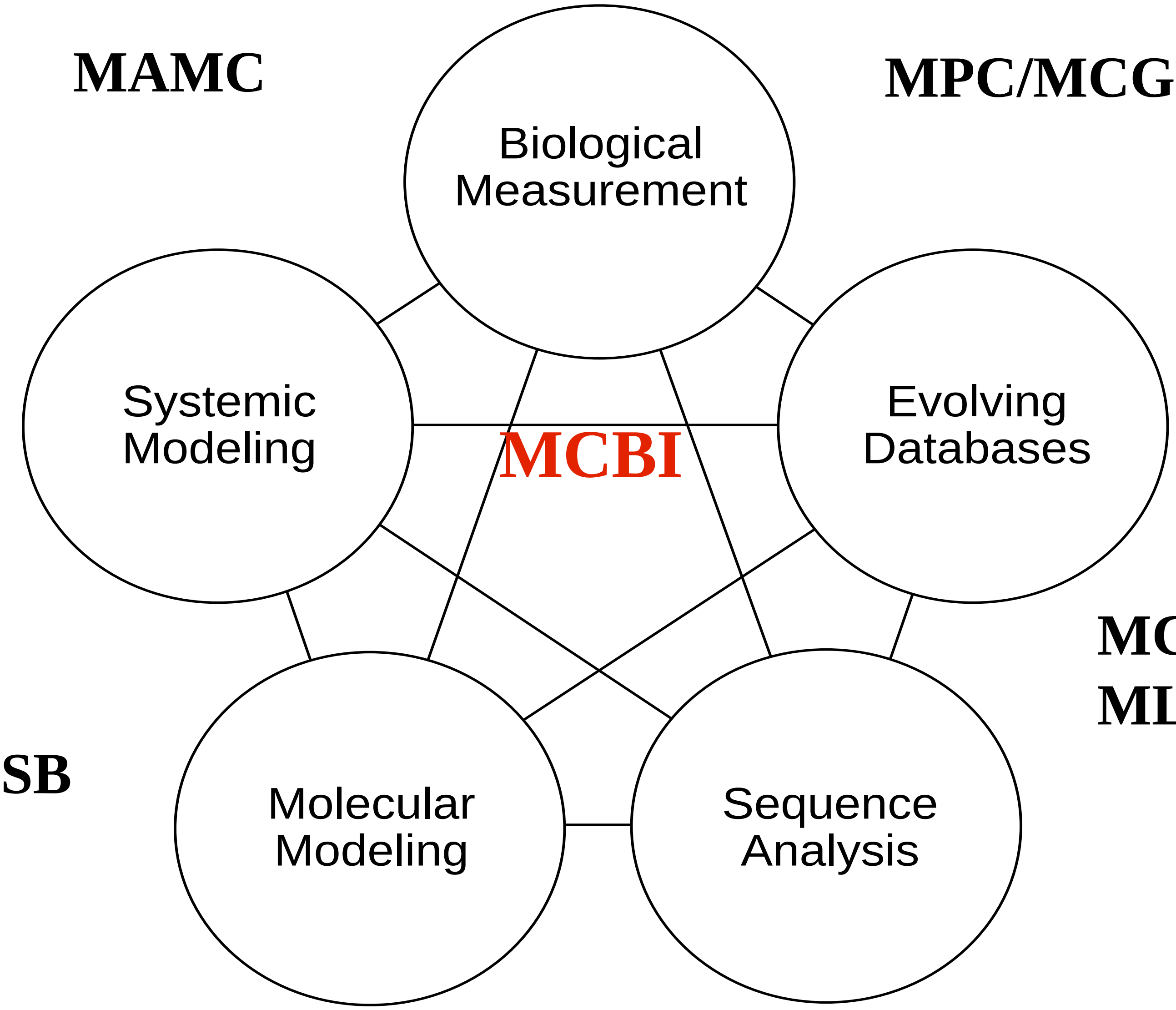
Principles and Experience that Can Help Bring the Michigan Center for Biological Information to Functionality and Sustainability

Brian D. Athey, Ph.D.

Assistant Professor
Director University of Michigan Visible Human Project
Department of Cell and Developmental Biology
University of Michigan Medical School
Interim Director MCBI
Office of VP Research University of Michigan
Ann Arbor, Michigan

July 23, 2001

MLSC Core Activities Map to the University of Michigan Bioinformatics Vision and the Central Role of the MCBI



MAMC

Biological Measurement

MPC/MCGT

Systemic Modeling

MCBI

Evolving Databases

- **CTA**
- **Private**
- **Public**

MCGT
MLSC Users

MCSB

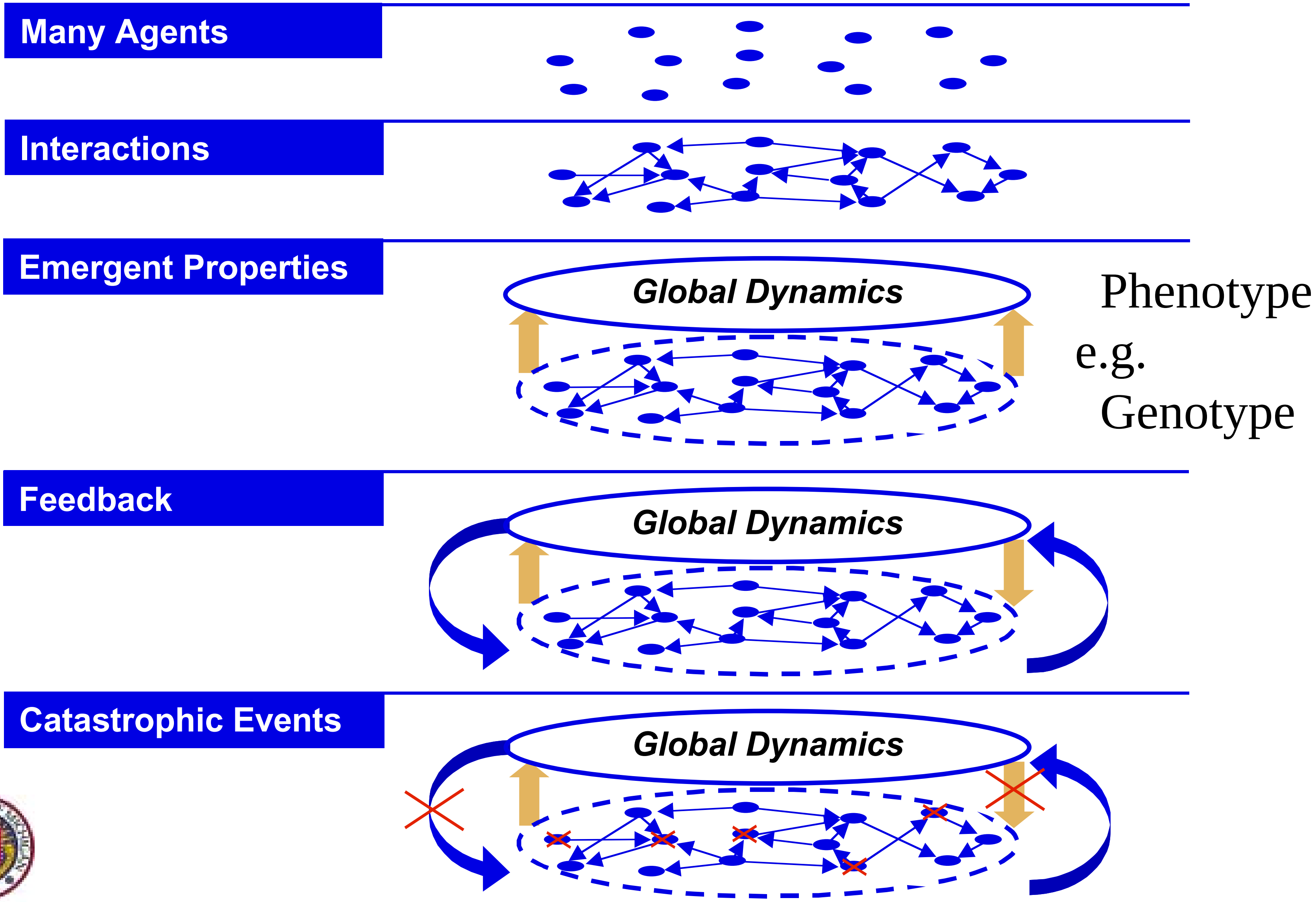
Molecular Modeling

Sequence Analysis

MCBI Core Priorities

- **Applications Layer:**
  - Bioinformatics tools and support
  - Celera and proprietary database access
  - Gene-Array Test-bed/Rosetta/GeneSpring/SpotFire
- **Middleware layer:**
  - Bioinformatics-specific integration middleware
    - Working ontological framework
    - Heterogeneous database architecture/LIMS integration/metadata scheme
  - Security vertically and horizontal integrated
- **IT Physical Layer and Networks:**
  - Central and distributed machine room(s) and network functionality
  - Core IT and Sensor inventory baseline and architecture
  - Communications (e.g interactive www and teleconferencing)

Complex Adaptive Systems (CAS): A Key Motivator of the University of Michigan Life Sciences Institute (LSI)



Many Agents

Interactions

Emergent Properties

Global Dynamics

Phenotype
e.g.
Genotype

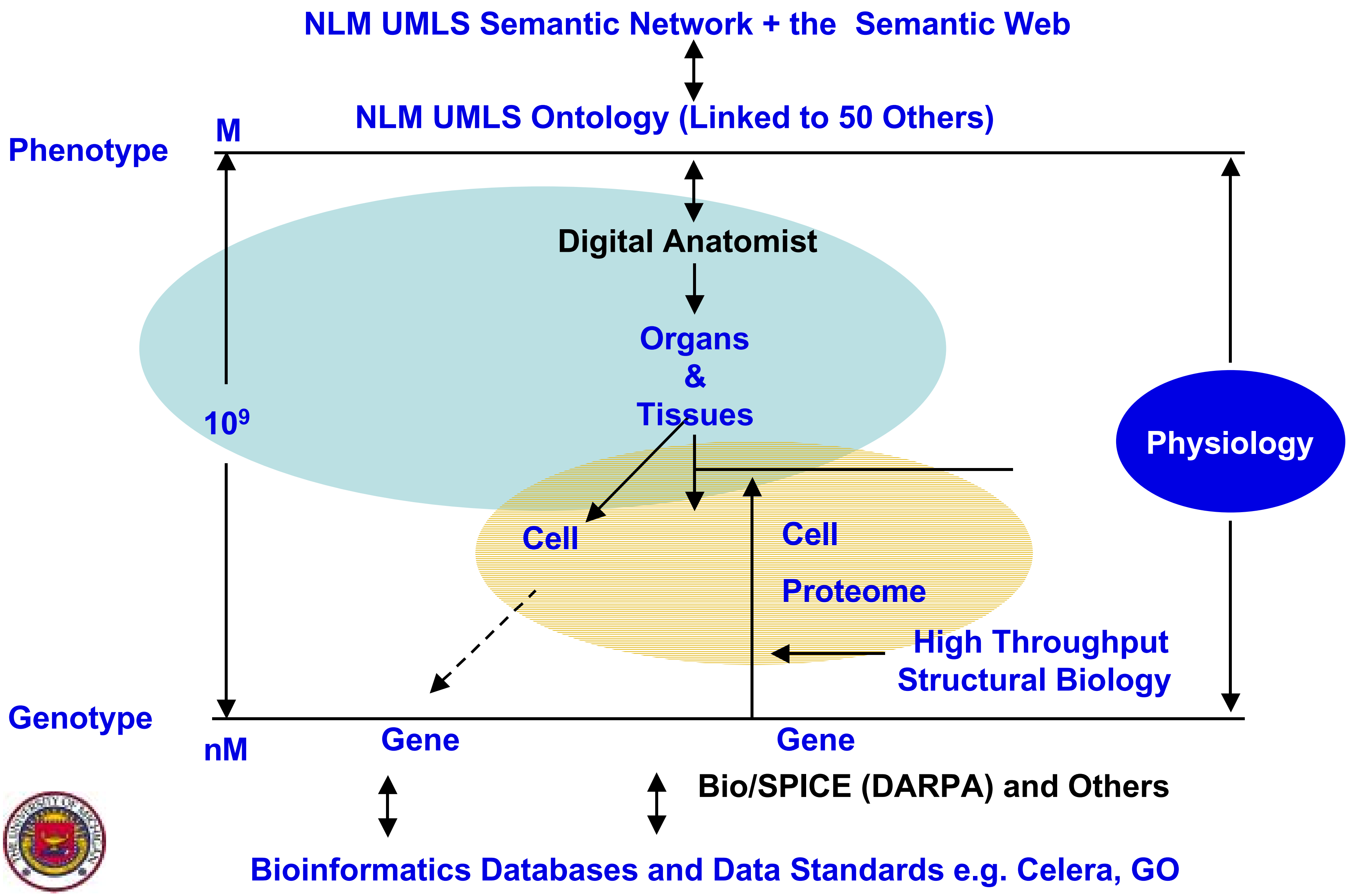
Feedback

Global Dynamics

Catastrophic Events

Global Dynamics

Ontology Integration and Biological Context



NLM UMLS Semantic Network + the Semantic Web

NLM UMLS Ontology (Linked to 50 Others)

Phenotype

M

10^9

Genotype

nM

Digital Anatomist

Organs
&
Tissues

Cell

Cell

Proteome

High Throughput
Structural Biology

Physiology

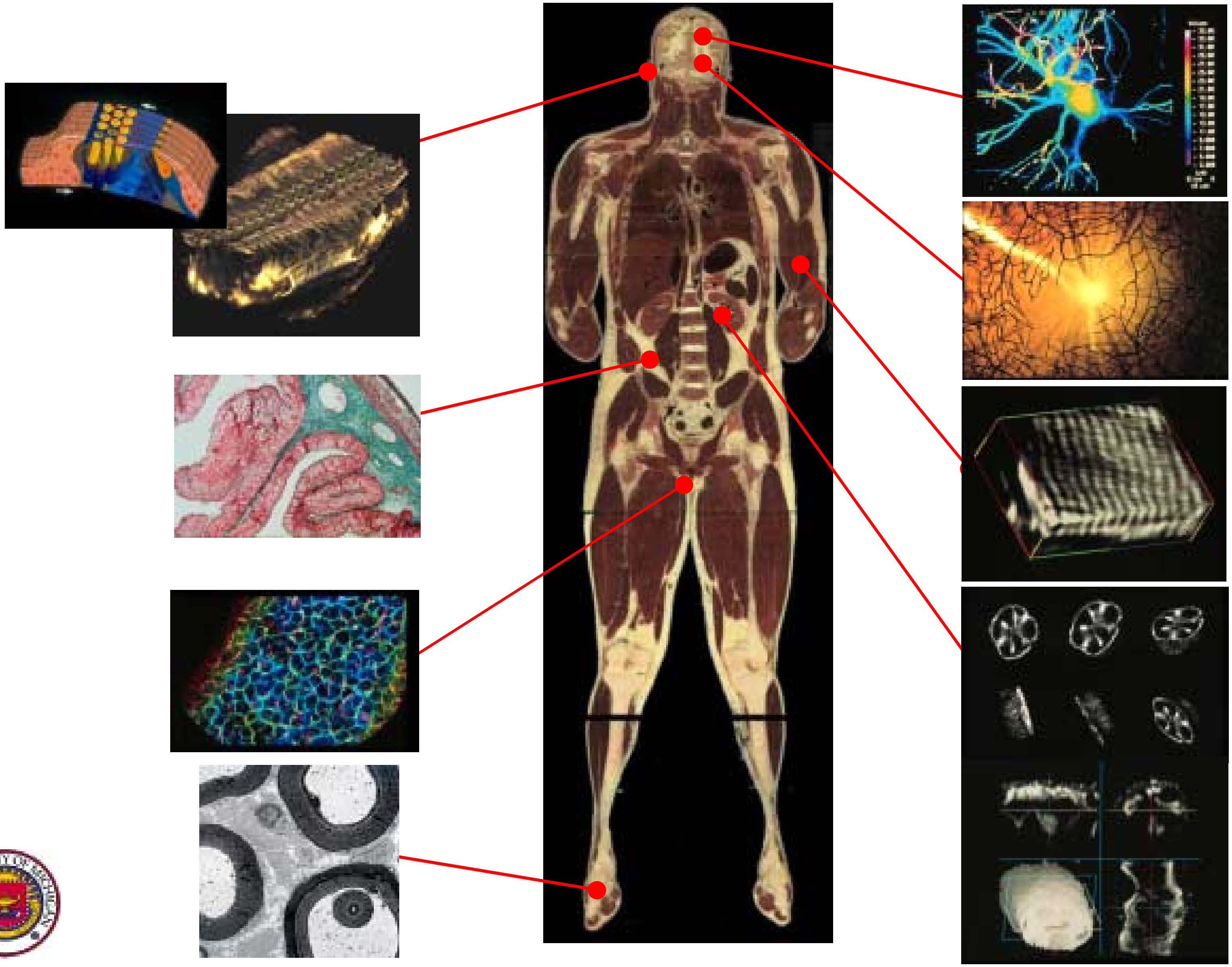
Gene

Gene

Bio/SPICE (DARPA) and Others

Bioinformatics Databases and Data Standards e.g. Celera, GO

Hierarchical Augmentation of the Visible Human Data Using Modern Microscopic Imaging



UM Visible Human Ontology Database and Server

Content
The database architecture allows for pluggable lexicons. Currently, the Terminologia Anatomica and Latin lexicons are supported, but other lexicons may be added. Customized learning modules are also supported. An instructor may use authoring software to create or edit learning modules of anatomical terms and relationships. Modules may contain any combinations of lexical terms.

Data Modeling
Supported relationships include structural, connection, clinical, and boundaries. Connection and Clinical relationships are one-to-many or one-to-one. Structural relationships are modeled as DAGs, or Directed Acyclic Graphs. That is, nodes share a boundary as a parent and child, and there is no looping.

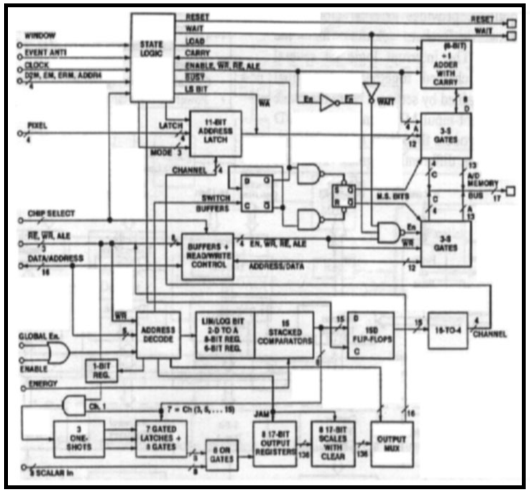
Extensions
The database can support multi-resolution data by extending the leaves of the DAG to include histological and sub-cellular information from other databases. Connections to those databases can occur by inclusion of pointers in the leaves of the UMVH Ontology database.

Hardware and Software
The UMVH database server is a Sun E250 running Solaris 8. The server has 1GB of RAM, 400GB of disk space, and 2 SPARC processors. Oracle 8i is running with the interMedia and Spatial extensions. InterMedia is used to store binary data like movies and images, as well as indexed text. The Spatial package allows compact storage and querying of points, lines, and curves for geometry landmark mappings.

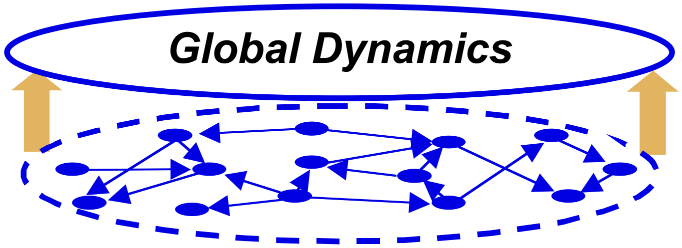
Alex Ade, July 2001

Bio/SPICE

Bio/SPICE is a biological data analysis, modeling workspace and database based loosely on SPICE tools used by Electrical Engineers for the design and analysis of their circuitry. It is a Consortium of 18 Contractors, lead by Lawrence Berkley Laboratory (LBL) and funded by DARPA for 5 years at $60M+. BioSPICE is taking on the Gene to Proteome Ontology problem, defining biological equivalents to electrical components (blue arrows).

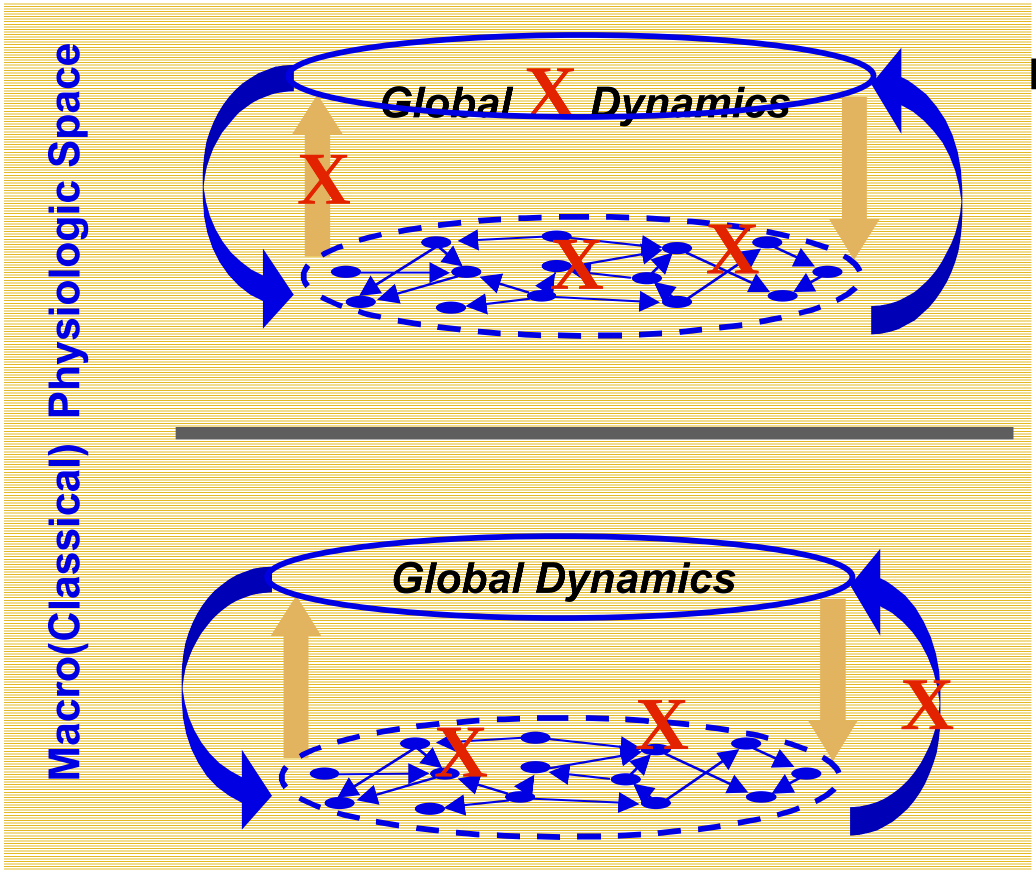


SPICE is an software environment to model and simulate electrical circuits



Bio/SPICE Analog to Circuit Diagram

Macro Physiological Ontology Example: Pathologic Phenotype



Lung Cancer

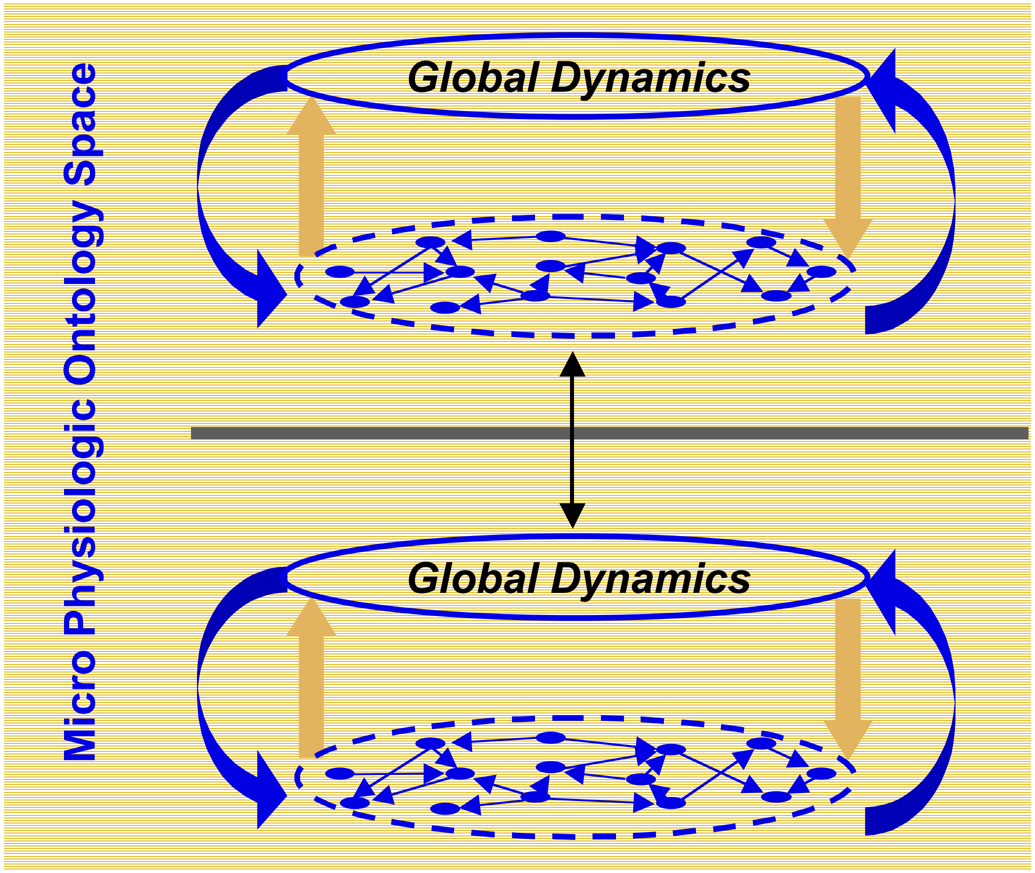
Abnormal Tissue
and System
Properties

Atypical Cellular
Adhesion and
Physiology

Tissue Specific
Gene Mutation Profile

Environmental Variable: Heavy Smoker

Micro Physiological Ontology Example: LSI Focus
Integrating MCBI, MPC, and MCGT



Signal Transduction Models and 3-D Cellular Distribution

Proteomics/Confocal Imaging

High Throughput Structural Biology

Gene Expression Profile

Biological Measurements and Modeling: Genes, SNPs, RFLPs, etc

Elements of High Throughput Structural Biology and Current MCSB Interests (Ranked Order)

Color Panels: Stevens and Wilson, Pipeline Flow from cDNA to Structure (Science 293, 519-520, 2001)

4.

Gene Database A_1 **Target Selection Pipeline** A_2 **Target Filter** A_3 **Target Prioritization** A_4

2.

Clone B_1 **Express** B_2 **Purify** B_3 **Crystallize** B_4

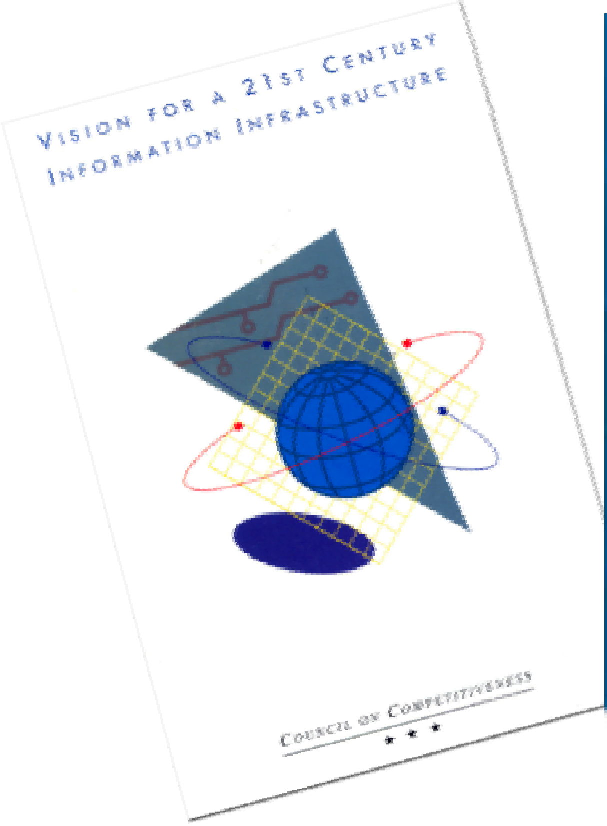
3.

Data C_1 **Solve Structure** C_2 **Refine** C_3 **Validate**

1.

1. Modeling<-------->Analysis<-------- >Simulation

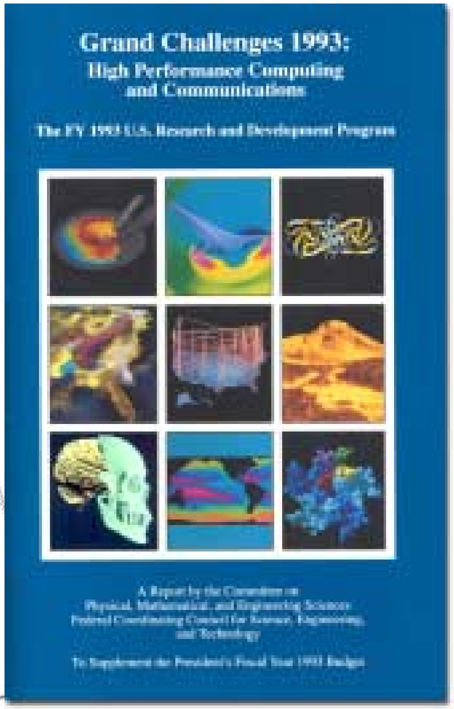
Evolution of Cutting Edge Network Connectivity for General Scientific and Life Sciences Applications



Vision for a 21st Century
Information Infrastructure

Council on Competitiveness

Late stage
NSFnet and
transition
1991-1993



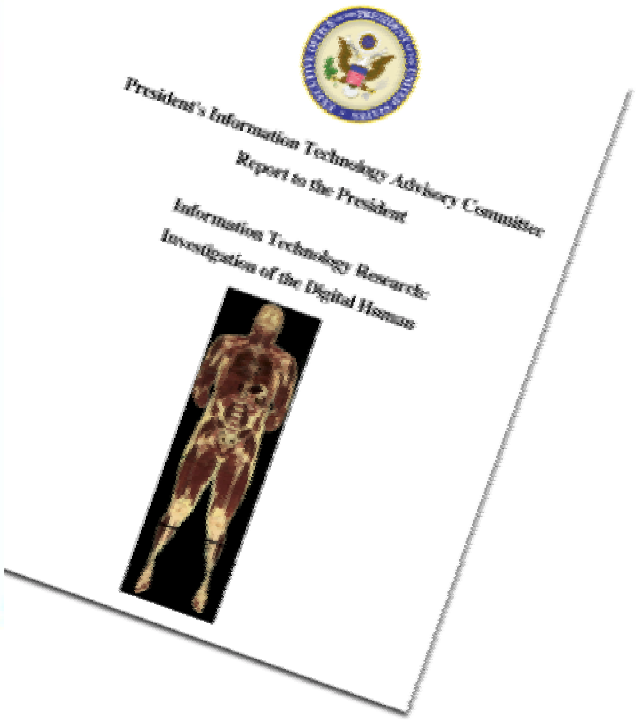
Grand Challenges 1993:
High Performance Computing
and Communications

The FY 1993 U.S. Research and Development Program

A Report by the Committee on
Physical, Mathematical, and Engineering Sciences
Federal Coordinating Council for Science, Engineering,
and Technology

To Supplement the President's Fiscal Year 1993 Budget

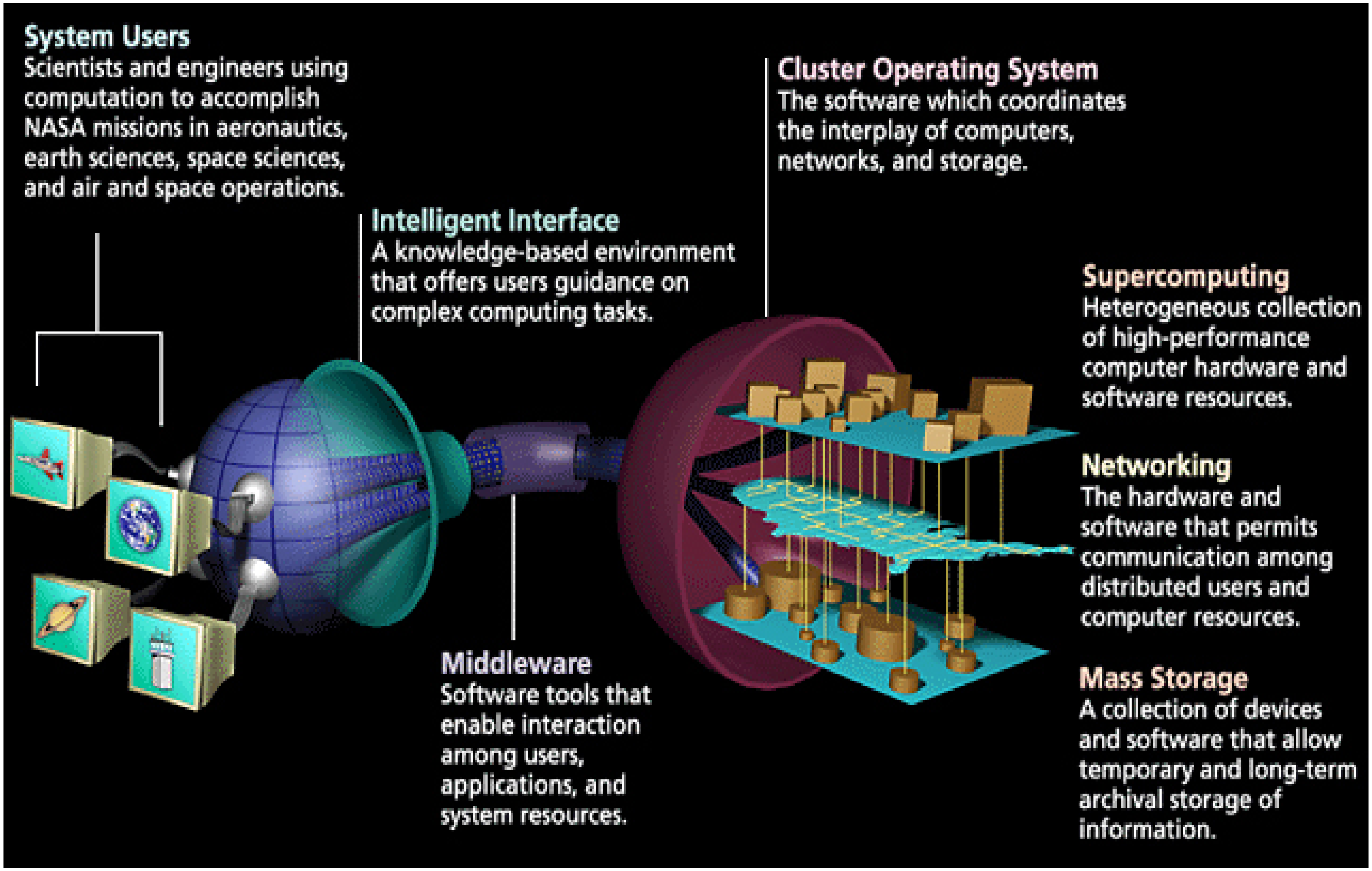
Internet HPCC
1993-1998



President's Information Technology Advisory Committee
Report to the President
Information Technology Research:
Investigation of the Digital Human

Next
Generation
Internet (NGI)
1998-2002

The Next Stage in Development is called a Highly Distributed Cyber Infrastructure or Access Grid



System Users
Scientists and engineers using computation to accomplish NASA missions in aeronautics, earth sciences, space sciences, and air and space operations.

Intelligent Interface
A knowledge-based environment that offers users guidance on complex computing tasks.

Middleware
Software tools that enable interaction among users, applications, and system resources.

Cluster Operating System
The software which coordinates the interplay of computers, networks, and storage.

Supercomputing
Heterogeneous collection of high-performance computer hardware and software resources.

Networking
The hardware and software that permits communication among distributed users and computer resources.

Mass Storage
A collection of devices and software that allow temporary and long-term archival storage of information.

Courtesy NASA

Semantic Web

"The Semantic Web is an extension of the current web

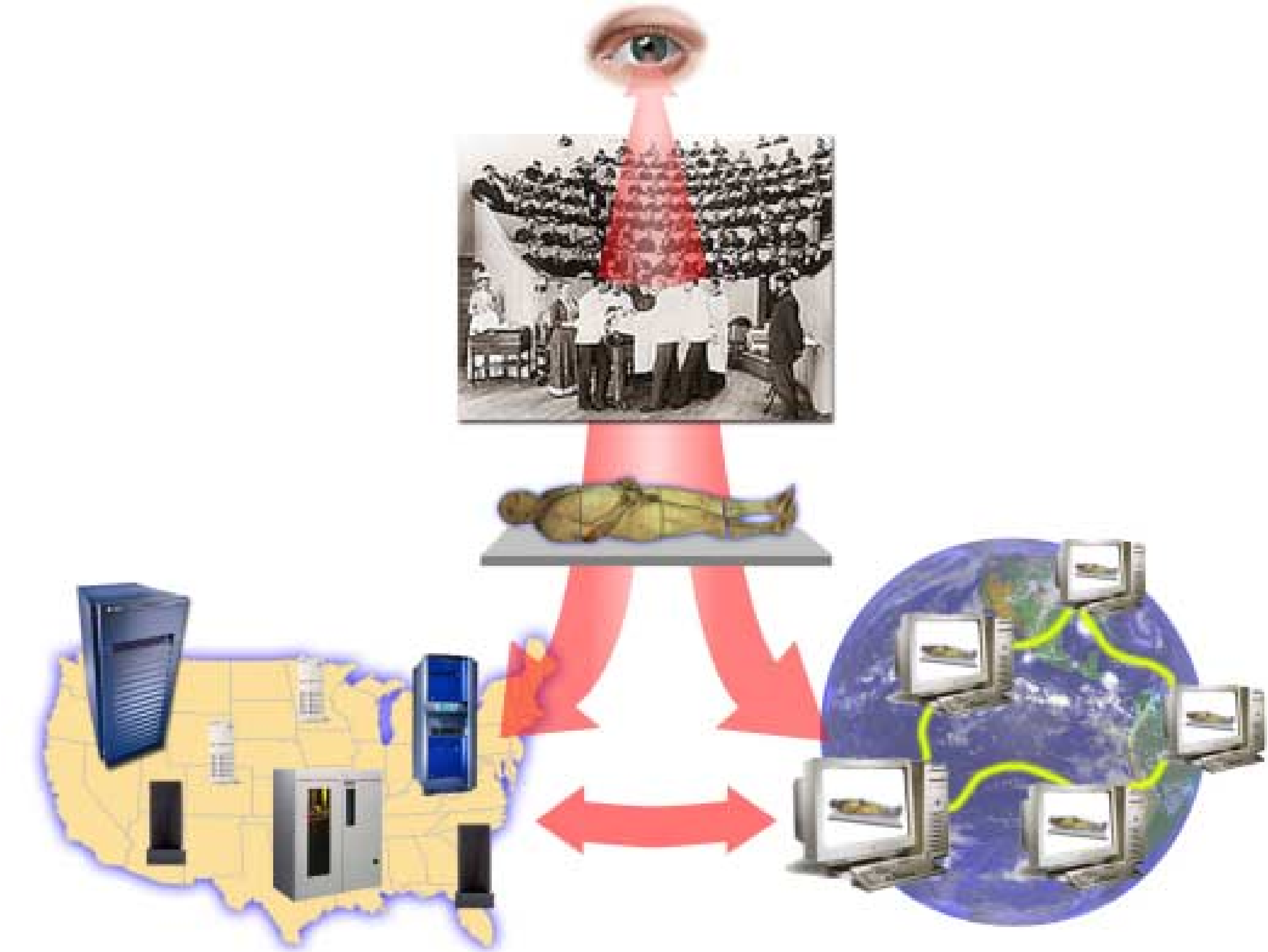
in which information is given well-defined meaning,

better enabling computers and people to work in cooperation."

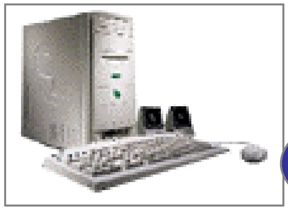
-- Tim Berners-Lee, James Hendler, Ora Lassila,

Illustrative Example: The NLM Visible Human Project

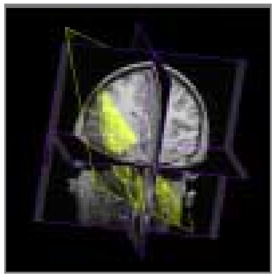
The Evolving Amphitheater of
Human Anatomy



Integration of VH, and Volume/Structure/ Instructional Databases



User Request



Arbitrary Section

Rotation-Translation Parameters

Navigable Volume Data

XML Tagged

Volume/Structure/Instructional Space

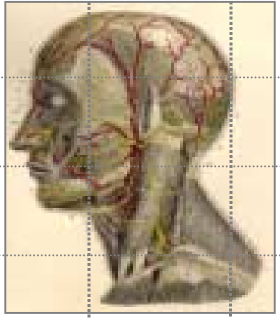
Area	Structure	Volume

UMLS MeSH TA

XML Tagged

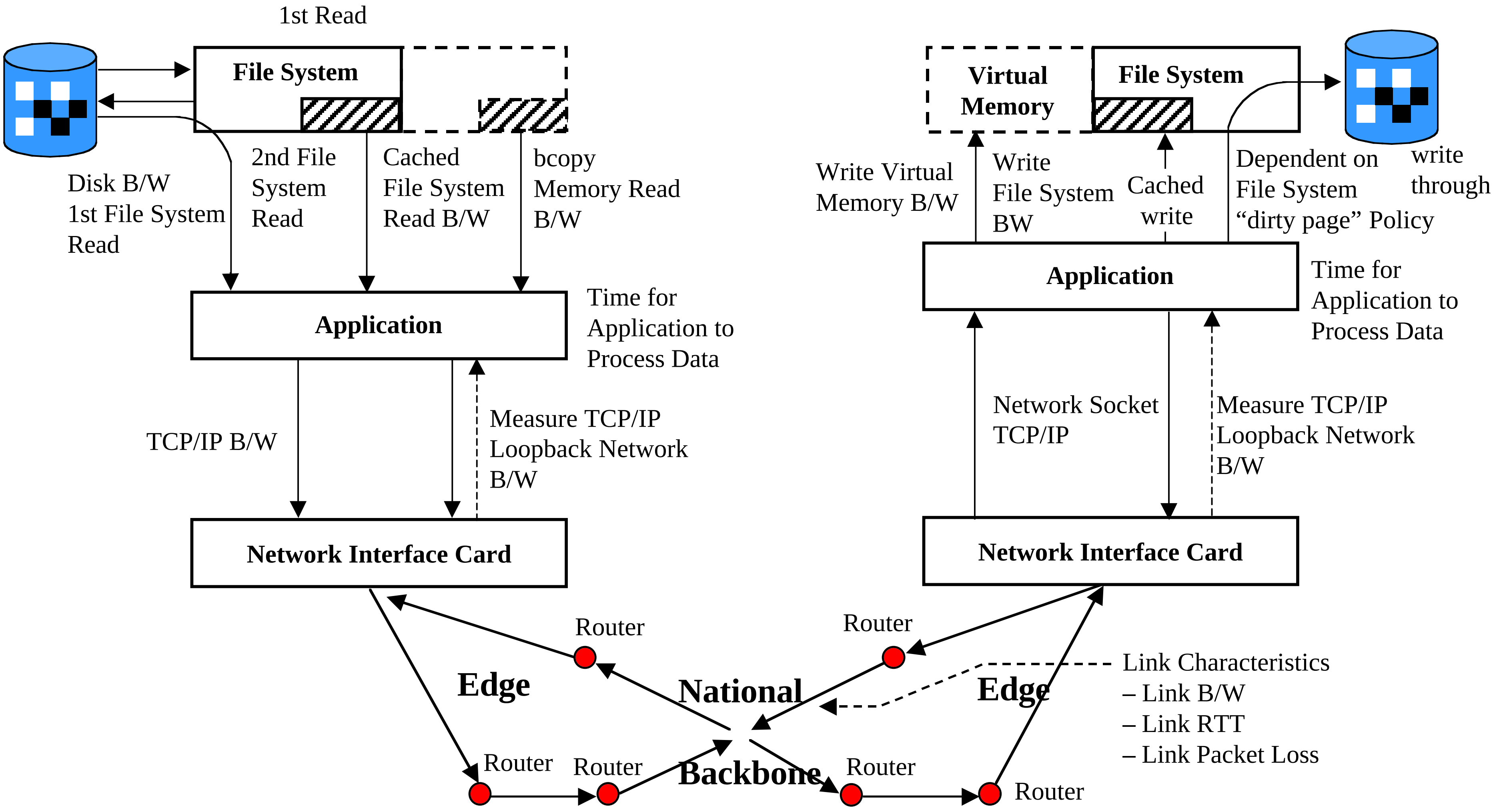
User Skill Instruction Area

	Dental	Surgery	Nursing	Anatomy	Allied Health
Audio					
Video					
Quizzes					
Tests					



Walter Meixner June, 2000

Currently Employed Measurement System for NGI Network Performance



1st Read

File System

Disk B/W
1st File System
Read

2nd File
System
Read

Cached
File System
Read B/W

bcopy
Memory Read
B/W

Application

Time for
Application to
Process Data

TCP/IP B/W

Measure TCP/IP
Loopback Network
B/W

Network Interface Card

Virtual Memory

File System

Write Virtual
Memory B/W

Write
File System
BW

Cached
write

Dependent on
File System
“dirty page” Policy

write
through

Application

Time for
Application to
Process Data

Network Socket
TCP/IP

Measure TCP/IP
Loopback Network
B/W

Network Interface Card

Router

Router

Edge

National

Edge

Backbone

Router

Router

Router

Router

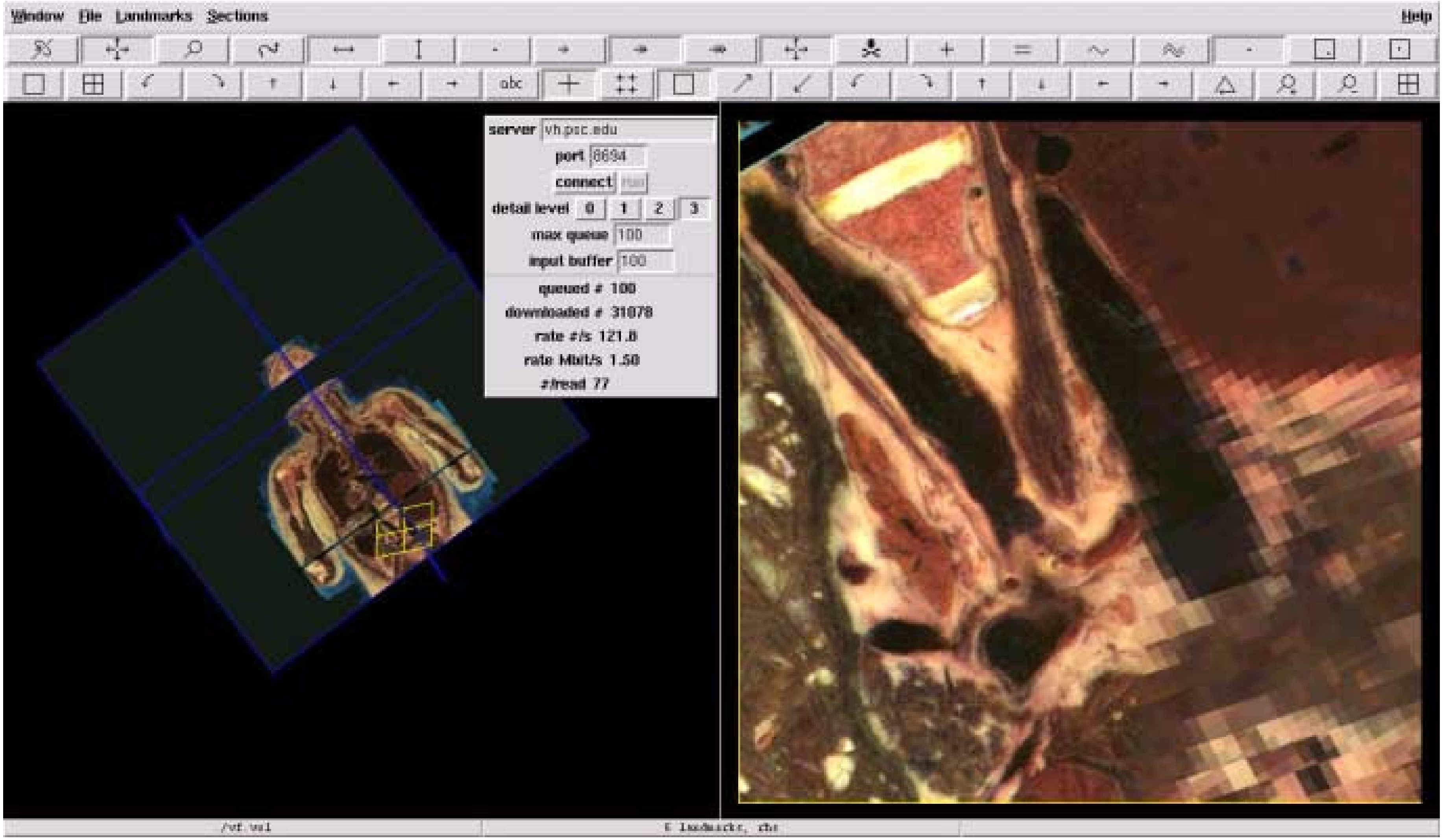
Link Characteristics
– Link B/W
– Link RTT
– Link Packet Loss

- Bi-Directional B/W
- Route Stability “Bi-Directional”
- Latency (RTT)
- Path Length (Hops)

} Internet
Characteristics

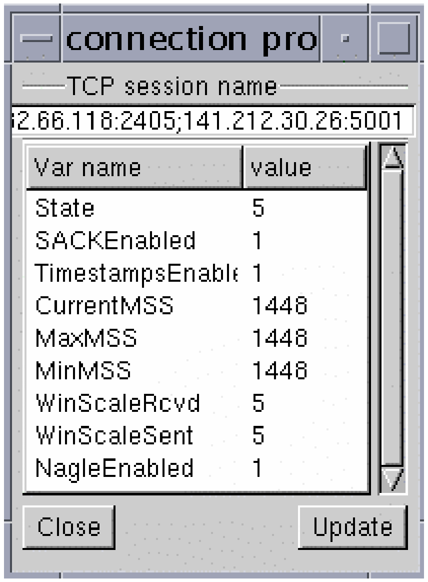
Tom Hacker, June 2000

Edgewarp 3.8: Part of a family of NGI Network-Enabled UM VH software tools



server vh.psc.edu
port 8694
connect
detail level 0 1 2 3
max queue 100
input buffer 100
queued # 100
downloaded # 31078
rate #/s 121.8
rate Mbit/s 1.50
#read 77

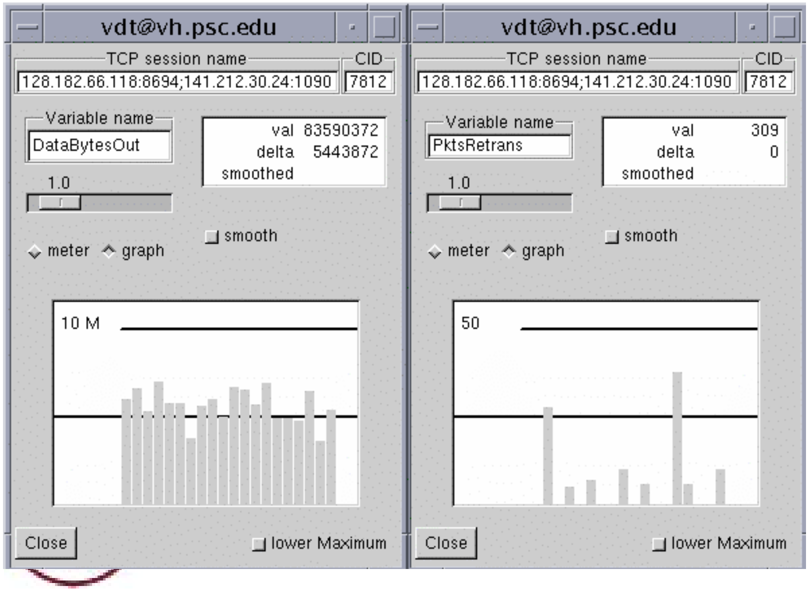
Web100: Well Tuned Host



TCP session name

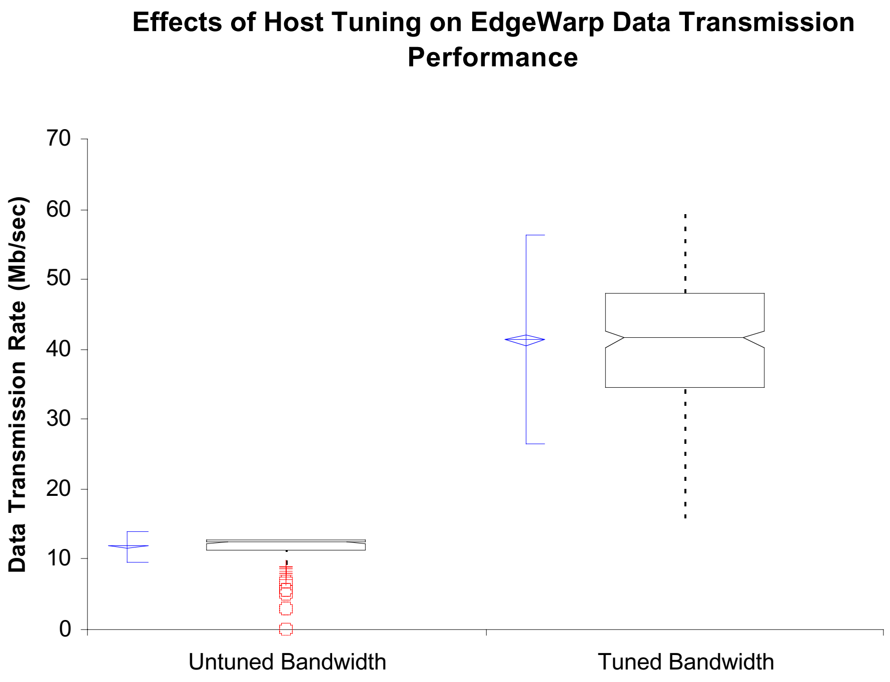
2.66.118:2405;141.212.30.26:5001

Var name	value
State	5
SACKEnabled	1
TimestampsEnable	1
CurrentMSS	1448
MaxMSS	1448
MinMSS	1448
WinScaleRcvd	5
WinScaleSent	5
NagleEnabled	1

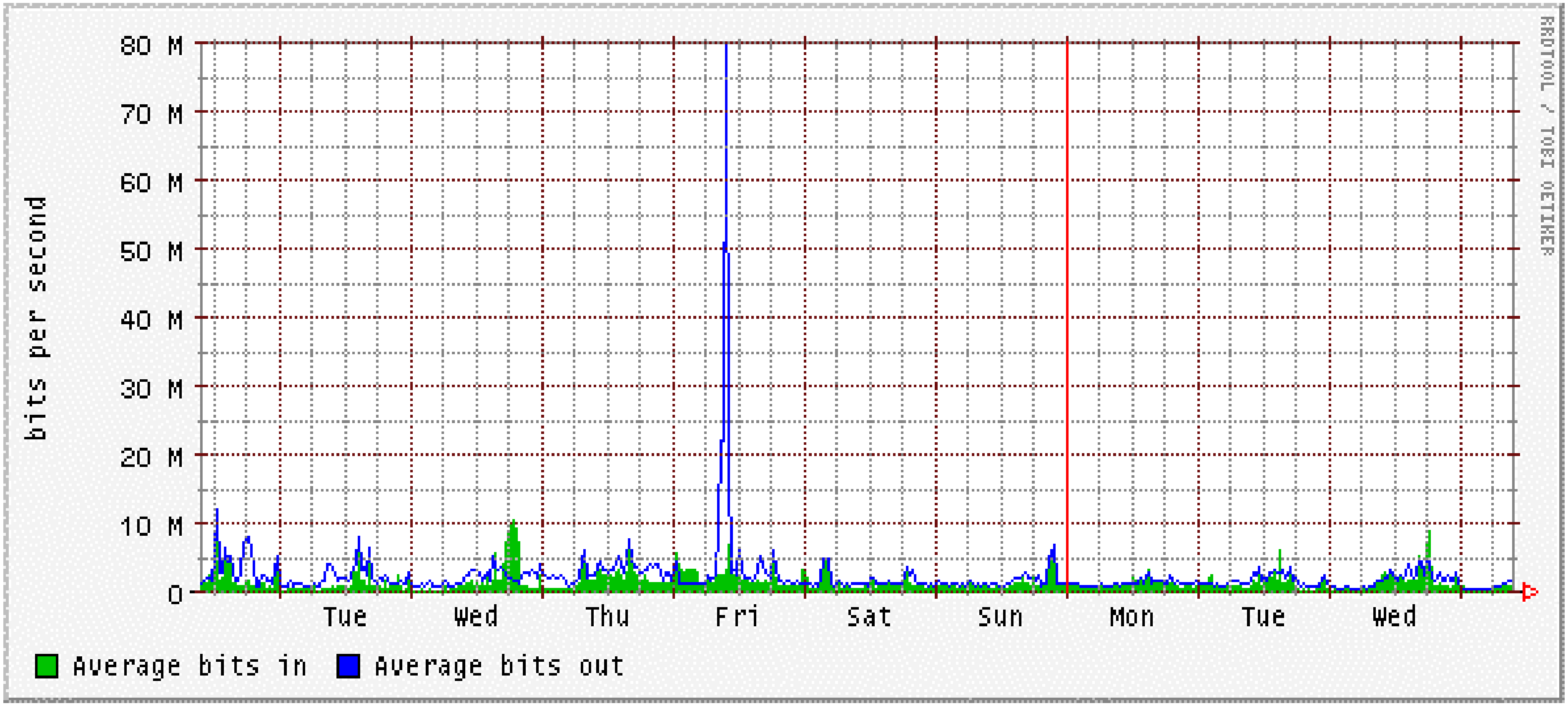


TCP session name: 128.182.66.118:8694;141.212.30.24:1090
CID: 7812
Variable name: DataBytesOut
val 83590372
delta 5443872
smoothed

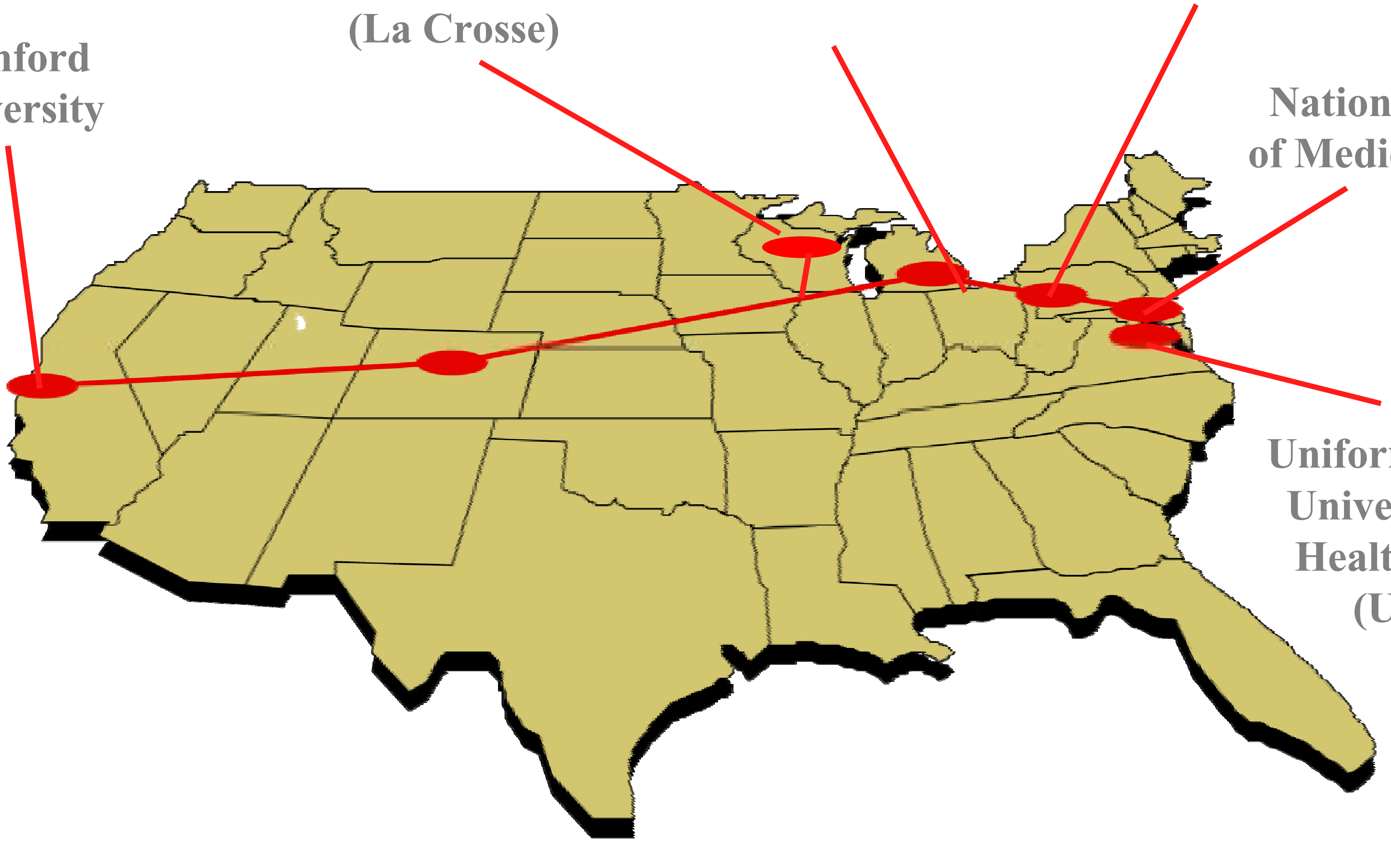
TCP session name: 128.182.66.118:8694;141.212.30.24:1090
CID: 7812
Variable name: PktsRetrans
val 309
delta 0
smoothed


Effects of Host Tuning on EdgeWarp Data Transmission Performance

Network Tuning results on the UM Campus Backbone (CPC to Physics)



Prototype NGI/I-2 Visible Human Collaboratory For Medical Education and Training



Stanford
University

University
Of
Wisconsin
(La Crosse)

University
of
Michigan

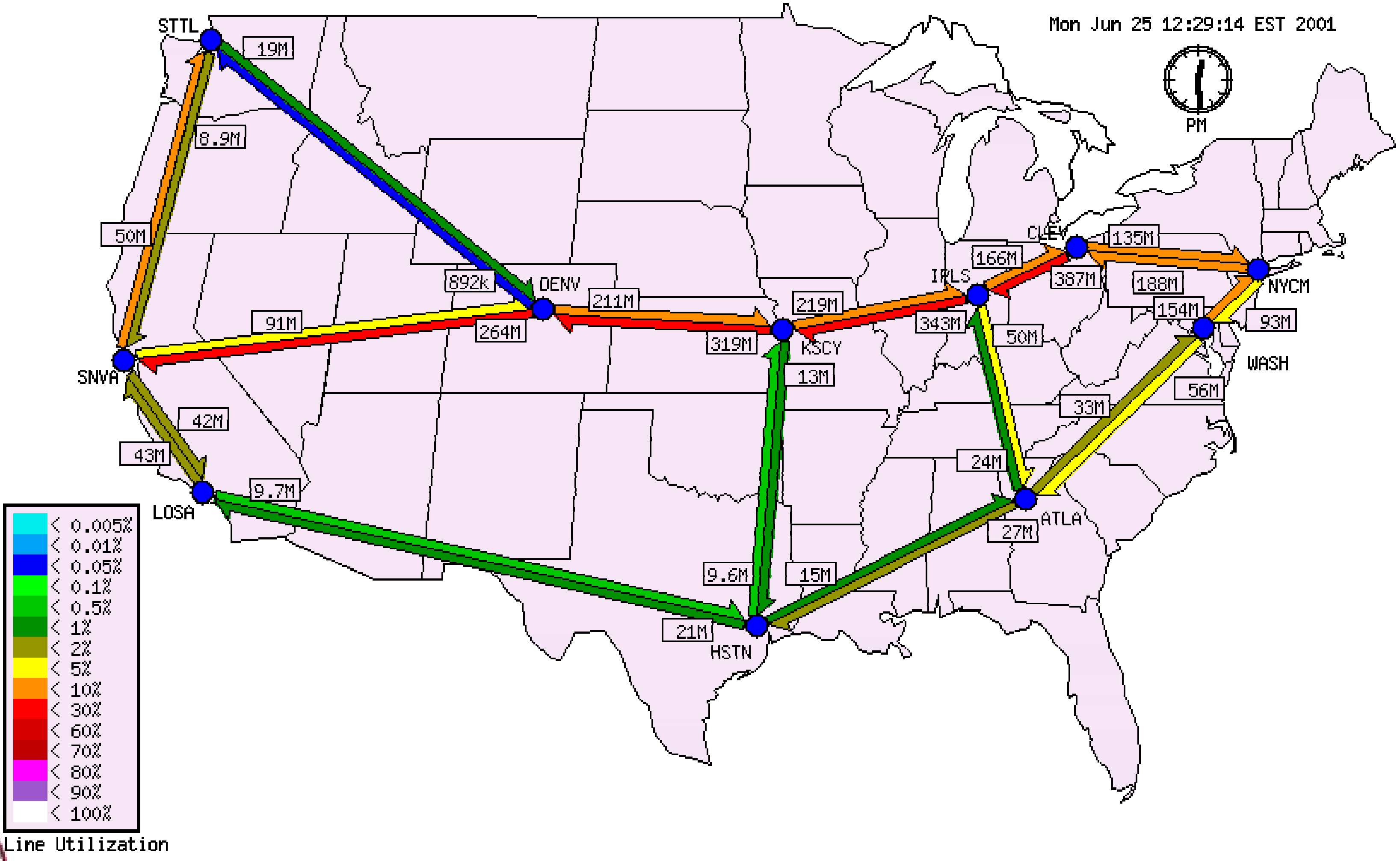
Pittsburgh Supercomputing
Center (PSC)

National Library
of Medicine (NLM)

Uniformed Service
University of the
Health Sciences
(USUHS)

— Abilene Network

National Demonstration at NREN (June 2001): Four Parallel VH Streams— Memory-to-Memory Transfer (UMich to Ames)



Deployment Test-bed and Production Network for the MCBI and the MLSC CTA

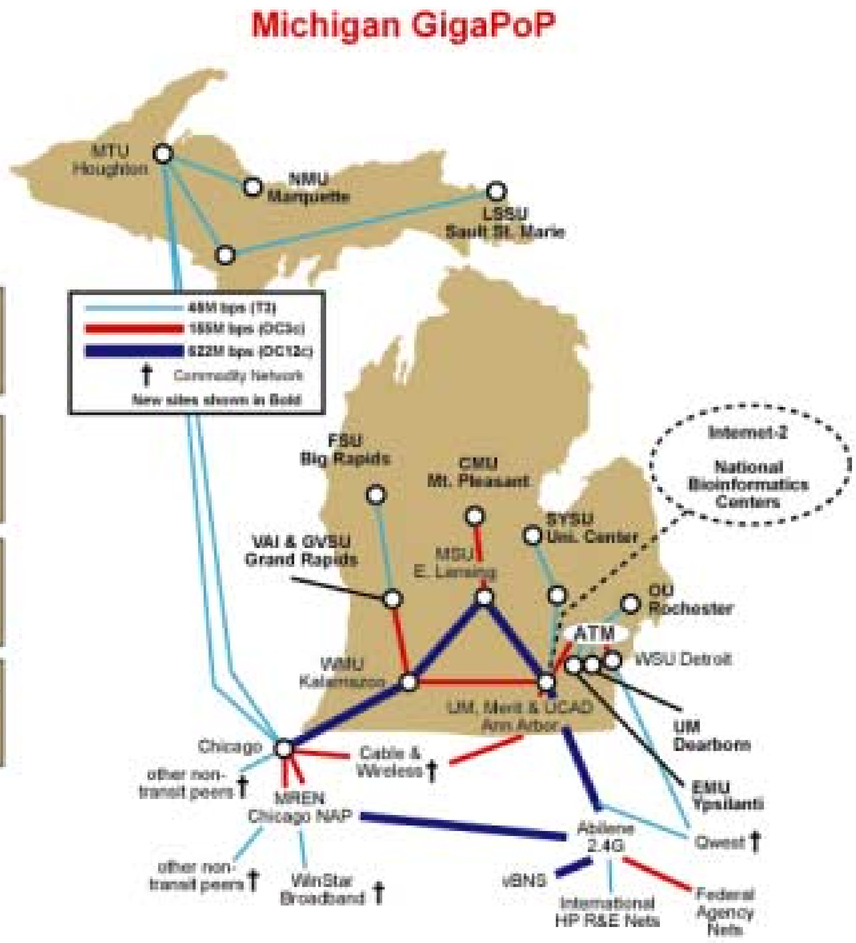

Michigan Life Sciences Corridor

Animal Models (VAI)

Genomics (WSU)

Proteomics (UMICH)

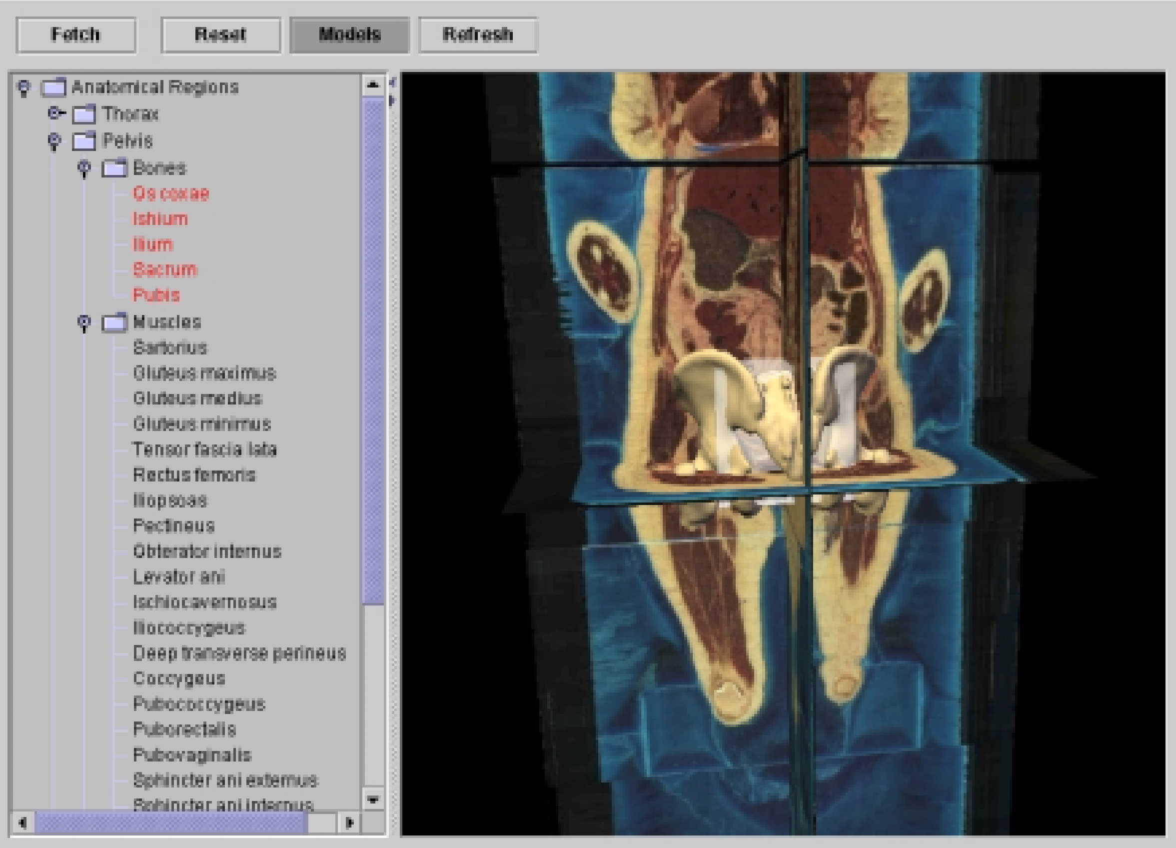
Structural Biology (MSU)


Michigan GigaPoP

Michigan Center for Biological Information Mission Statement

The Michigan Center for Biological Information (MCBI) will provide **secure functional access to a full range of Computational Biology resources** to life sciences researchers and businesses throughout the State of Michigan. **MCBI will provide specialized personnel and assistance, software, modeling and algorithmic expertise, bioinformatics database access, and computational support.** A core mission of the MCBI is to support and project Michigan Life Sciences Corridor (MLSC) infrastructure by helping to **integrate** and support the Cores and Users in the **Core Technology Alliance** (CTA). In this regard, the MCBI will maintain a digital library and archive of the information and data resources of the CTA core and node laboratories. To sustain its activity, the MCBI will participate in a variety of academic, government, and corporate partnerships. The MCBI is hosted by the University of Michigan.

On to the Demonstrations



Emerging Pair-Wise CTA Core Partnerships: Leveraging Special Talents and Activities of the Alliance

- MCBI will leverage and extend MPC Bioinformatics efforts for the corridor as a whole.
- MCBI and MCSB will partner to develop a distributed statewide High Performance Computing image leveraging mutual interests and capabilities in cluster computing.
- MCBI and VARI will lead IT infrastructure design and implementation. Corporate partnerships are being considered (e.g. IBM).
- MCBI will leverage MAMC mouse and rat Genomics expertise.
- MCBI and WSU node will partner to bring specialized bioinformatics expertise to the Corridor.
- MCBI and MCGT must integrate to make the Gene Array Test-bed a success.