Life Sciences and Cyberinfrastructure: a perspective from Indiana University

Dr. Craig A. Stewart
Fulbright Senior Specialist
ZIH, Technische Universität Dresden

Associate Vice President,
Research & Academic Computing;
Chief Operational Officer, Pervasive Technology Labs
stewart@iu.edu

Outline

- Introduction and the situation in Indiana
- Some life science successes based on IU cyberinfrastructure
- Basic computer science research
- IU Cyberinfrastructure
- It takes more than just good science – some thoughts about strategy and execution

- **What is Cyberinfrastructure?**
  - Cyberinfrastructure is a group of high performance computing systems, massive data archives and data resources, visualization systems, advanced instruments, and people all linked together by high speed networks to accomplish tasks and achieve breaktroughs in understanding that would not otherwise be possible
- **What are the life sciences?**
  - Biology, organic chemistry, analytic chemistry, many areas of psychology, some areas of geography and geology, environmental sciences but typically not atmospheric sciences/global environmental change

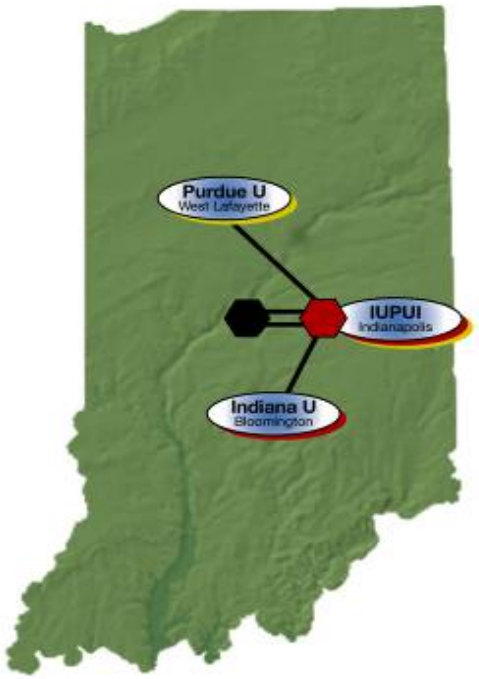
The situation in Indiana

- Indiana's economy is traditionally based on steel and heavy industry
- Indiana has been a national leader: personal bankruptcies, mortgage foreclosures, job losses

- Indiana has a strong tradition in life science industries
- Since the mid 1990s Indiana has developed a strong presence in Information Technology

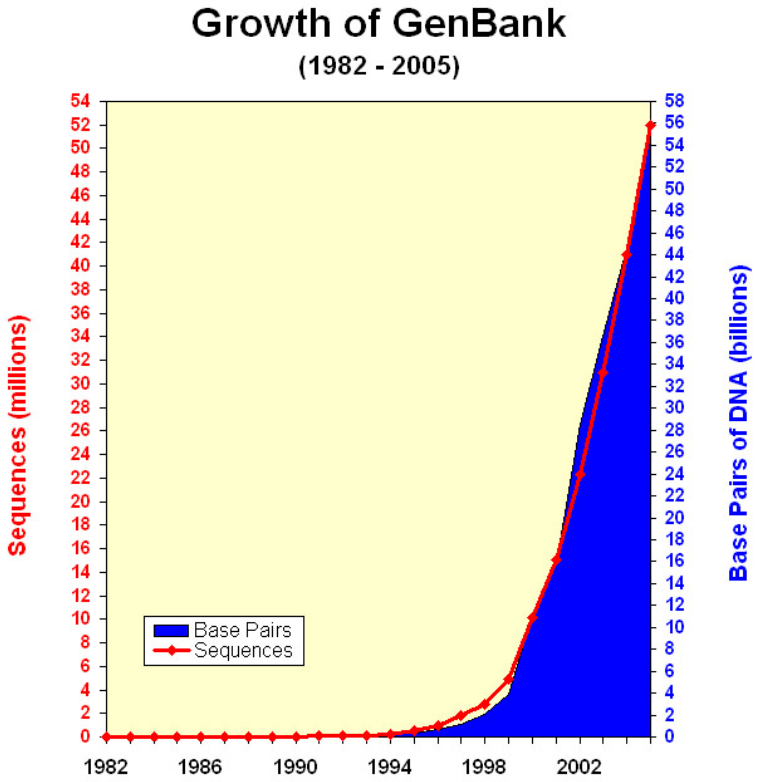
Life sciences & Cyberinfrastructure strategy for the State of Indiana

- The State of Indiana has set a strategy based on a combination of life sciences and information technology (cyberinfrastructure) supported by the State Government, private industry, public consortia (e.g. Biocrossroads.org), private charitable trusts, and the State's main universities.
- Indiana University set a strategic goal in the mid 1990s to become a "leader in absolute terms in the development, deployment, and use of information technology"
- IU has long tradition in life sciences
- Indiana University has most recently created a life science strategic plan, and identified Life Sciences and Information Technology as the two leading goals for the University



Why Life Sciences and Cyberinfrastructure?

- Growth of basic scientific data
- Possibilities of developing new medical therapies ($!)
- Growth area for high performance computing
- Need for management of clinical data (Baycol example). If you are taking any medication, you are part of an experiment!



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

So why isn't this easy?

- A large portion of the High Performance Computing (HPC) community knows relatively little about life sciences
- A large portion of the global life sciences research community knows relatively little about HPC
- Real need for scientific advancement in theory and in data
- In the land of the blind….
- So the real challenge for the HPC and life science communities is to learn to work together for mutual benefit and to improve the quality of life of those who pay the bills.
- The key for HPC centers: working with "real users" to deliver real innovation and results

IU in a nutshell

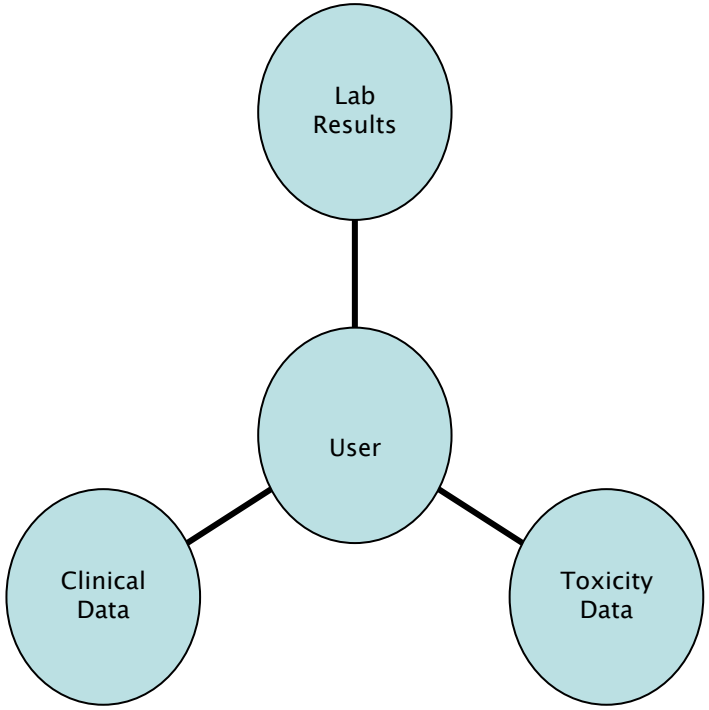
- $2B Annual Budget
- One university with 8 campuses; 90.000 students, 3.900 faculty
- 878 degree programs, including nation's 2nd largest school of medicine
- IT Organization:
  - Vice President for IT & Provost: Michael A. McRobbie
  - CIO: Bradley C. Wheeler
  - 4 Divisions: Telecommunications, Teaching & Learning, University Information Services, Research and Academic Computing
  - Several offices: Security, Human Resources, Communications
  - >$100M/year IT budget

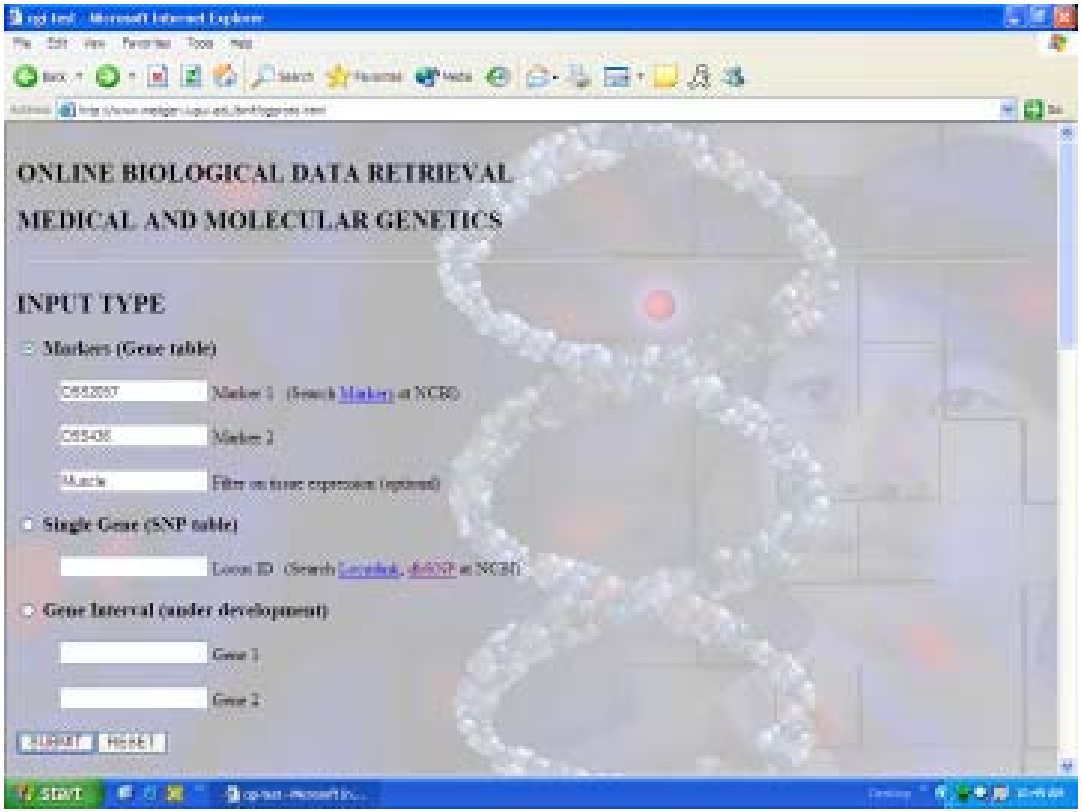
Timeline of key events

- 1997: IU sets goal to become leader in absolute terms in IT
- 1997: IU acquires 64 GFLOPS SGI Origin2000, upgrades IBM SP2
- 1999: $35M grant from Lilly Endowment to create Pervasive Technology Labs
- 1999: Indiana Governor Frank O'Bannon approves I-Light network connection Purdue, Indiana, Bloomington
- 2000: $155M grant from Lilly Endowment for Indiana Genomics Initiative
- 2001: IU announces first US university-owned supercomputer with > TFLOPS peak
- 2003: IU announces 2.2 TFLOPS distributed Linux cluster, first distributed Linux cluster with > 1 TFLOPS achieved
- 2003: IU awarded grant to become part of the TeraGrid
- 2004: $35M grant from Lilly Endowment for METACyt
- 2005: IU announces strategic plan for Life Sciences, announces 20.4 TFLOPS BladeCenter

Some life science innovations that involve cyberinfrastructure

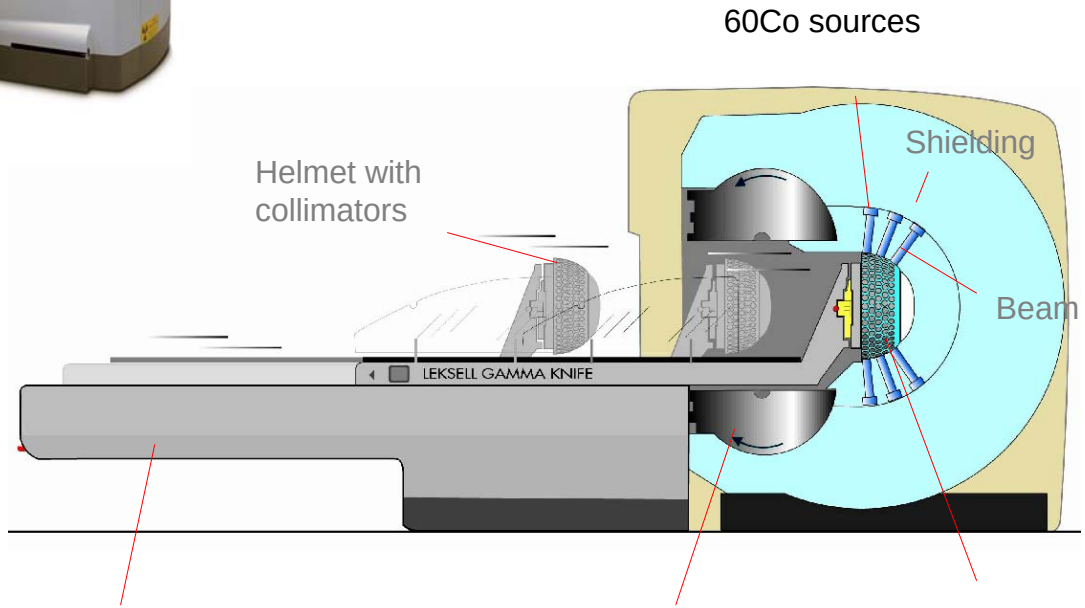
Information access - Idealized View





Leskell Gamma Knife

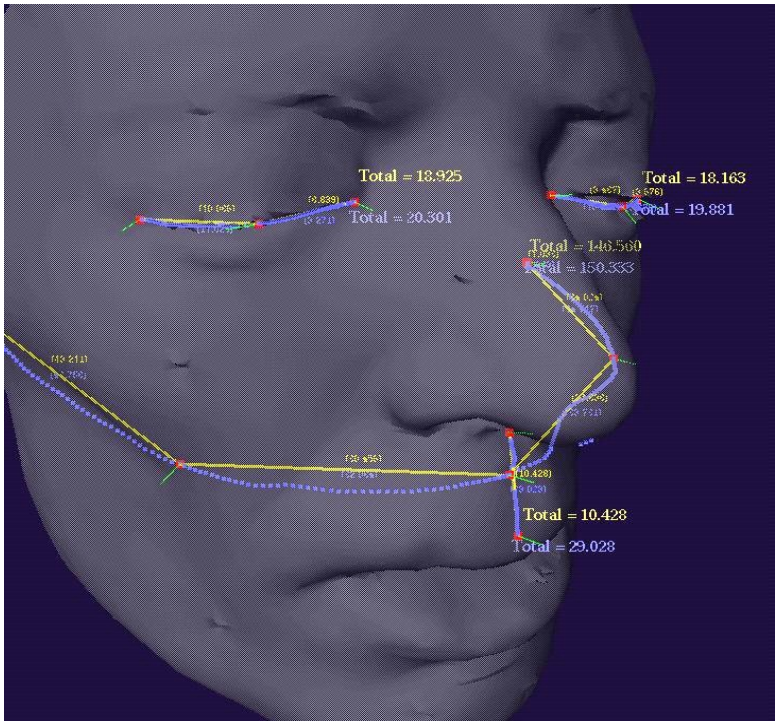
- Used to treat inoperable tumors
- Lance Armstrong was treated in IU's Gamma Knife



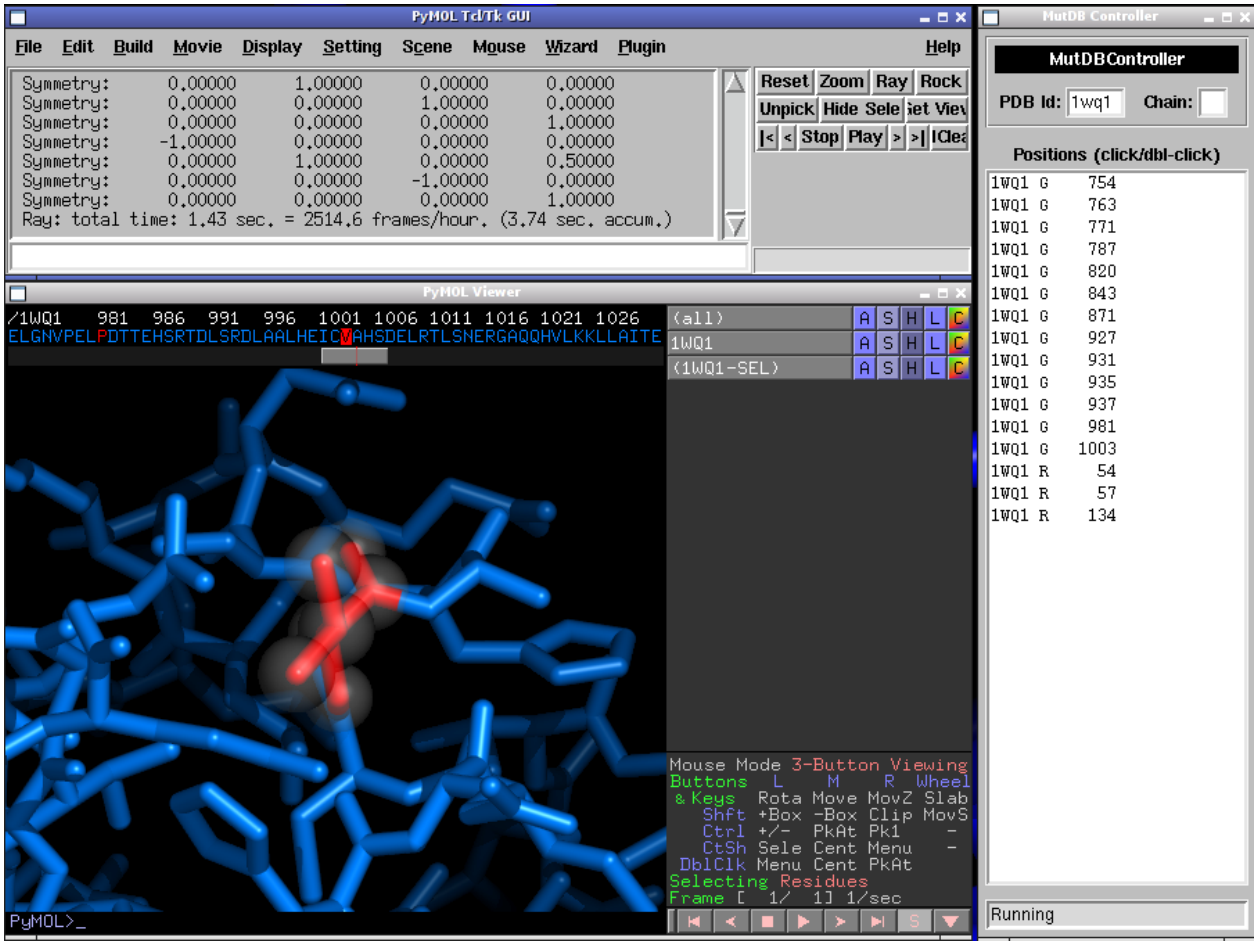
Gamma Knife

- An idealized head model is used for target planning
- When treatment fails to be successful, the primary problem thought to be targeting
- Solution: use a model of individual patient's head to plan targeting

Collaborative Initiative on Fetal Alcohol Spectrum Disorder

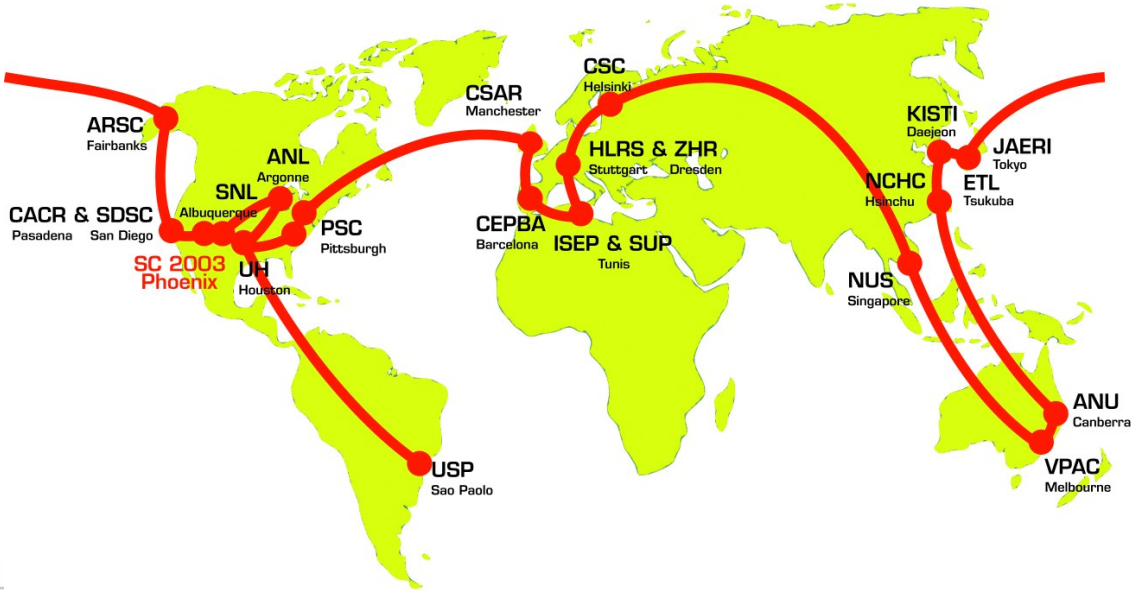


MutDB


PyMOL Tcl/Tk GUI
File Edit Build Movie Display Setting Scene Mouse Wizard Plugin Help
Symmetry: 0.00000 1.00000 0.00000 0.00000
Symmetry: 0.00000 0.00000 1.00000 0.00000
Symmetry: 0.00000 0.00000 0.00000 1.00000
Symmetry: -1.00000 0.00000 0.00000 0.00000
Symmetry: 0.00000 1.00000 0.00000 0.50000
Symmetry: 0.00000 0.00000 -1.00000 0.00000
Symmetry: 0.00000 0.00000 0.00000 1.00000
Ray: total time: 1.43 sec. = 2514.6 frames/hour. (3.74 sec. accum.)
Reset Zoom Ray Rock
Unpick Hide Sele
Stop Play
PyMOL Viewer
/1WQ1 981 986 991 996 1001 1006 1011 1016 1021 1026
ELGNVPELPDTTEHSRTDLSRDLAALHEICMAHSDELRTLSNERGAQQHVLKKLLAITE
(all)
1WQ1
(1WQ1-SEL)
Mouse Mode 3-Button Viewing
Selecting Residues
Frame [1/ 1] 1/sec
PyMOL>_
MutDB Controller
MutDBController
PDB Id: 1wq1 Chain:
Positions (click/dbl-click)
1WQ1 G 754
1WQ1 G 763
1WQ1 G 771
1WQ1 G 787
1WQ1 G 820
1WQ1 G 843
1WQ1 G 871
1WQ1 G 927
1WQ1 G 931
1WQ1 G 935
1WQ1 G 937
1WQ1 G 981
1WQ1 G 1003
1WQ1 R 54
1WQ1 R 57
1WQ1 R 134
Running

Global Analysis of Arthropod Evolution

- Winner, “Most geographically distributed application,” High Performance Computing Challenge at SC2003.
- Created a global grid of computers including 14 systems; 8 types of systems; 6+ vendors; 641 processors; 9 countries, 6 continents (every continent except Antarctica).
- Demonstration accomplished computationally intensive evolutionary research



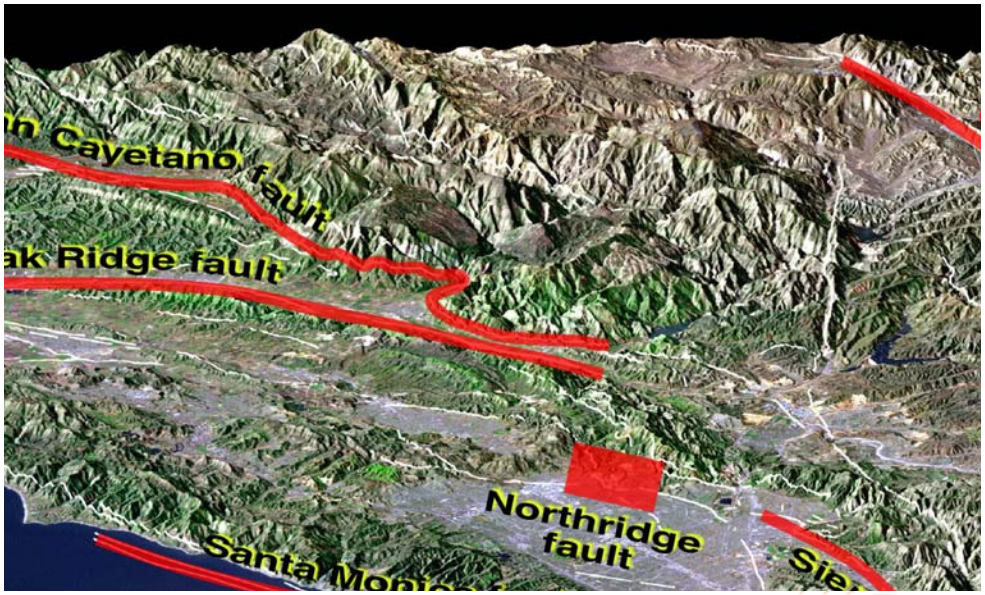
Basic computer science research and cyberinfrastructure

Pervasive Technology Labs

- *The mission of the Pervasive Technology Labs at Indiana University is to:*
- *perform leading-edge research based on the pervasiveness of information technology in our world, creating new inventions, devices, and software that extend the capabilities of information technology in advanced research and everyday lives;*
- *attract, encourage, educate, and retain the workforce of tomorrow for the State of Indiana and educate the residents of the State generally about the value of advanced technology;*
- *accelerate economic growth in the State of Indiana through the commercialization of new software and inventions; and*
- *develop an income stream through external grant funding and technology transfer revenue that will lead toward self-sustainability for the Labs.*
- *In carrying out its mission, the Pervasive Technology Labs will help Indiana University attain a position of international leadership in information technology research and enhance the prosperity of the entire State of Indiana.*

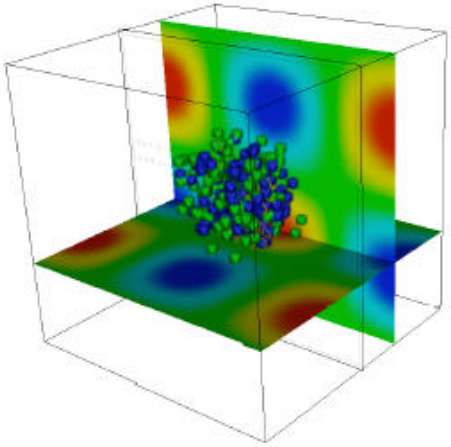
Community Grids Lab

- Director:
  - Geoffrey Fox
- Focus on Collaboration
  - Using the network to help groups work together.
- Information and Computing "Grid"
- Science Applications
  - Biocomplexity
  - Earthquake Science
  - Fusion Energy

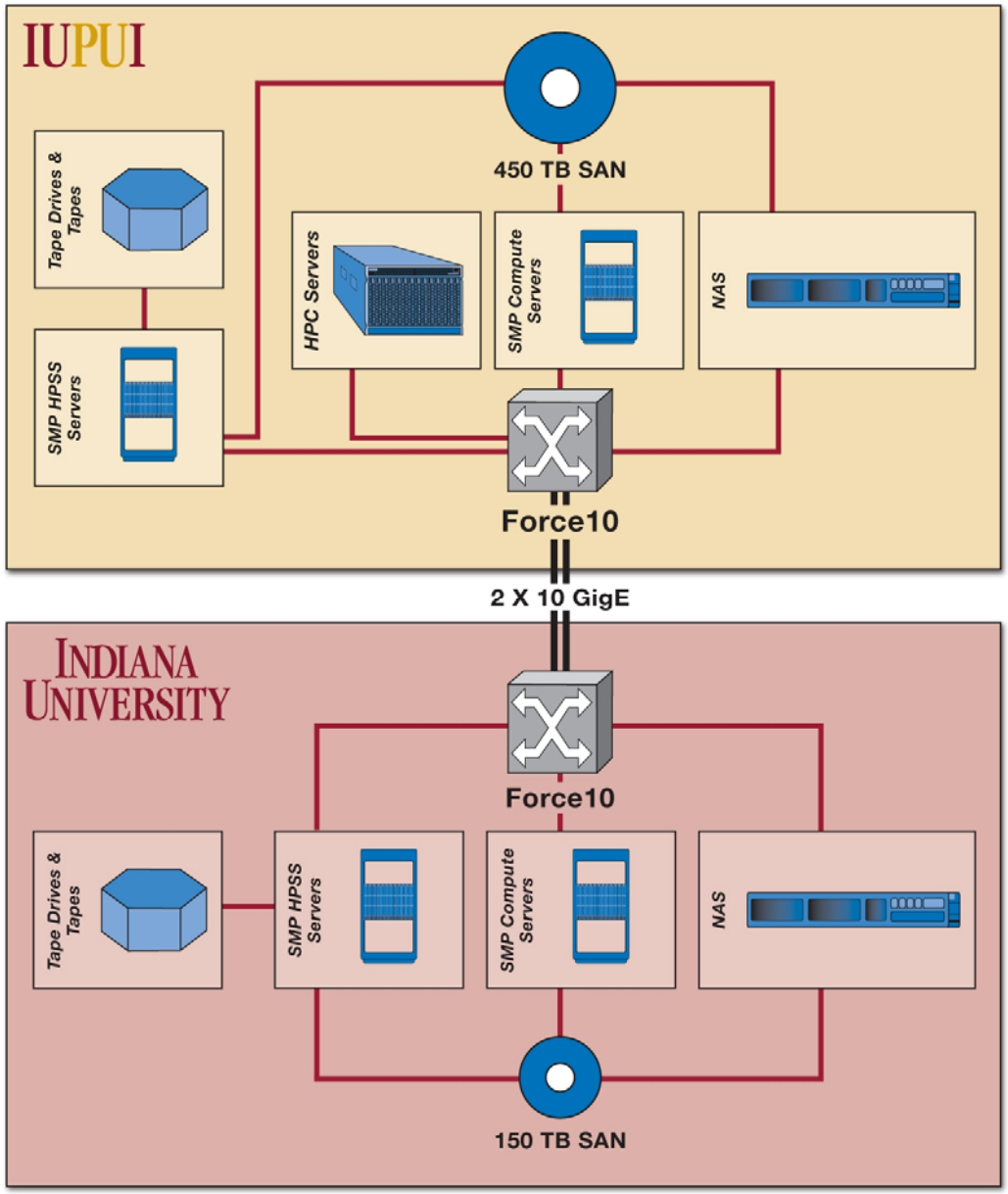


Open Systems Lab

- Director
  - Andrew Lumsdaine
- Software for Supercomputers and Science
    - Tools for building very complex systems.
    - Open MPI
    - Boost graph library

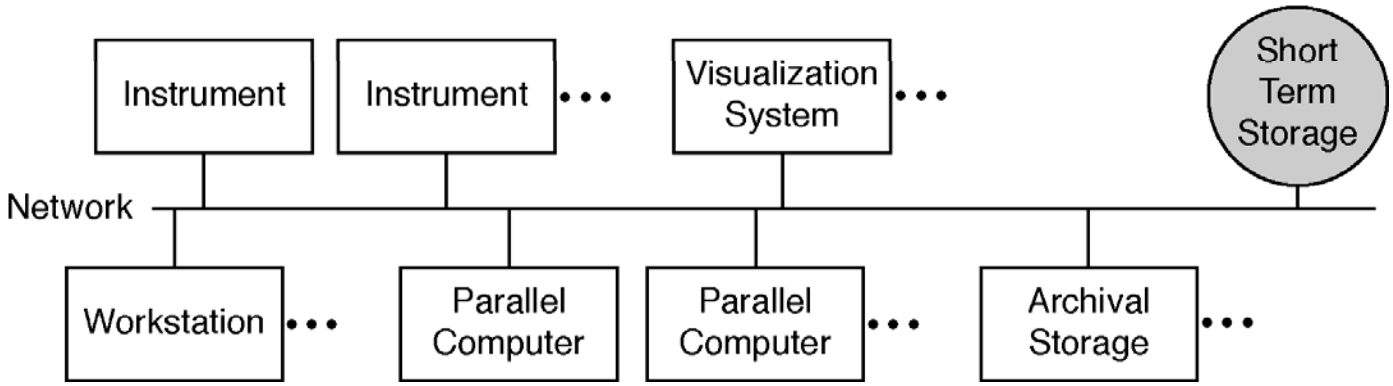


IU Cyberinfrastructure


IUPUI
450 TB SAN
Tape Drives & Tapes
HPC Servers
SMP Compute Servers
NAS
SMP HPSS Servers
Force10
2 X 10 GigE
INDIANA UNIVERSITY
Force10
Tape Drives & Tapes
SMP HPSS Servers
SMP Compute Servers
NAS
150 TB SAN

Data Capacitor

- Short term management of massive amounts of data
- Data produced by instruments, data as part of workflows, data stored temporarily, summarized, and thrown away



Basic system statistics

- Computational systems
  - IBM BladeCenter Cluster: 20.4 TFLOPS (512 dual processor JS21 Blades, each with two dual-core PowerPC 970 MPs). 8 GB RAM per Blade
  - IBM p575 cluster: 1.6 TFLOPS (8-way Power-5 based nodes – half with 16 GB RAM, half with 32 GB RAM)
- 1.15 PB spinning disk
  - 650 as SAN supporting research systems:
    - GPFS parallel disk system
    - Network Attached Storage
    - AFS Cell
  - 500 TB for Data Capacitor

Challenges in HPC

- Management of the size of the system
  - Reliability and resilience
  - Electrical and cooling facilities
- Management of use of the system
  - Levels of parallelism
  - Reliability and resilience (Open MPI)
  - Performance (OTF, Vampir NG)

Long term data management

- Life science data is essentially irreplaceable
- Duplicate copies of archived data kept in Bloomington and Indiana
- ~ 1 PB of spinning disk overall,
- Currently 2 PB of tape and adding 1 PB per year
- Metadata services, management of provenance, and management of availability of data are key areas of focus for us in the future
- Data provided via web services will be a key matter as well
- Putting management and retrieval of massive data stores, and use of that infrastructure in massive simulations, to produce meaningful results and important innovation is a key problem

TeraGrid

- Collaborative development of new computer technologies *and delivery of new scientific innovations*
- Linked systems – massive, dynamic

It takes more than just science – some thoughts about strategy and execution

Tech transfer by RAC: John-E-Box

Invented by John N. Huffman, John C. Huffman, & Eric Wernert, IU

Creating the 21st Century Workforce

- RAC/UITS
  - 2004 Grace Hopper Celebration
  - 2005 Richard Tapia Conference
- PTL
  - Outreach to Native Americans
  - Graduate Students
- PTL & RAC
  - Outreach to lay public
  - Indianapolis Museum of Art, Indiana State Museum

Promoting Indiana - SuperComputing Conference

- IU and Purdue collaboration on booths starting in 2000
- Excellent national attention
- Helped build many collaborations, including successful TeraGrid proposal

Research & Academic Computing Strategy

Our Work

Our Objective

Front Office

Back Office

Reliable Services

Co-Creating the Future

Researcher Consulting & Education

Systems Administration

Grant Initiation, Collaboration, Fulfillment

Engineering Computing Frontiers

INDIANA UNIVERSITY INFORMATION TECHNOLOGY SERVICES

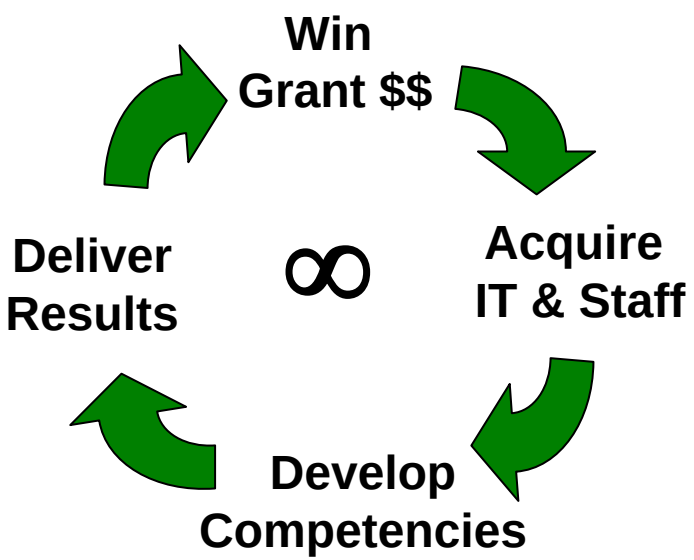
RESEARCH AND ACADEMIC COMPUTING

ZIH

Zentrum für Informationsdienste

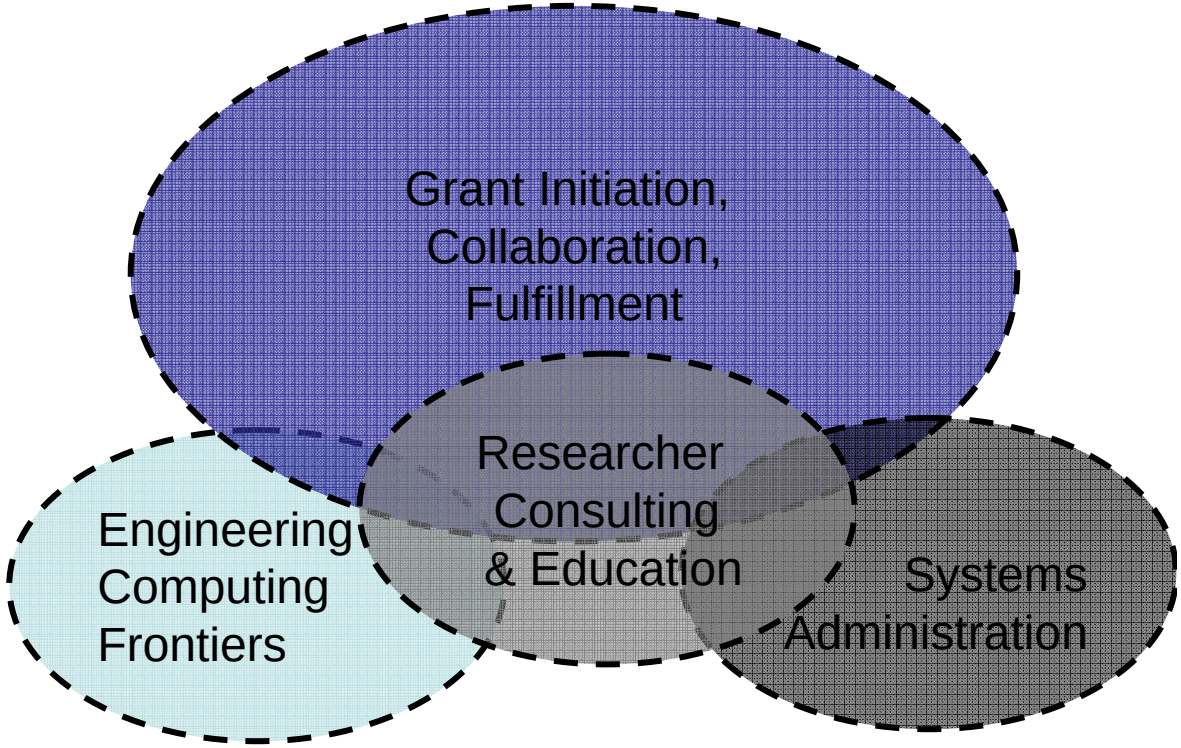
Dr. Kate Pilachoski, Professor of Astronomy

“Double External Funding by AY10-11”



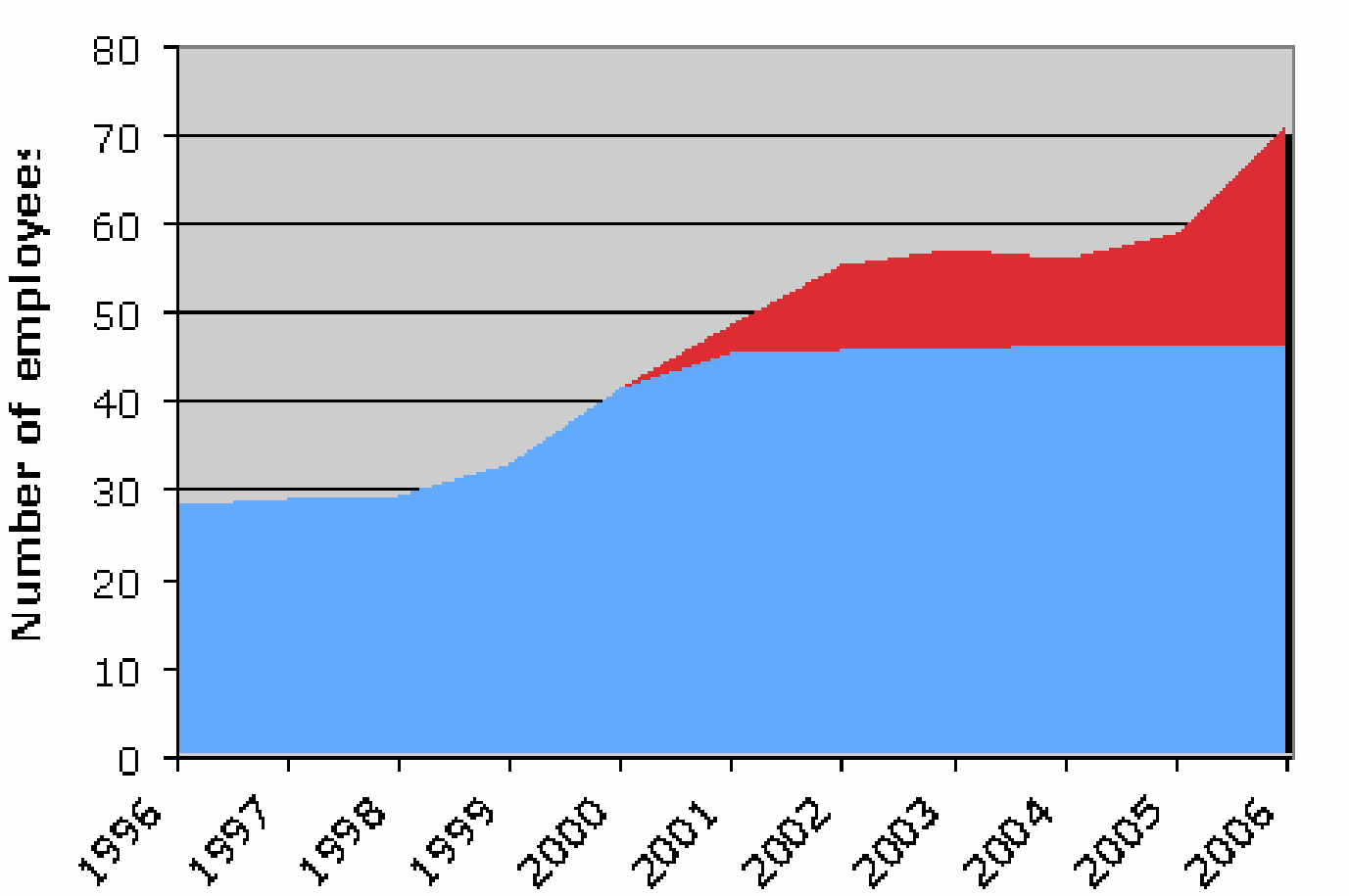
Ever Advancing Frontiers…

- High Performance Computing
- Mass Research Storage
- Visualization
- Networks (Telecom)
- Consulting (Stat, Linux)
- Digital Libraries



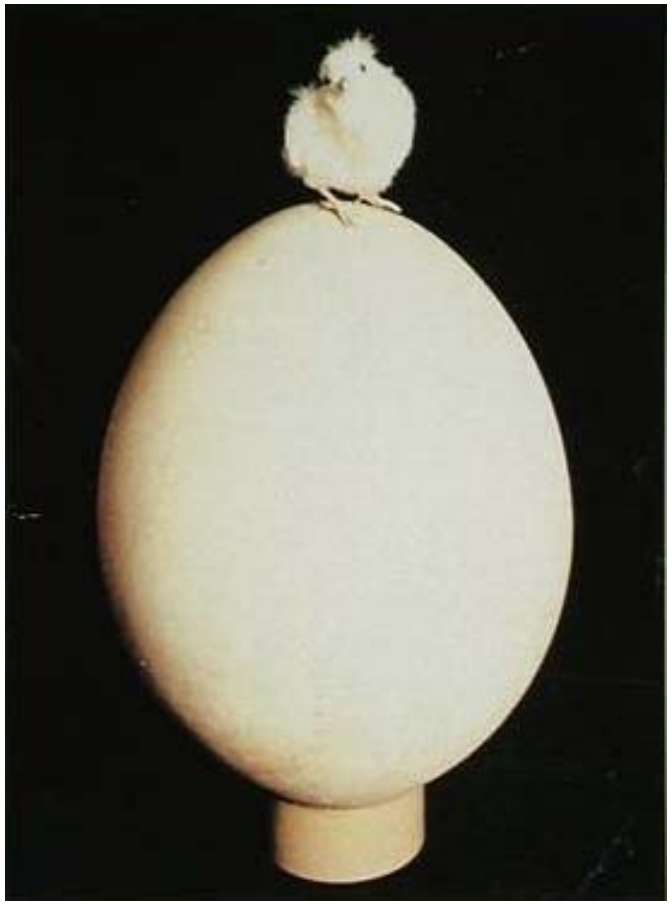
RAC Works via Relationships & Technical and Domain Competence

Execution: Funding and Staffing



Some final thoughts on life science and cyberinfrastructure collaborations

- Chicken and egg problem, or bank robbery?
- There are lots of opportunities open for HPC centers willing to take the effort to cultivate relationships with biologists and biomedical researchers – but we as HPC exerts will have to go where the biologists are
- There are LOTS of opportunities available for universities willing to commit to information technology and life sciences as joint strategies
- The joint IT/Life sciences strategy for IU is working for the State of Indiana
- There are many similarities between the strategies of TU-D and IU!



Acknowledgments

- Funding for projects described in this talk has come from the National Science Foundation, National Institutes of Health, Lilly Endowment, Inc., State of Indiana (particularly through support of I-light Initiative and the 21st Century Fund)
- The work described here was made possible by the faculty, students, and staff of Indiana University. Thanks especially to the staff of RAC, CPO, Telecommunications, PTL, UITS generally, the participants in the Indiana Genomics Initiative, and the participants in the METACyt Initiative.
- Several of the slides and ideas presented here were developed by colleagues or collaborators – the Research and Academic Computing Division of UITS in general, and Dick Repasky in particular.
- Stewart's visit to Dresden is funded in part by the Center for the International Exchange of Scholars, the Technical University of Dresden, and Indiana University. THANKS!

For additional info

- rac.uits.indiana.edu/
- www.iu.teragrid.org/
- Life Sciences Strategic Plan - http://lifesciences.iu.edu/