational**
- catalogued data**
- instrument system**
- human collaborator**

**resources of science and engineering in order to facilitate the solution of large-scale, complex, multi-institutional / multi-disciplinary data and computational based problems.**

**These resources should be accessible through problem solving environments appropriate to the target user community.**

## **IPG Goals**

**Independent, but consistent, collections of tools and services that support applications and interactions at all levels of abstraction.**

**These tools and services installed and operated so as to integrate NASA computational, storage, and instrument facilities across multiple sites in order to provide an infrastructure capable of routinely addressing larger scale, more diverse, more transient problems than is possible today.**

**Initial IPG will integrate computing and data storage resources at ARC, LeRC, LaRC, and ICASE into a single grid system.**

# **Grid Architecture**

**The grid may be envisioned as a layered set of basic services and middleware that supports different styles of usage (e.g. different programming paradigms).**

**However, the implementation is that of a continuum of hierarchically related, independent and interdependent services, each of which performs a specific function, and may rely on other grid services to accomplish its function.**

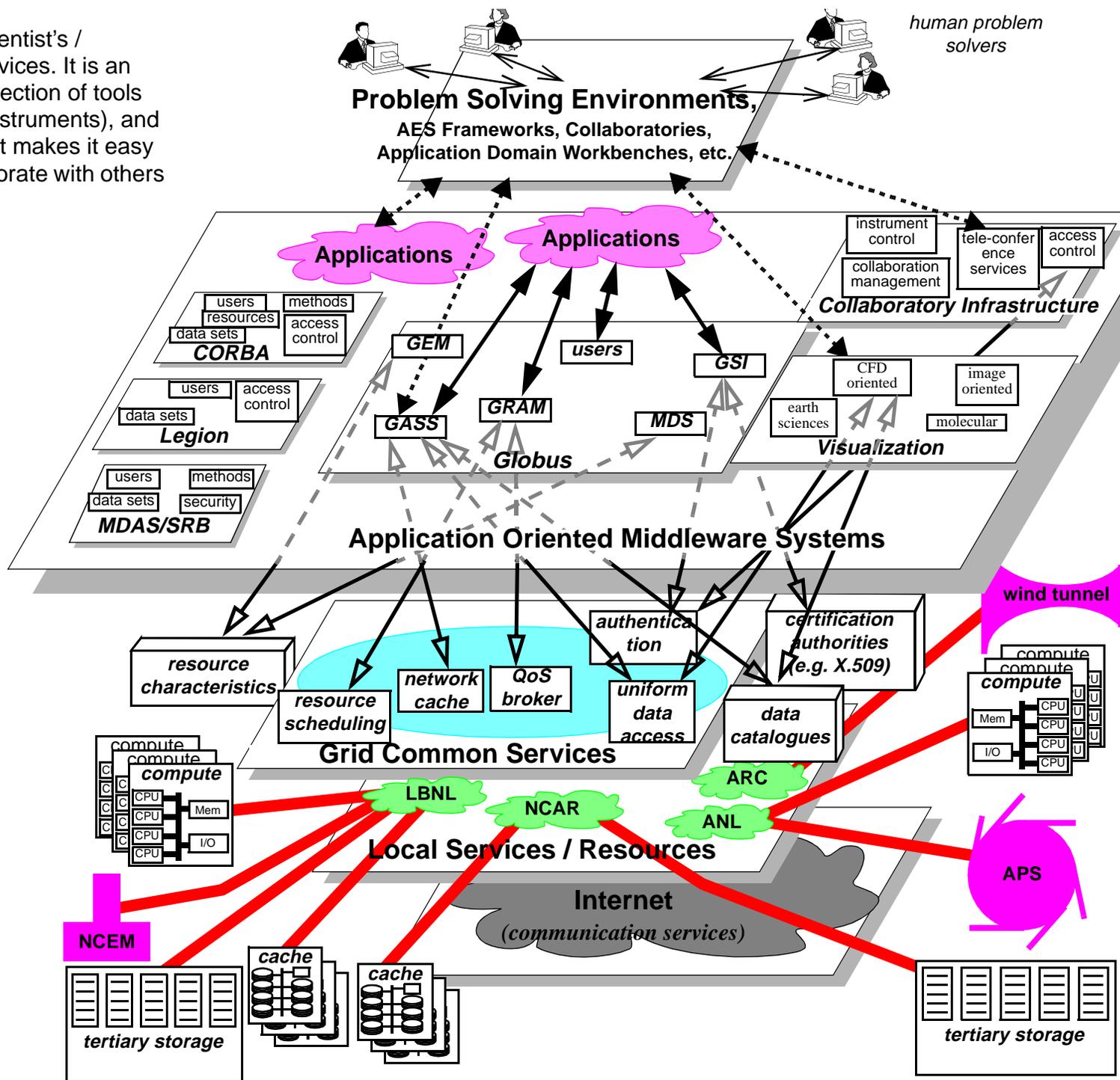
**Further, the “layered” model should not obscure the fact that these “layers” are not just APIs, but usually a whole collection of functions and autonomous management systems in order to provide the “service” at a given “layer.”**

The PSE layer provides the scientist's / engineer's interface to Grid services. It is an application domain-specific collection of tools (e.g. simulations, databases, instruments), and a "workbench" environment that makes it easy to use those tools and to collaborate with others working on the same problem.

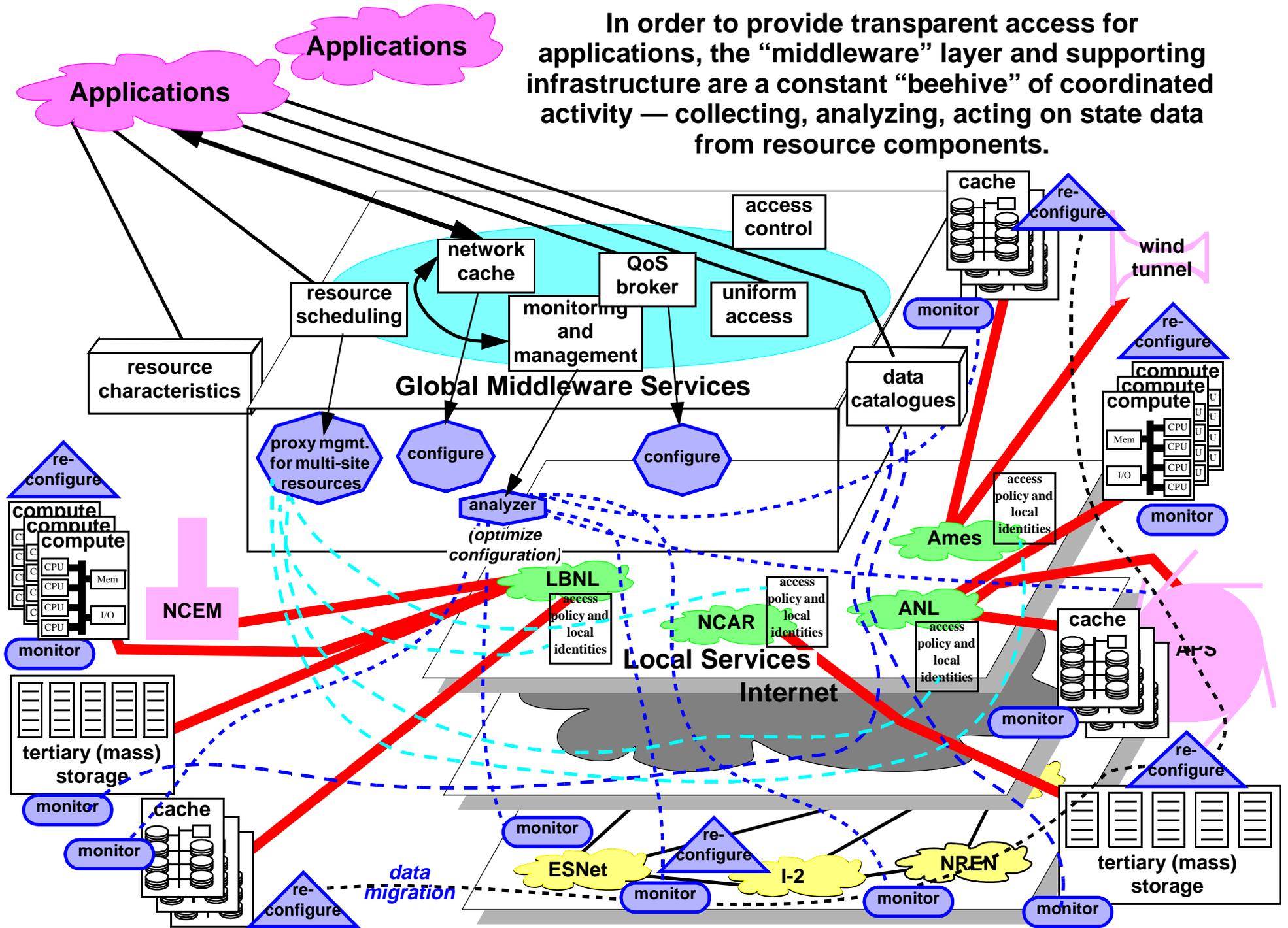
The middleware layer provides different styles of service interfaces for application developers to access the basic Grid services.

Grid services are "standard" interfaces for the functions needed to build and manage distributed applications of all sorts.

Most "resources" are "local" and will have their own resource managers and use policies. It is the use mechanisms and interfaces for the local resources that the Grid common services are intended to homogenize.



In order to provide transparent access for applications, the “middleware” layer and supporting infrastructure are a constant “beehive” of coordinated activity — collecting, analyzing, acting on state data from resource components.



# **Functionality and Approach**

**Major functionality goals for IPG include:**

- **Support for “easy” construction of application domain Problem Solving Environments**
- **Routine application domain use of large-scale, widely distributed computing systems**
- **Routine use and management of high rate data streams and massive data sets in widely distributed environments**
- **Integrated, computer mediated human collaboration**
- **Having the entire grid user environment appear like the IMac / Windows / X workstation on the user’s desk**

# **Technology Goals (all in wide area network environments)**

## **◆ Usability**

- **Toolkits for constructing application “frameworks” / Problem Solving Environments**
- **Easy management of resource use-conditions by stakeholders**

## **◆ Programability**

- **Support for multiple programming styles<sup>\*</sup> in multi-platform computing systems**
- **Uniform model for data access<sup>\*\*</sup>**
- **Generalized fault and data/process migration management in distributed systems**

---

<sup>\*</sup>Current examples include Fortran plus MPI (for message passing in parallel systems), C++ coupled to object oriented databases (for structured data storage and management), CORBA (being used to facilitate construction of composable systems - “reusable modules”), and general object oriented approaches such as Java/Jini and Legion are being experimented with.

<sup>\*\*</sup>What does a uniform access interface look like that incorporates MPI/IO files, legacy data sets in deep archives, Web based documents, and instrument data streams. (E.g. Vanderbilt’s work.)

## ◆ Scalability

- **Brokering<sup>\*</sup>, reservation, quality-of-service, and policy based access control for all resources (computing, data, instruments, and collaborators)**

## ◆ Operability

- **Tools and techniques for grid administration and operations (system administration, security, accounting, etc.)**

\*“Brokering” = given a description of the resources needed to solve a problem, locate candidates in the grid and negotiate for their use.

# **Implementing IPG**

**We are not starting from scratch: There is a sizable academic/government/commercial R&D computer science community working toward grid computing environments. *However Ames is making unique and substantial contributions:***

**First, through work in related technology areas. For example:**

- **on-line instruments (remote Wind Tunnel access - Darwin)**
- **NREN projects (OhioView, telepresence, network quality-of-service R&D, distributed applications)**
- **PBS and Metacenters (shared batch queuing system)**
- **Highly distributed storage systems**
- **Data mining and visualization**
- **Uniform data access R&D**
- **Numerical algorithms for widely distributed systems**

**Second, Ames is leading the effort to build the first prototype-production grid system that will provide distributed computation and tertiary storage access in wide area networks: Information Power Grid.**

**IPG is a sizable collaboration lead by NAS/ARC, and directly involving:**

- **LeRC and LaRC, DoD Major Shared Resource Centers**
- **both NSF high performance computing consortia: Alliance (NCSA) and NPACI (SDSC)**
- **additional universities**
- **industrial partners (e.g. Cisco, SGI, Sun)**

**Third, there is an implementation plan that includes the following elements:**

## **Two year goals:**

- **An operational IPG “prototype-production,” heterogeneous, distributed environment that provides access to computing, data, and instrument resources at several NASA Centers**
- **At least two significant Aerospace Engineering Systems applications operating in IPG**
- **A prototype Aerospace Engineering Systems collaborative workbench**
- ***Support for real-time access to scientific and engineering instrumentation systems***

# **IPG Two Year Grid System Objectives**

**1) Grid runtime environment I: A robust, usable, widely deployed, Globus-based, distributed computing framework, providing:**

- **resource management**

(standardized interfaces to various local resource management systems (GRAM) and that manage allocation of collections of resources (DUROC))

- **remote access**

(transparent remote access to files (GASS and RIO))

- **execution environment management**

(executable code and library staging (GEM))

- **security**

(single sign-on, authentication, authorization, and privacy within the Globus system)

- **fault detection**

(services supporting fault detection and recovery into Globus applications (HBM))

- **information infrastructure**  
(Global access to information about the state and configuration of system components of an application (MDS))
- **Grid programming services**  
(support for writing parallel-distributed programs, monitoring, etc. (MPICH-G))
- **Grid administration**  
(tools for the Grid operators)

## 2) Grid runtime environment II: Augmented services

- **a global shell**  
(various task models, workflow mgmt., signals, I/O mgmt., etc.)
- **facilities for resource time based scheduling, matching, QoS**  
(support for co-scheduling, relative priority, “Matchmaker,” etc., for all resources)
- **policy based access control**
- **application performance monitoring**
- **distributed debuggers**
- **support for Java, CORBA, and DCOM**
- **support for commodity platforms**  
(NT cycle stealing and Linux and NT clusters)

### **3) Grid advanced application support: Auxiliary services integrated into the Globus Grid framework:**

- **high-performance, high-capacity, metadata based digital data repositories**
- **collaborative and data exploration tools**
- **autonomous agent framework providing support functions for both humans and resources**
- **autonomous infrastructure management**
- **knowledge based, generalized fault management services and tools**
- **Identification and implementation of the services needed to support collaboration tools and PSE/workbench systems, and composable workbench components**

### **4) Integration of instrument systems**

# **Two Year Operational Objectives**

- 1) Persistent, “large-scale,” heterogeneous, distributed testbed that enables applications that cannot be done today**
  - significant CPU resources**  
**( $\geq 50\%$  of all non-COSMO, NAS resources + IPG systems at LeRC, LaRC, and ICASE ( $\cong 30$  Gflop) + resources from Alliance and NPACI)**
  - stable and predictable application environment**  
**(long term resource usage policies + operational support)**
  - widely distributed**  
**(usable resources at multiple sites)**
  - running “stable” Globus “beta” release**  
**(new services will be incorporated into the testbed, which will have the direct involvement of the Globus developers)**

- **“bullet proofing” and validating the beta releases prior to installation in the testbed**  
**(developers testbed: clean up, validate, and integrate into the beta distribution)**
- **user documentation and consulting**  
**(users will both be supported and be partners in the IPG evolution)**
- **operations and system admin. services and documentation**  
**(testbed will also be for training systems and operations people)**

## **2) Comprehensive benchmark suite**

- **develop performance measurement tools for the Grid**
- **“regression” suite for monitoring stability of Grid functionality**
- **provide programming examples**

### **3) AES proto-application**

- **provide support for AES applications (airframe and propulsion)**
- **begin development of an AES framework and “workbench”**

### **4) Operating in COSMO environment**

- **establish COSMO criteria and mechanism for transition to production**

### **5) An operating “standards” process with “community” and industry involvement**

- **establish computing industry involvement**
- **establish aerospace industry involvement**
- **facilitate community “standards” process**

## **6) Advanced heterogeneous computing environment**

- maintain IPG as a viable testbed for prototyping advanced distributed systems and prototype applications**
- use IPG testbed to evaluate and promote advanced computing platforms**

# **Two Year R&D Objectives**

- 1) Distributed-parallel algorithms in the grid environment**
- 2) Ultra high-speed distributed systems**
  - Network R&D**
  - Infrastructure design, monitoring, and testing**
  - Platform and I/O subsystems**
  - Data management systems**
  - Applications architectures and algorithms**
  - High data rate instrument systems**
  - Proto-applications**
- 3) Advanced program execution environment**
  - Network and resource-aware adaptation**

# **IPG Four Year Objectives**

- 1) A high performance, widely distributed, global file system**
  - provide a completely uniform view of the Grid application environment regardless of location**
  
- 2) Transparent fault tolerance and reconfiguration**
  - the Grid will dynamically reconfigure itself and provide sufficient assistance and information to running applications that they can also reconfigure**
  
- 3) Numerical techniques optimized for parallel-distributed environments**
  - significant progress in developing new numerical algorithms optimized for the Grid environment**
  
- 4) Infrastructure security**
  - operational security for the communications, systems, and other resources of the Grid**

- 5) Distributed object programming paradigm (probably Java based) integrated into the Grid framework**
- 6) Facilities for coupled computational simulation and experiments, and computational steering**
- 7) High capacity, high performance, prototype-production metadata based data repositories that support data management, mining, and exploration for AES and DAO**
- 8) Prototype ASE Framework components**
  - collaborative workbench building environments that enable coordinated, multiple component model operation**

# **IPG Six Year Objectives**

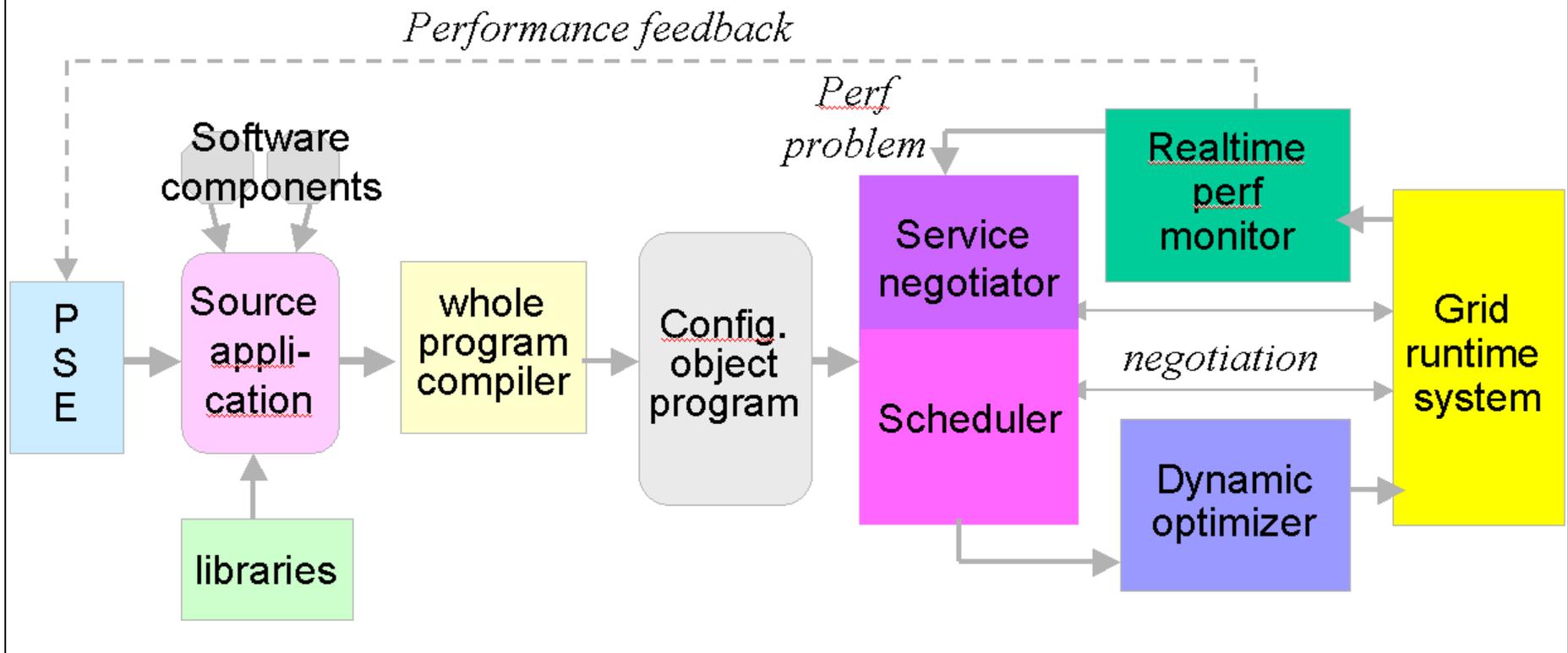
- 1) Advanced programming paradigms including adaptive compilation and resource driven dynamic linking**
- 2) Application performance model based resource scheduling**

**These objectives are closely aligned with, and will be coordinated with NSF's Next Generation Software program:**

**“The NGS program fosters multidisciplinary software research under two components: Technology for Performance Engineered Systems (TPES), and Complex Application Design and Support Systems (CADSS). The overall thrust of NGS will be research and development for new software technologies integrated across the systems' architectural layers, and supporting the design and the operation cycle of applications, computing and communications systems, and delivering quality of service (QoS). The TPES component will support research for methods and tools leading to the development of performance frameworks for modeling, measurement, analysis, evaluation and prediction of performance of complex computing and communications systems, and of the applications executing on such systems. The CADSS component will support research on novel software for the development and run-time support of complex applications executing on complex computing platforms; CADSS fostered technology breaks down traditional barriers in existing software components in the application development, support and runtime layers, and will leverage TPES**

# • Grid-aware Programming

- development of adaptive poly-applications
- integration of schedulers, PSEs and other tools



(Berman, Darema, Gannon, Kennedy, et al.)

developed technology for delivering QoS.

See <http://www.nsf.gov/cgi-bin/getpub?nsf998>

### **3) Prototype ASE Framework**

- **coordinated operation of multiple AES component models in a framework that represents and can (partially) evaluate an operational vehicle**

***IPG is putting significant resources into both R&D, and testbed construction and operation, in about equal proportion.***

***R&D will be an on-going IPG activity in order to ensure that the grid continuously evolves through incorporating leading-edge technology.***

# **IPG Next Generation Internet Projects**

***Every one of the previous example applications is an integral part of a high-speed network testbed:***

**High-speed, wide area networks are an essential and inseparable aspect of grids.**

- ◆ **IPG will enable the applications envisioned for NGI by providing the distributed systems infrastructure.**
- ◆ **NGI will enable IPG through R&D and testbeds for several key network communications technologies.**

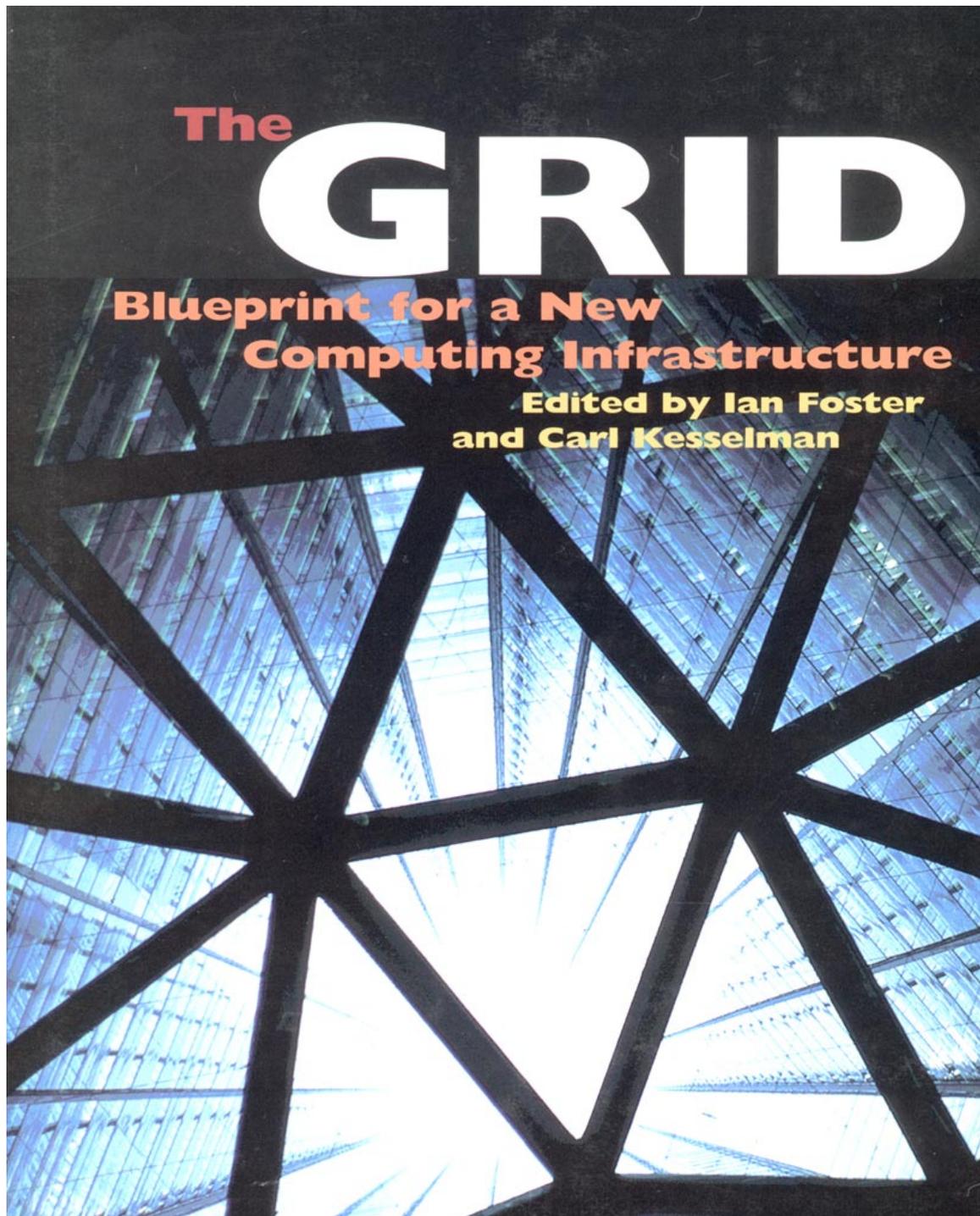
## **IPG NGI projects:**

- ◆ **Very high data-rate applications that will use IPG as infrastructure, which, in turn, uses NGI for network capacity:**
  - **OC-48 (2.5 Gb/s) JPL ↔ ARC ↔ Boeing, Seattle**
    - aviation safety “help desk” (*next talk*)
    - remote operation of wind tunnels (*next talk*)
    - high data-rate access to historical wind tunnel data
    - image processing and SAR processing with JPL
    - remote VR
- ◆ **NGI technology R&D supporting IPG infrastructure**
  - **Network quality-of-service and reservable bandwidth**
  - **High data-rate and high volume wide area data transport**

# **Areas of New R&D for Data Intensive Applications**

- **Services for locating the data resources needed to solve a given problem**
- **Integration of tertiary storage systems with digital libraries / metadata catalogue systems in widely distributed environments**
- **Widely distributed storage architectures that incorporate resource and bandwidth QoS**
- **Integration of policy based access control**
- **Integration of data location management with uniform data access techniques**
- **Fault tolerant distributed storage and catalogue systems**

**Development of automated cataloguing techniques and self-describing data semantics from metadata**



*The Grid is a consistent, open, and standardized environment providing collaborative problem solving that involves using distributed, aggregated, and high performance, computing and large data archive resources, and real-time, high data rate instrumentation, and computer mediated, human collaboration.*

# **References and Notes**

- [1] *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8. [http://www.mkp.com/books\\_catalog/1-55860-475-8.asp](http://www.mkp.com/books_catalog/1-55860-475-8.asp)
- [2] Globus is a R&D grid system that is the starting point for IPG. Globus is described at [www.globus.org](http://www.globus.org)
- [3] “Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments,” W. Johnston, Jin G., C. Larsen, J. Lee, G. Hoo, M. Thompson, and B. Tierney (LBNL) and J. Terdiman (Kaiser Permanente Division of Research). Invited paper, International Journal of Digital Libraries - Special Issue on “Digital Libraries in Medicine”. May, 1998. <http://www-itg.lbl.gov/WALDO/>
- [4] DPSS: The Distributed-Parallel Storage System (DPSS) is a scalable, high-performance, distributed-parallel data storage system developed in the MAGIC Testbed. The DPSS is a collection of wide area distributed disk servers which operate in parallel to provide logical block level access to large data sets. Operated primarily as a network-based cache, the architecture supports cooperation among independently owned resources to provide fast, large-scale, on-demand storage to support data handling, simulation, and computation in a high-speed wide-area network-based internetworked environment. See <http://www-didc.lbl.gov/DPSS> .
- [5] Clipper: The goal of the Clipper project is software systems and testbed environments that result in a collection of independent but architecturally consistent service components that will enhance the ability of applications and systems to construct and use widely distributed, high-performance data and computing infrastructure. Such middleware should support

high-speed access and integrated views for multiple data archives; resource discovery and automated brokering; comprehensive real-time monitoring and performance trend analysis of the networked subsystems, including the storage, computing, and middleware components, and; flexible and distributed management of access control and policy enforcement for multi-administrative domain resources. See <http://www-itg.lbl.gov/~johnston/Clipper>

[6] MAGIC: “The MAGIC Gigabit Network.” See: <http://www.magic.net>

**<http://nas.nasa.gov/~wej/home/IPG>**