

# High Performance Computing in Medicine and Pathology

Joel Saltz - Department of Pathology, JHMI  
University of Maryland Computer Science

## University of Maryland

Asmara Afework

Mike Beynon

Charlie Chang

John Davis

Renato Ferreira

Bongki Moon

Kilian Stoffel

Alan Sussman

Mustafa Uysal

## Johns Hopkins Medical Institutions

Angelo Demarzo

Jim Dick

Bill Merz

Robert Miller

Jerry Rottman

Mark Silberman

Kilian Stoffel

Merwyn Taylor

# Goals

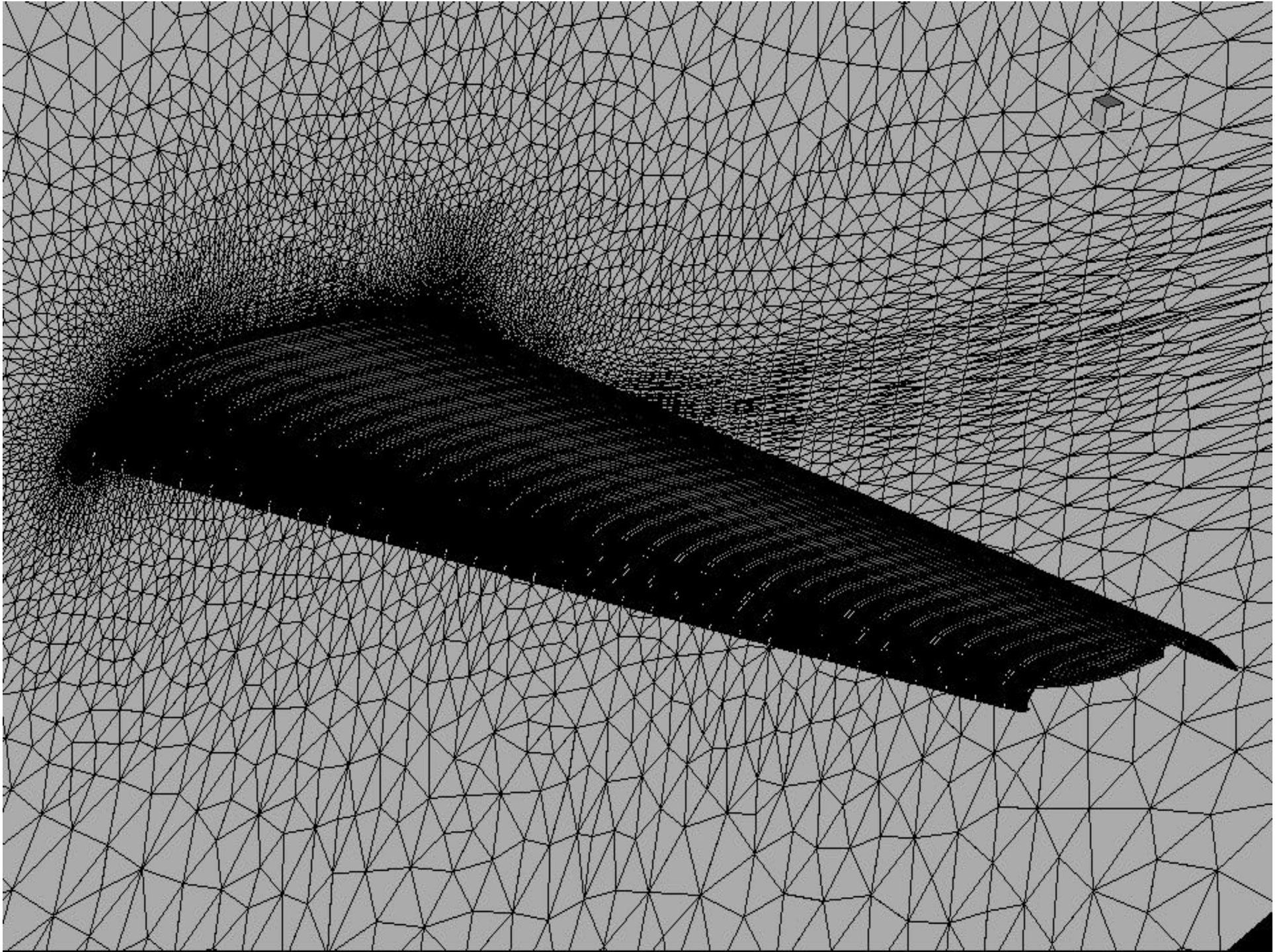
- Applications from different fields have common computational characteristics
- Exploit knowledge of common computational characteristics, and knowledge of advanced computer architecture
  - build tools to optimize performance of groups of problems on advanced computer architectures

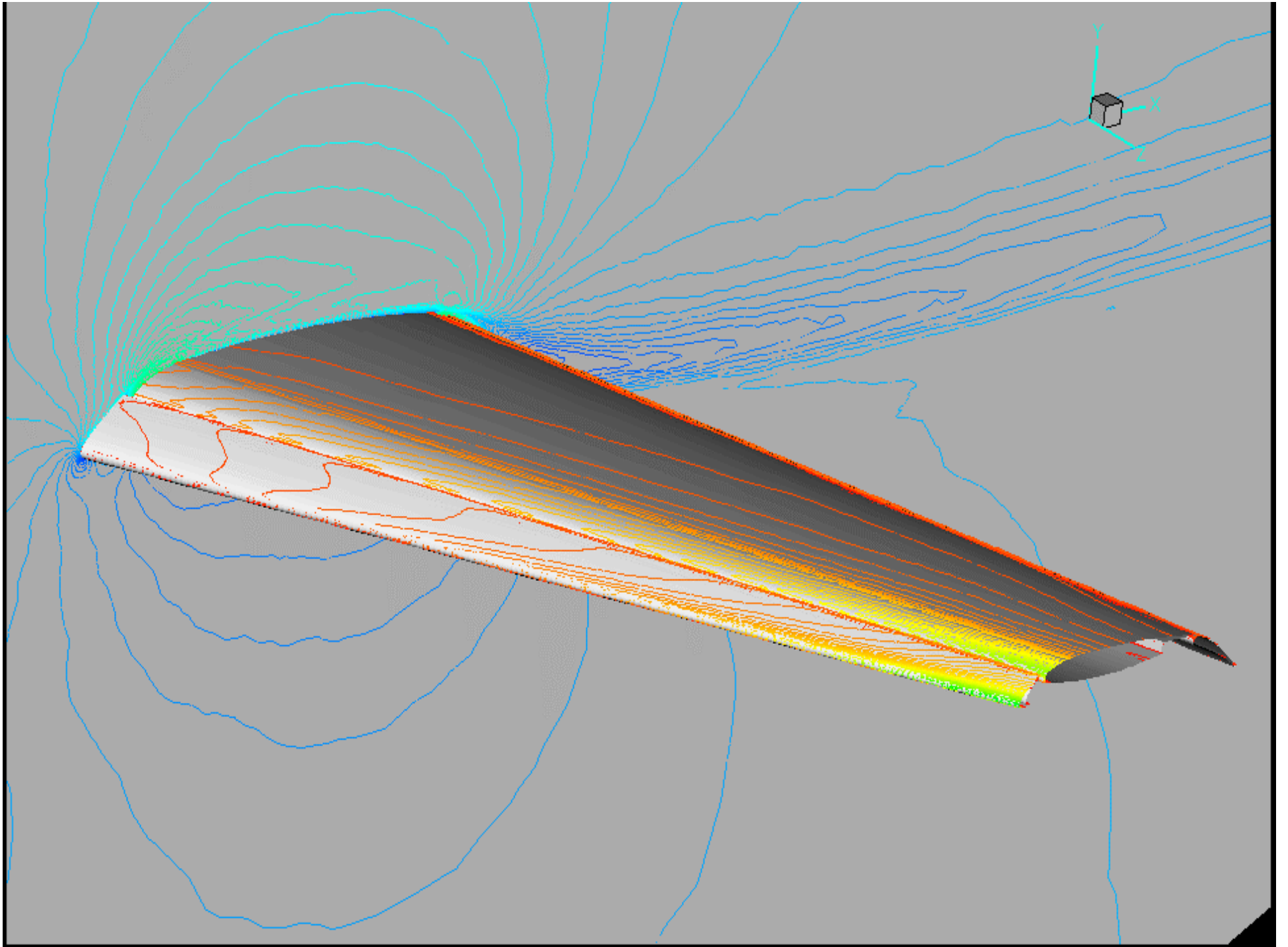
# Roadmap

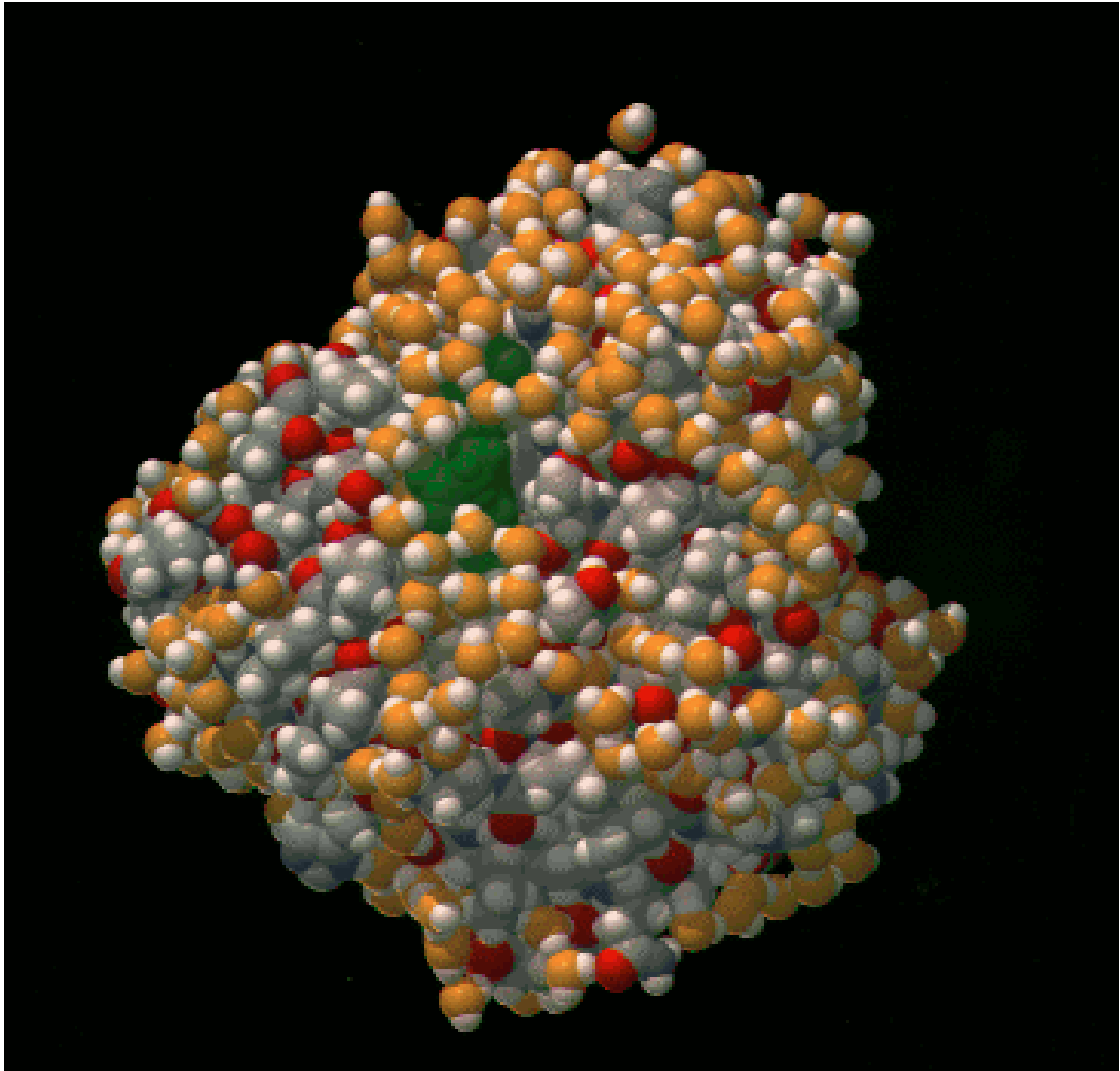
- Ambitious Simulations, Sensor Data Processing and Data Mining
- Computer Architecture Made Ridiculously Simple
- High Performance Databases and Systems Software

# Ambitious simulations

- Design of aircraft - integrate structures, aerodynamics, engine design
- Simulation of conventional or nuclear explosions
- War game simulation
- Computational chemistry
  - relationship between chemical composition and structure
- Detailed simulations of cardiac function, hemodynamics and cardiac conduction
- Detailed simulation of physiological neural network







# Sensor Data

- Sensors obtain data about the world
  - satellite sensors acquire data at different wavelengths from earth, planets and universe
  - radiology devices acquire biomedical datasets
  - microscope with a robotic stage acquires
  - synthetic aperture radar produces datasets used in military operations

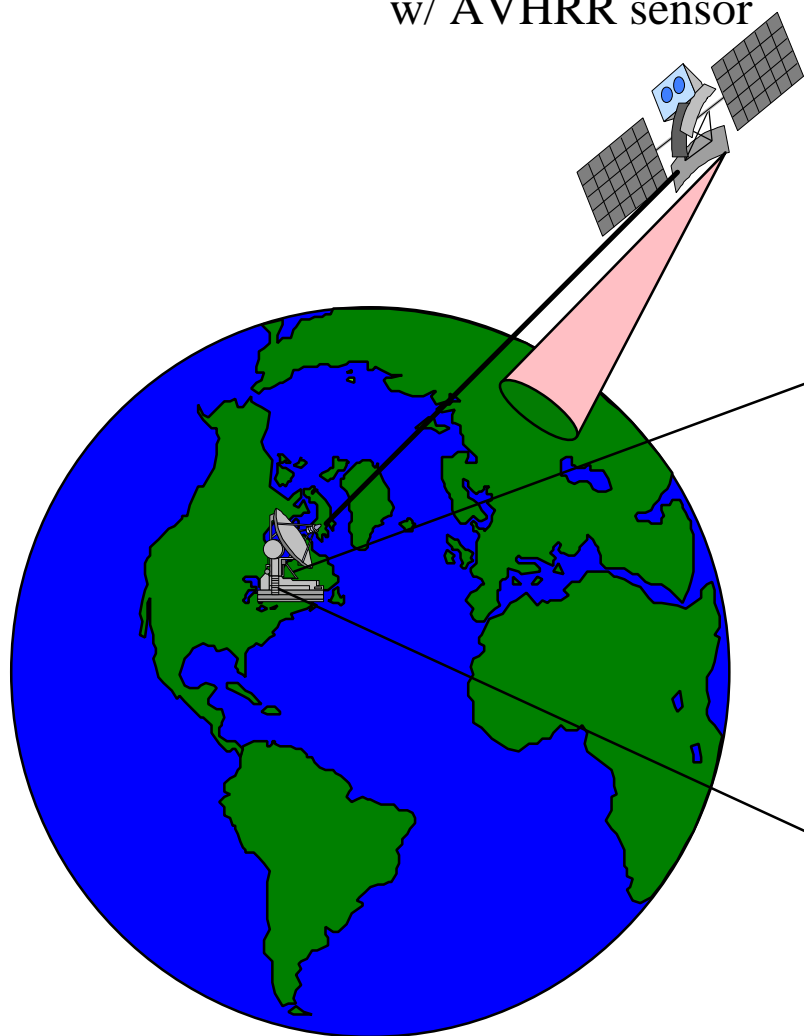


# Satellite Data

- Land cover and changes in land use
- Year-in-advance rainfall and temperature predictions
- Long term climate predictions
- Short, medium and long term flood prediction

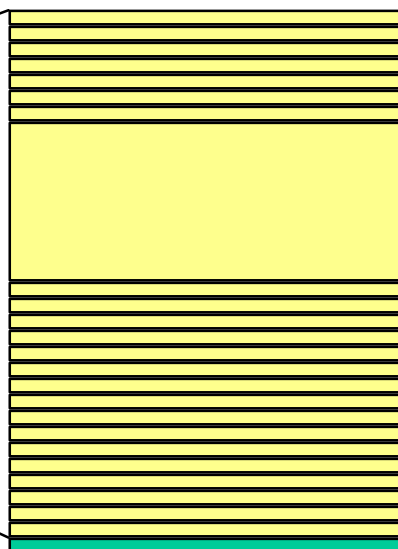
# Processing Remotely Sensed Data

NOAA Tiros-N  
w/ AVHRR sensor



## AVHRR Level 1 Data

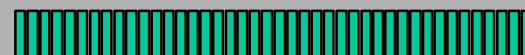
- As the TIROS-N satellite orbits, the *Advanced Very High Resolution Radiometer* (AVHRR) sensor scans perpendicular to the satellite's track.
- At regular intervals along a scan line measurements are gathered to form an *instantaneous field of view* (IFOV).
- Scan lines are aggregated into Level 1 data sets.



A single file of *Global Area Coverage* (GAC) data represents:

- ~one full earth orbit.
- ~110 minutes.
- ~40 megabytes.
- ~15,000 scan lines.

One scan line is 409 IFOV's



# Dynamic Digital Microscopy

- Single 200X spot at a single depth of focus requires a resolution of (roughly) 1000 X 1000 pixels with three bytes of RGB color information per pixel -- 3MB
- Completely cover the slide requires a grid of about 50 X 70 such 200X spots
  - for two field depths, uncompressed file size of 21 GB
- One year's surgical pathology data at JHMI -- total uncompressed storage  $10^{16}$  bytes
  - *Equivalent to 1 million 10 Gbyte hard drives!*

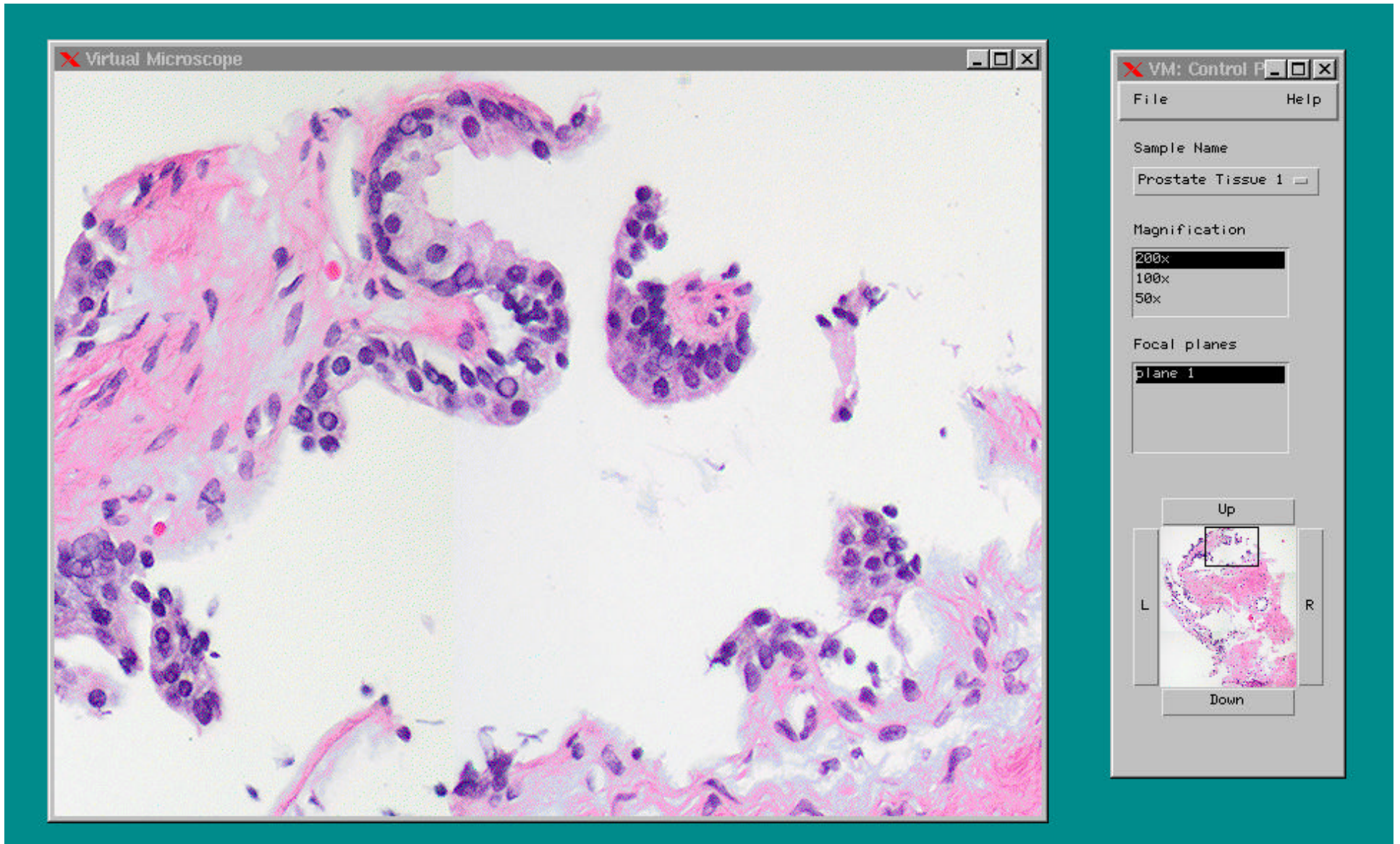
# Goals: Virtual Microscope

- World wide access to global collection of full digitized cases
  - no need to rely on slides available at local institution
  - compare specimens with earlier cases
- Research community can explore significance of morphology
  - compositing and 3-D reconstruction
  - quantitative immunohistochemistry analysis
  - run analysis programs on common data sets
- Links to relevant portions of digitized cases from medical information systems, electronic publications and textbooks.

# Training and Conference Environment

- Collection of institutions can carry out didactic conferences or collaborate on diagnosis
- Needle in haystack training
  - crucial skill often involves locating portions of case with interesting findings
- Multiple sites can
  - extends behavior of microscope - *independently* cruise through virtual pathology cases
    - users can change focus, magnification and move virtual stage
    - can track expert's examination of slide
  - capture whole case exploration process,
    - support combined medical record, demographic, pathology queries

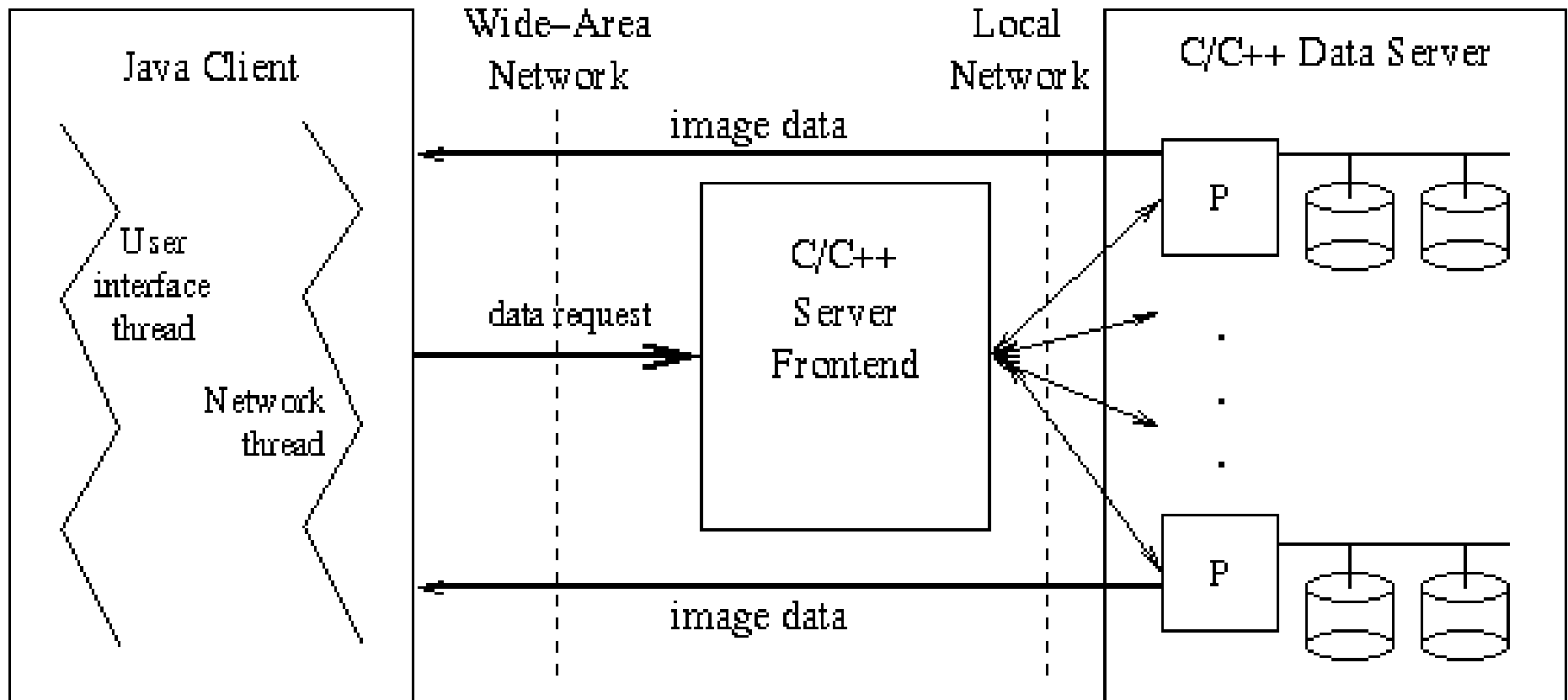
# Virtual Microscope Client



# Virtual Microscope Design

- Java based client
  - Client tested using Sun JDK and Microsoft J++
- Two part server (Currently runs on Windows, Solaris, AIX)
  - Front end -- accepts client queries, schedules queries and forwards to back end
  - Back end -- runs on multiple processors, retrieves and processes data from multiple disks

# Virtual Microscope Architecture



**Back end**

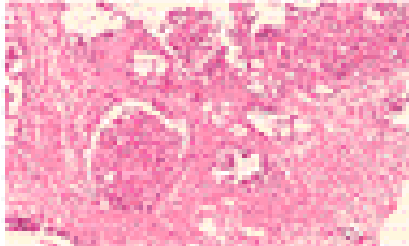


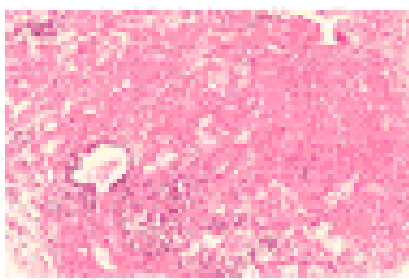
Virtual Microscope Test Slides - Microsoft Internet Explorer

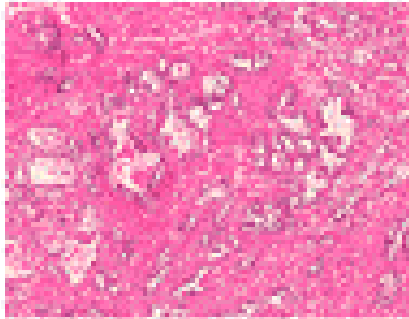
File Edit View Go Favorites Help

Address <http://www.cs.umd.edu/users/ftfang/vmicroscope/>

## Virtual Microscope Test Slides

 Breast tissue 1: (The diagnosis is atypical ductal epithelial hyperplasia). The clinical history is: The patient is a 63 year old woman with a prior history of total abdominal hysterectomy and bilateral salpingo-oophorectomy for well-differentiated endometrial carcinoma and benign breast cysts who is found to have focally clustered microcalcifications on routine mammography. An incisional biopsy is performed. The attached slide is a representative section of this patient's histology.

 Breast tissue 3: (The diagnosis is non-atypical ductal epithelial hyperplasia (papillomatous) arising in the setting of fibrocystic disease of the breast). The clinical history is: The patient is a 35 year old with no significant past medical history but a family history of breast and colon cancer in first degree relatives who noticed a rubbery firm mass in the right breast on routine self examination. The patient underwent incisional biopsy of a 2.5 cm. mass, histology showed fibrocystic changes (cysts, apocrine metaplasia, stromal fibrosis) in most sections. One section is submitted for your review.

 Prostate tissue 2: (The diagnosis is adenocarcinoma of the prostate, Gleason grade 3+3=6, and background benign prostatic hyperplasia). The clinical history is: The patient is a 68 year old man with a prior history of colorectal adenomas, diverticulitis for which he underwent a left hemicolectomy, and several basal cell carcinomas of the skin and a family history of prostate carcinoma who presented with complaint of urge incontinence, nocturia, frequency, diminished force of his urinary stream and occasional dysuria. A screening prostatic specific antigen assay performed three months prior to the current evaluation was 5.5 (ng/ml); subsequent repeat biopsies showed no tumor.

Internet zone

# Data Mining and On line Analytical Processing

- Definitions
- Examples from microbiology and infection control

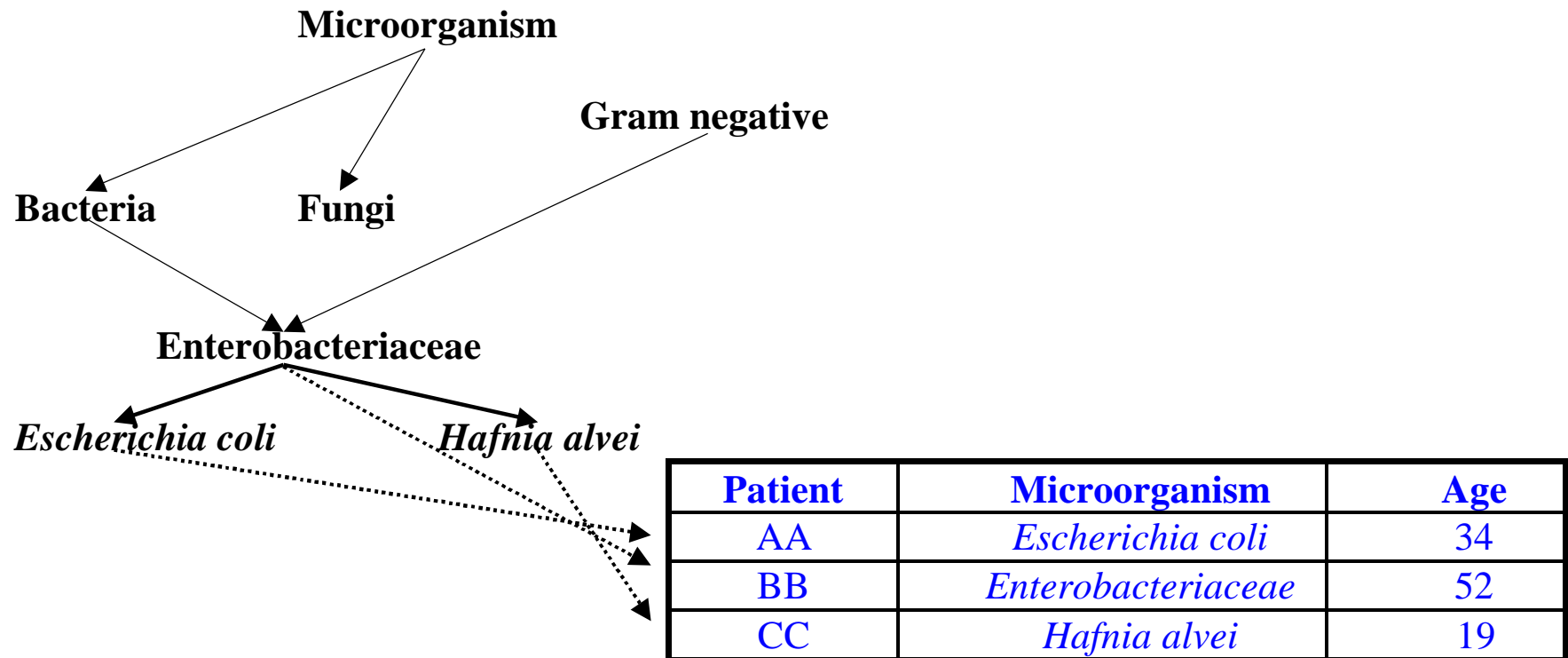
# Microbiology Queries

- For all classes of aerobic bacteria (ranging from general categories like gram negative rods to specific organisms like *Yersinia enterocolitica*) characterize resistance to all beta lactam antibiotics
- Compare effectiveness of different possible antibiotic protocols for neutropenic oncology patients
  - e.g. Trimethoprim/Sulfa + piperacillin v.s. Trimethoprim/Sulfa + piperacillin/tazobactam
- Screen for changes over time in antibiotic susceptibility

# Hierarchies and Databases

- Domain specific knowledge represented by hierarchies
- Graphical user interface or programmer API (currently two different versions) make it possible for the user to select portions of a hierarchy
- User can use a hierarchy to select a patient subset
  - user can then use other hierarchies to carry out data cube operation
  - e.g. for oncology patients, tabulate third generation cephalosporin antibiotic resistance in all classes of non-fermenter microorganisms

# Relationship between an Ontology and a Database



**DataBrowser**

Add to Query  
Clear Query  
Send Query

AND  
 OR

Start Date   
End Date

DATE

- ▶ Locations
- ▶ MICs
- ▼ Organisms
  - ▼ bacteria
    - ▼ bacterial
      - ▼ aerobe
        - ▶ gp aerobe
        - ▼ gn aerobe
          - ▼ nr aerobe
            - ▶ ebact
            - ▼ vibriona
              - ▶ AEROMONAS SPECIES
              - ▶ VIBRIO SPECIES
              - PLESEOMONAS SHIGELLOIDES

```
[[  
Location IN ('373CL','379CL','393CL')  
OR (  
Location IN ('ONC','ONC-2','OC-N3','GSONC','OC-S  
AND
```

**DataBrowser**

AND  
 OR

Start Date:   DATE  
 End Date:

- ▶ tetracyclines
- ▶ macrolides
- ▼ beta lactam
  - ▼ penicillins
    - ▶ penicillinase resistant pen
    - ▶ extended spectrum pen
    - ▼ penicillin inhibitor combination
      - ▼ pip/tazobactam
        - ▼ A\_P/T
          - ▶ P/T susceptible
          - ▶ P/T intermediat
          - ▶ P/T resisant
      - ▶ cephalosporins
      - carbepenems
      - monobactams

```

AND((((
  Field4 IN ('ebact','ESCH','ESCO','E0157'
  Field4 IN ('nonfermenter','PSEU','PSAE','
Field17 IN ('T_S_ENF_R','2+/38+'))))
OR ((
  Field4 IN ('PSEU','PSAE','PSAL','PSFL','F
Field17 IN ('T_S_PS_R','2+/38+'))))
AND((((
  Field4 IN ('ebact','ESCH','ESCO','E0157'
  Field4 IN ('nonfermenter','PSEU','PSAE','
Field21 IN ('P_T_ENF_R','64+/4+'))))
OR ((
  Field4 IN ('PSEU','PSAE','PSAL','PSFL','F
Field21 IN ('P_T_PS_R','64+/4+'))))
  
```

```

SELECT PNO, Location, Field1 AS Specimen_no, Field2 AS Isolate_no, Field3 AS Specimen_date,
Field4 AS Organism_code, Field5 AS AMP, Field6 AS TCR, Field7 AS PIP, Field8 AS CFZ, Field9 AS
CFR, Field10 AS CTZ, Field11 AS CXM, Field12 AS GEN, Field13 AS TOB, Field14 AS AMI,
Field15 AS TET, Field16 AS SUL, Field17 AS T_S, Field18 AS CIP, Field19 AS FUR, Field20 AS T_C,
Field21 AS P_T, Field22 AS TIC, Field23 AS MET, Field24 AS PEN, Field25 AS VAN, Field26 AS CLN,
Field27 AS ERY, Field28 AS OXA, Field29 AS XXX, Field30 AS YYY, Field31 AS ZZZ,
Field32 AS CZO, Field33 AS CFU
FROM Moss
WHERE ((
Location IN ('373CL','379CL','393CL'))
OR (
Location IN ('ONC','ONC-2','OC-N3','GSONC','OC-S3','OC-A3','372CL','392CL'))))
AND((
Field4 IN ('nr aerobe','ebact','ESCH','ESCO','EO157','ESFE','ESHE','ESVU','ENTE','ENAE','ENAG'
,'ENAM','ENAS','ENCL','ENGE','ENSA','ENTA','CITR','CIAM','CIBR','CIDI','CIFR','CISE','CIWE'
,'CIYO','KLEB','KLOR','KLOX','KLOZ','KLPN','KLRH','KLUY','KLAS','KLCR','EDWA','EDHO'
,'EDTA','ERWI','HAFN','HAAL','morganella','MOMO','PROT','PRMI','PRPE','PRVU','SPP','PROV'
,'PRAL','PRRE','PRRU','PRST','SALM','SAAG','SAAR','SABA','SABR','SACS','SACH','SADN'
,'SADE','SAEN','SAGA','SAGB','SAGC','SAGD','SAEG','SAHE','SAIN','SALO','SAMA','SANE'
,'SAOH','SAOR','SAPA','SAST','SASE','SATH','SATYP','SATYM','SERR','SEFI','SEFO','SELI'
,'SEMA','SEOD1','SEOD2','SEPL','SERU','SHIG','SHBO','SHDY','SHFL','SHSO','YERS','YEEN'
,'YEFR','YEIN','YEPE','YEPS','YERU','CEDE','CENE','CELA','BUAQ','EWAM','KOTR','LEAD'
,'leminorella','LEGR','LERI','leminorella sp','MOWI','RAAQ','TAPT','YORE','vibriona','AERO'
,'AECA','AEHY','AESO','AEVE','VIBR','VIAL','VICH','VIFL','VIFU','VIHO','VIME','VIMI','VIPA'
,'VIVU','PLSH','vibrionaceae','nonfermenter','PSEU','PSAE','PSAL','PSFL','PSMAL','PSME','PSPI'
,'PSPSA','PSPSM','PSPUT','PSST','burkholderia','BUCE','PSCE','BUGL','COMSP','COAC','PSAC'
,'COTE','PSTE','FLAV','FLGL','FLII','FLME','FLOD','ALCA','ALDE','ALFA','ALOD','ALXY',

```



Cube Interface

nr\_aerobe [S | R (total)]

ms  
bacteria  
- bacteria1  
 - aerobe  
 + gp\_aerobe  
 - gn\_aerobe  
 - nr\_aerobe  
 + ebact  
 + vibriona  
 + nonfermenter  
 + nr\_fast  
 + nc\_aerobe  
 + spiro  
 + mycoplasma  
 + afb  
 + anaerobe

antibiotics  
+ aminoglycosides [81.01 6.92 12.06 (2412)]  
+ quinolones [79.18 8.43 12.39 (759)]  
+ tetracyclines [14.23 37.38 48.39 (808)]  
+ macrolides [32.0 0.0 68.0 (25)]  
+ glycopeptides [82.61 0.0 17.39 (23)]  
- beta lactam [42.1 14.15 43.75 (5772)]  
 - penicillins [32.61 15.88 51.51 (2551)]  
 + penicillinase resistant pen [52.0 0.0 48.0 (25)]  
 + extended spectrum pen [32.7 15.94 51.37 (2416)]  
 + penicillin inhibitor combination [26.36 18.18 55.46 (11)]  
 + cephalosporins [49.61 12.79 37.6 (3221)]  
+ Trim/Sulfa [62.21 8.67 29.12 (807)]

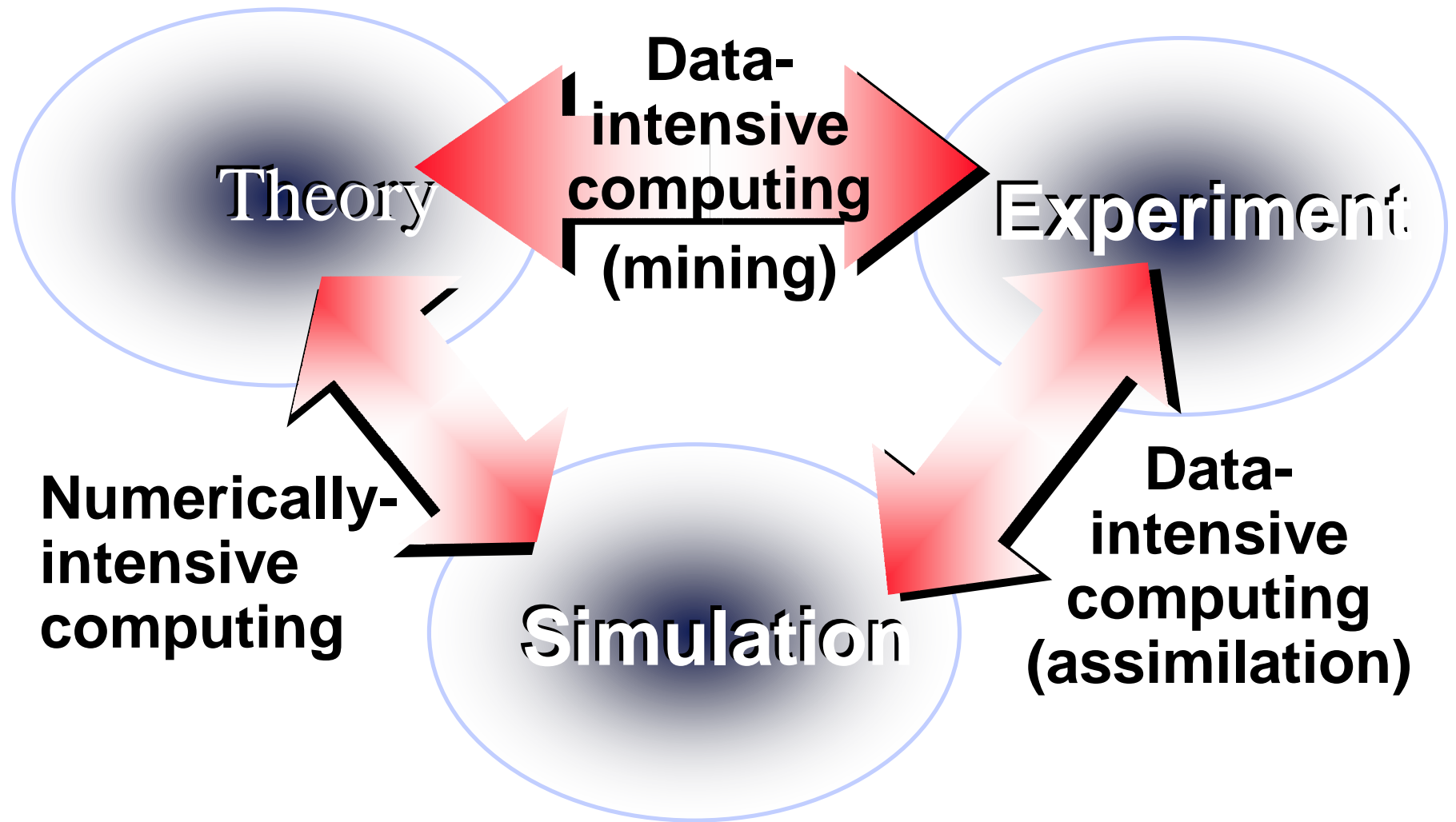
Calculate Distribution

Evaluate Cube

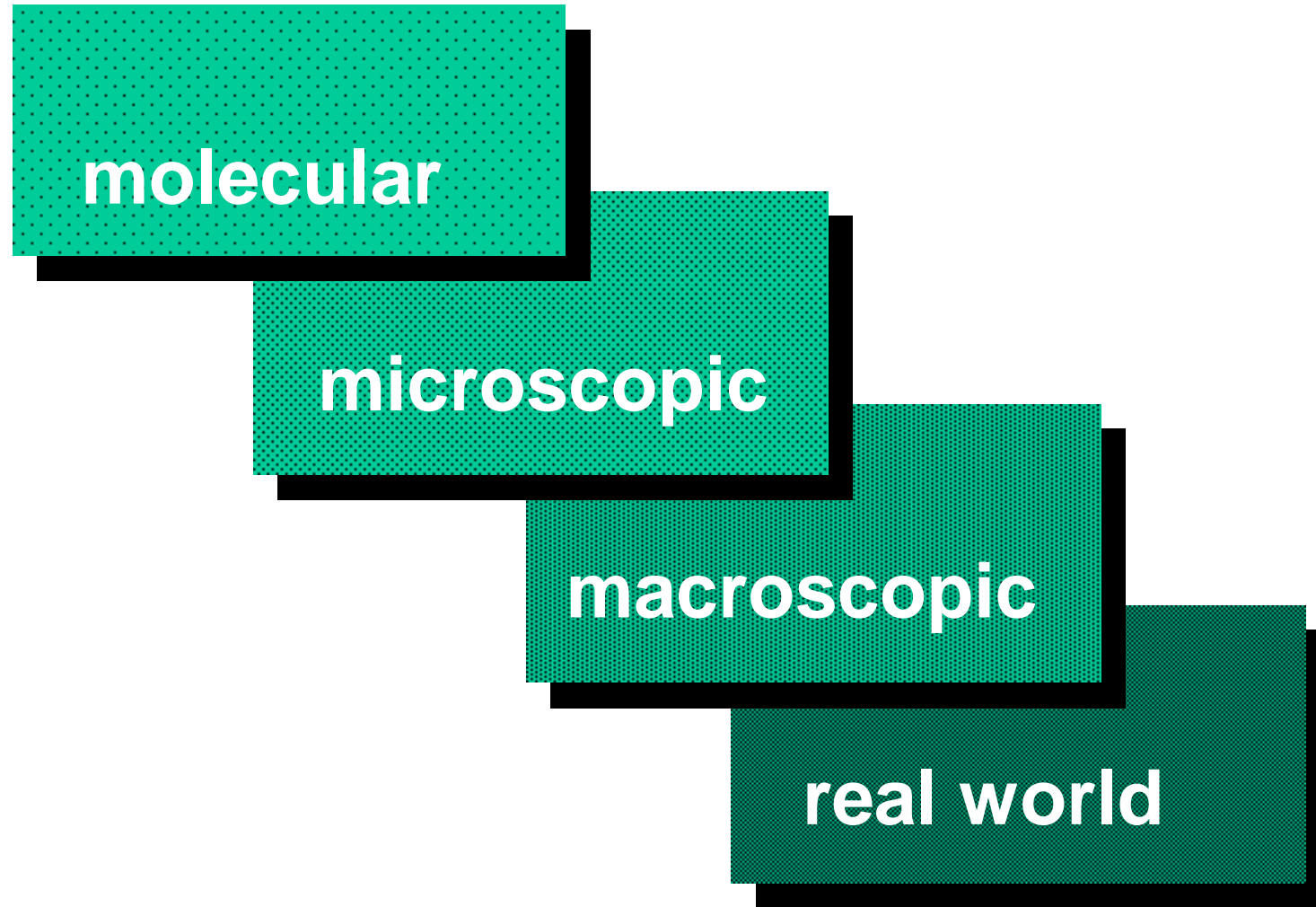
Calculate Distribution

EXIT

# Simulation is Synergistic



# The Next Challenge: Integration Across Scales



# Roadmap

- Ambitious Simulations, Sensor Data Processing and Data Mining
- Computer Architecture Made Ridiculously Simple
- High Performance Databases and Systems Software

# Parallelism

- Parallelism key to design of high performance processors
  - at any time, microprocessor carries out simultaneous work on 10-100 calculations
  - several assembly lines in operation at once
    - assembly lines for adds, for multiplies, for logical operations

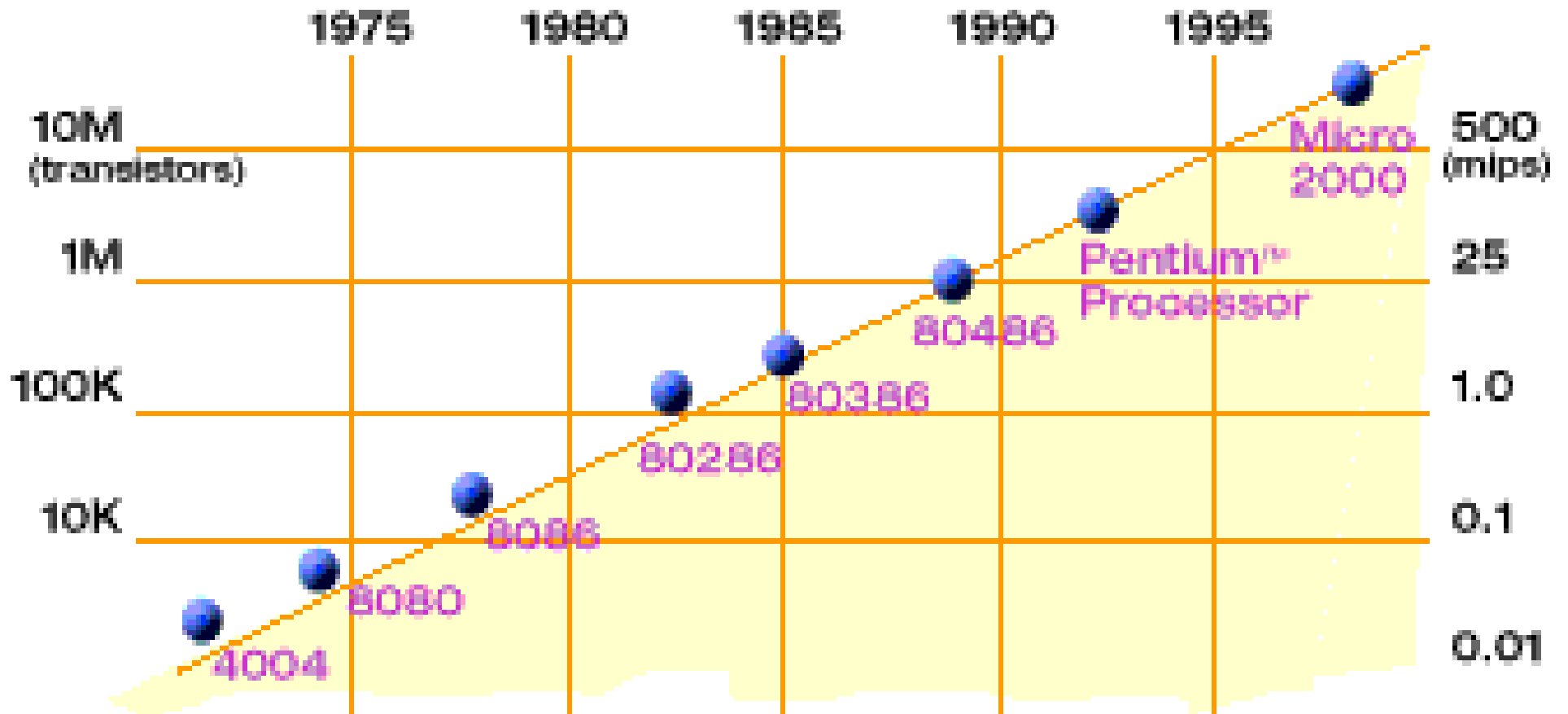
# Parallelism: Many Fast Processors Join Forces

- A large number of computers join forces to solve a single problem
  - chop the data and computations into pieces
  - farm a piece out to each processor
  - collect results and announce solution
- Challenge
  - chopping a problem judiciously can be tricky
    - pieces need to be similar in size
    - pieces are rarely independent
      - choose chopping strategy that minimizes costs of exchanging information

# Dramatic Improvements in Microprocessor Performance

- Moore's Law:
  - Since 1971 microprocessor performance, number of transistors
    - roughly doubles every 18-24 months
- Reductions in transistor size allow:
  - *Increase in parallelism* - large number of transistors make possible computational “assembly lines”
  - *Increase in speed* - speed of light limits speed that can be attained by a large machine

# Moore's Law

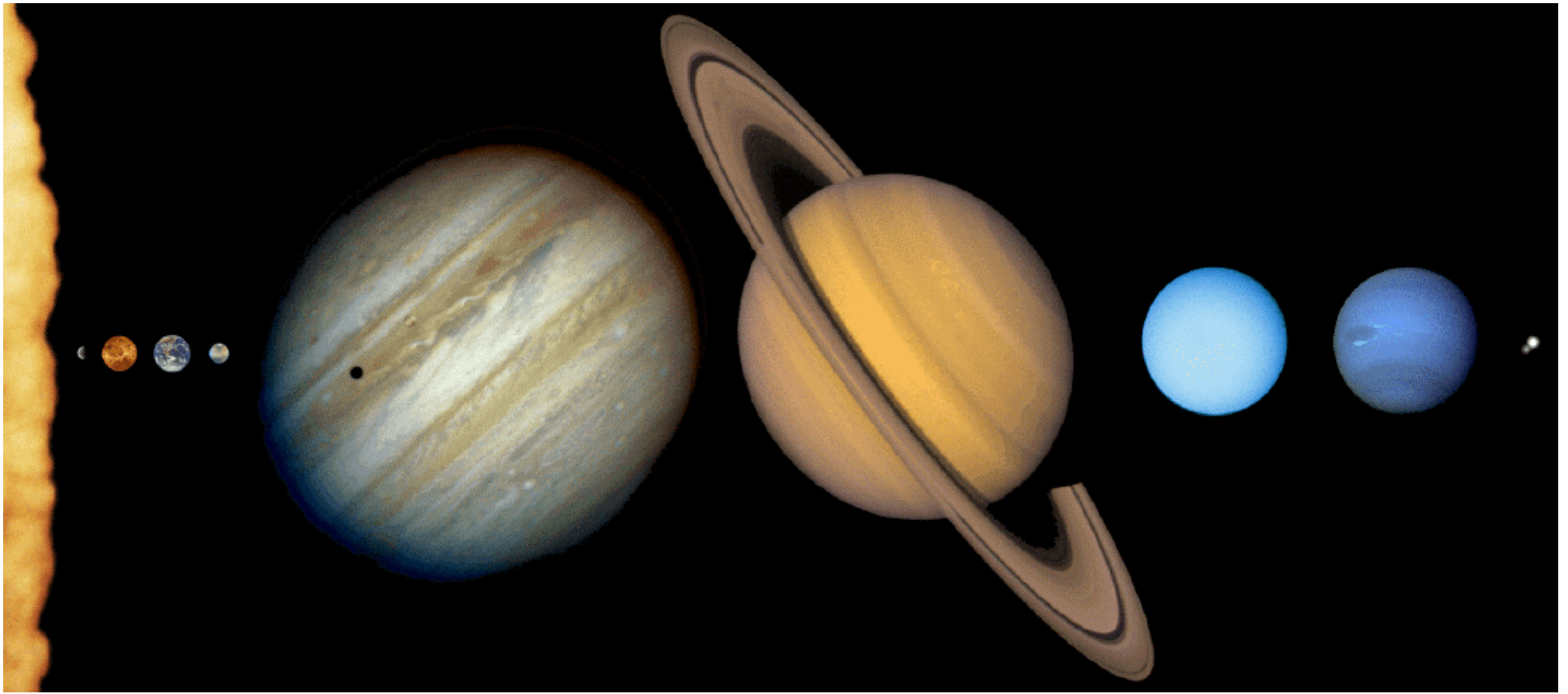


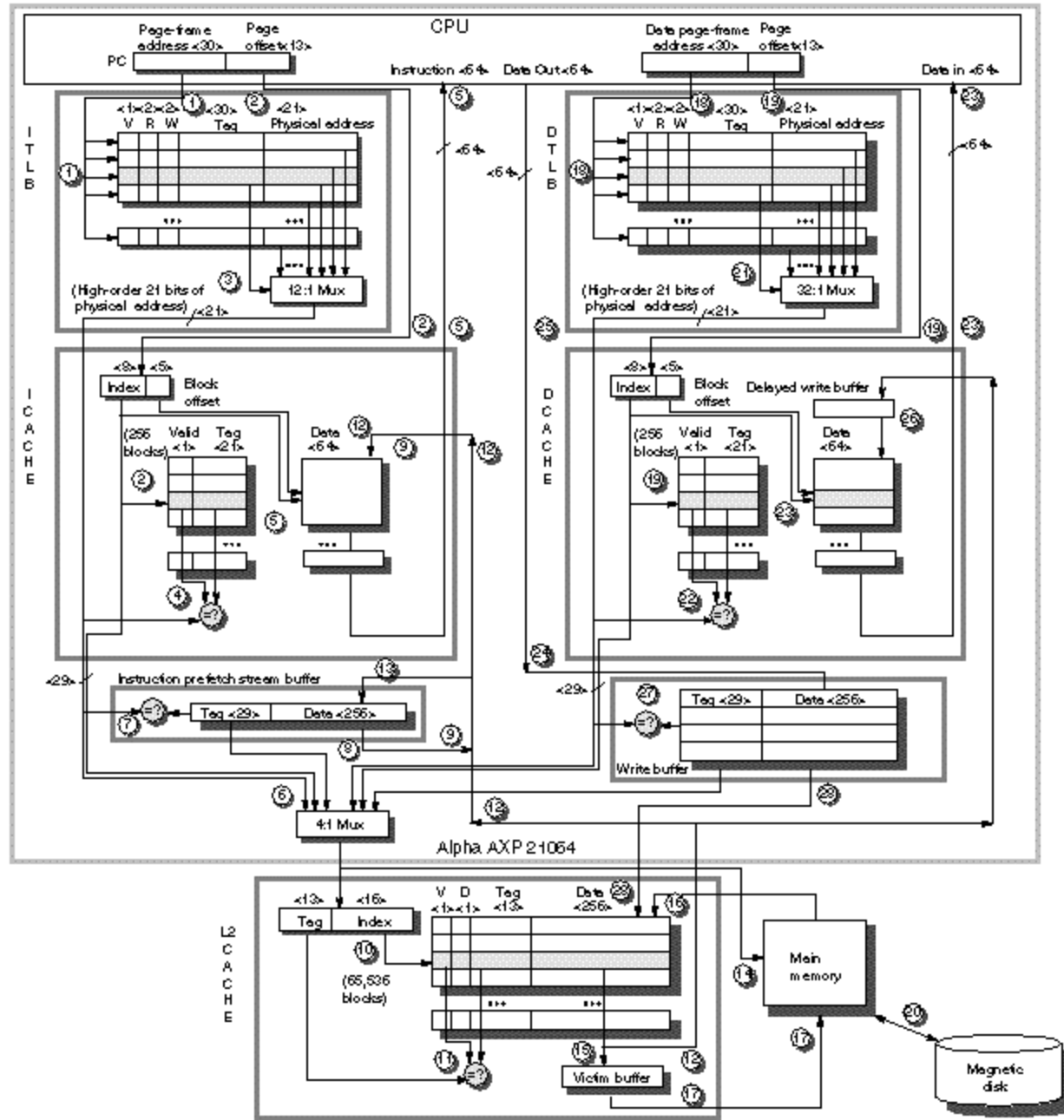


# Memory Hierarchy

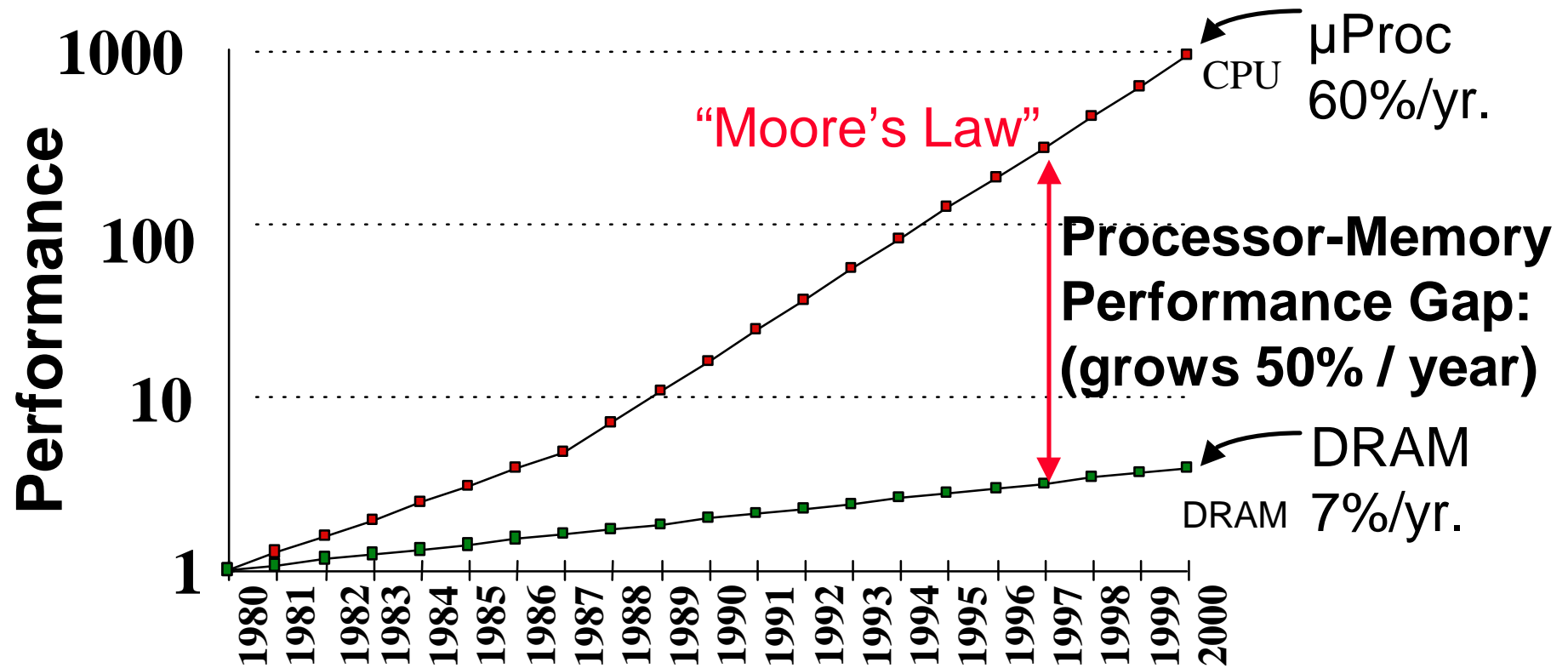
- Registers, cache, RAM, disk, tape are all types of memory
- Fast memory is expensive, slow memory is cheap
  - Cost:
    - registers > cache > RAM > disk > tape
  - Capacity
    - registers < cache < RAM < disk < tape
    - $10^4 :: 10^6 :: 10^9 :: 10^{11} :: 10^{13}$
  - Time needed to get data:
    - registers < cache < RAM < disk < tape
    - $1 :: 2 :: 50 :: 10^6 :: 10^{10}$
    - *JHH to Fell's point v.s. Earth to Saturn*

# Registers, Disks, Tapes and the Solar System





# Processor Memory Performance Gap



# Multiprocessor Performance Goals

- Late 1980's through late 1990's performance goal was to achieve a rate of  $10^{12}$  floating point operations per second “*teraflop*” on real applications
- Major research effort
  - Several billion dollars funding
  - Sponsored by National Science Foundation, Advanced Research Projects Agency, Department of Defense, Department of Energy, National Institutes of Health

Gflop/s

1000

750

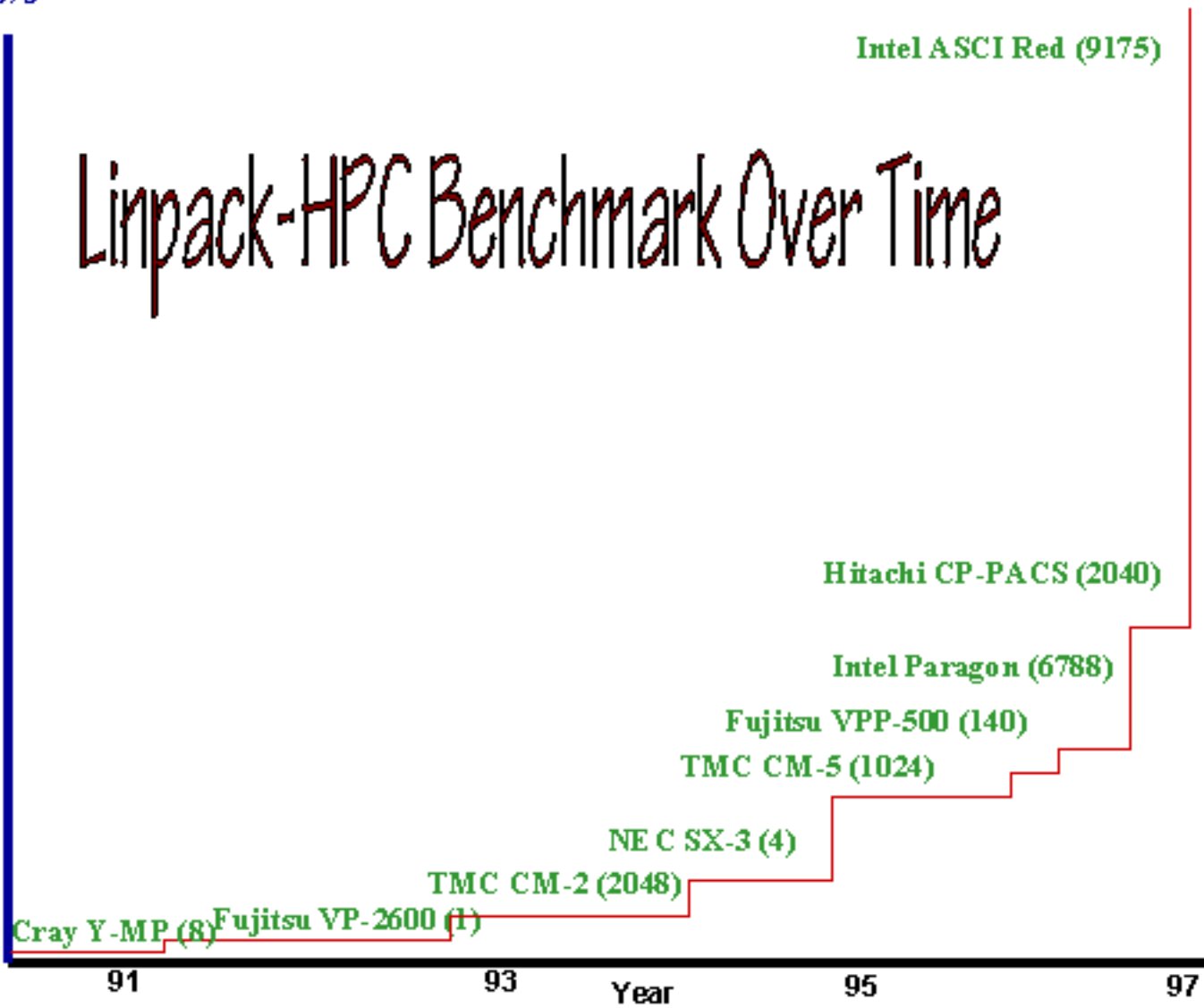
500

250

100

0

# Linpac-HPC Benchmark Over Time



# Roadmap

- Ambitious Simulations, Sensor Data Processing and Data Mining
- Computer Architecture Made Ridiculously Simple
- High Performance Databases and Systems Software

# High Performance on Real Problems

- Programming tools, compilers, databases optimize performance
  - Intraprocessor parallelism
    - keep “assembly lines” busy
  - Multicomputer parallelism
    - break portions of problem into pieces
    - minimize movement of data between computers
  - Memory hierarchy
    - minimize cost of moving data between registers, cache, main memory, disk, tape



# Active Data Repository

- Generalized parallel data server
- Target many application areas
  - Data server for the virtual microscope
  - Generation of data products from low level satellite sensor data
  - Compute engine for Data Cube, data mining, on line analytical processing
  - Post-processing, analysis and exploration of data generated by large scientific simulations
    - Chesapeake Bay simulation, Oil reservoir simulation
- Future applications
  - Pathology and radiology: 3-D image reconstruction, interactive data exploration

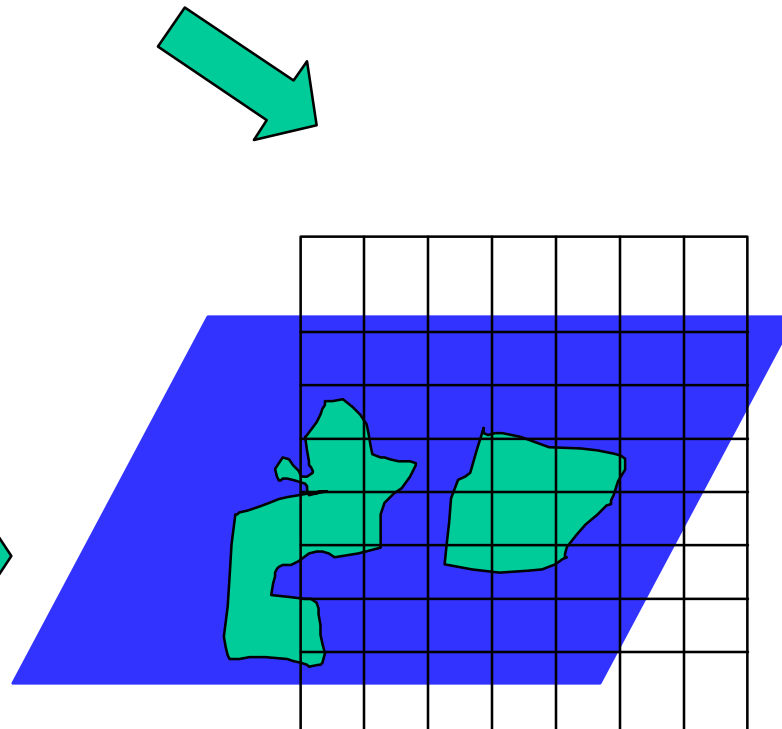
# Active Data Repository

- System software to carry out coordinated queries, retrieval and processing in very large persistent data structures
  - Integrate and overlap a wide range of user-defined operations, in particular, order-independent operations with data retrieval functions
  - Support optimized associative access to multiresolution and irregular persistent data structures
    - Targeted operations include projections, interpolations, range queries, data postprocessing and compositing
    - Proxy server architecture for servicing large numbers of remote users

# Example Projection Query

Output grid onto  
which projection  
is carried out

Specify portion of raw  
sensor data corresponding  
to some search criterion



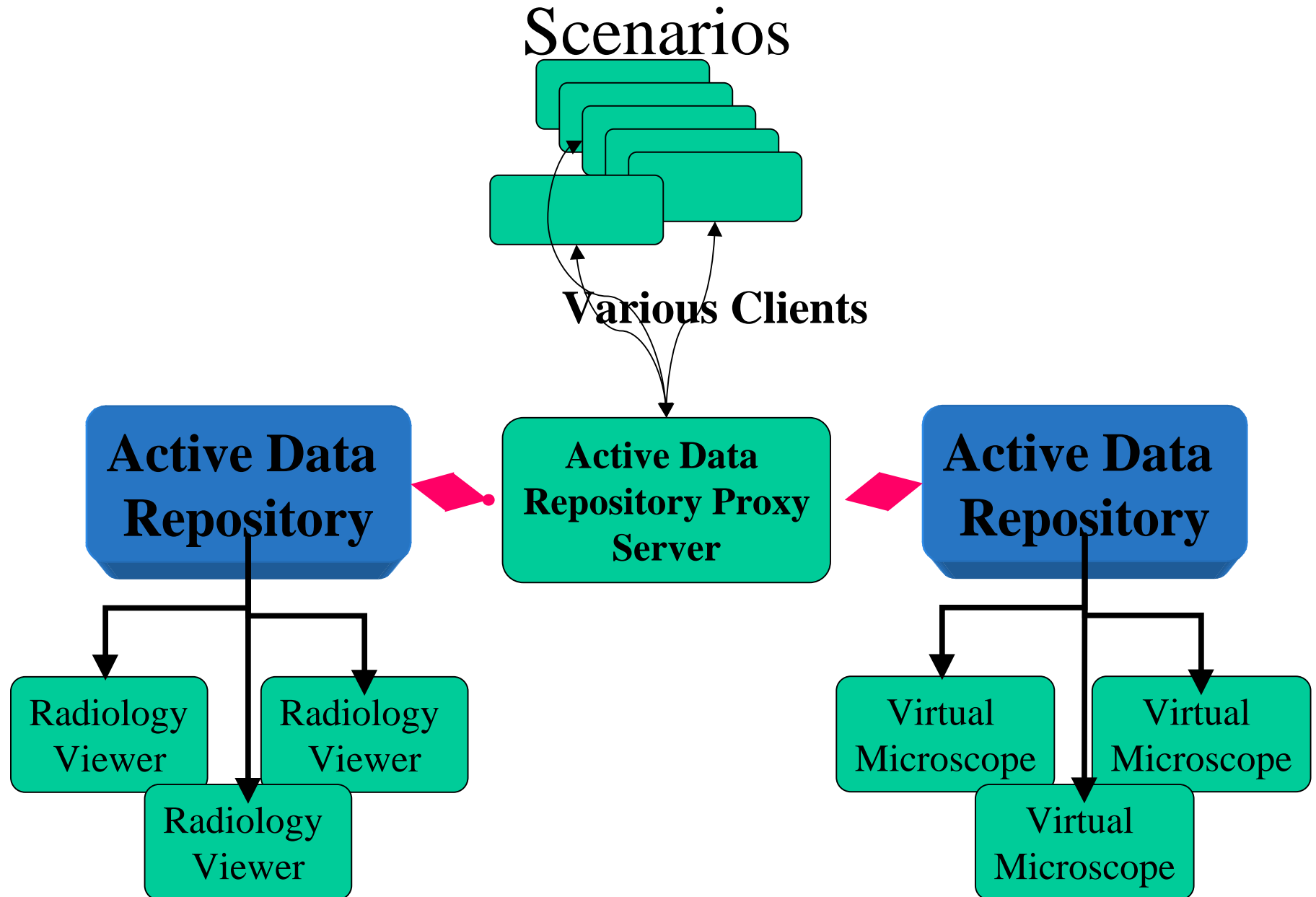
# Examples of Services

- Highly optimized software to perform computations, and to move data to and from disks and to and from processors
  - software exploits parallelism by overlapping computing, communicating data between processors and moving data to and from disks
- Clustering/Declustering software balances two competing tasks
  - spread dataset between disks to maximize parallelism
  - cluster data to reduce cost of moving data from each disk
  - Results in Ph.D. thesis, publications of Bongki Moon (started at U. Arizona last fall)

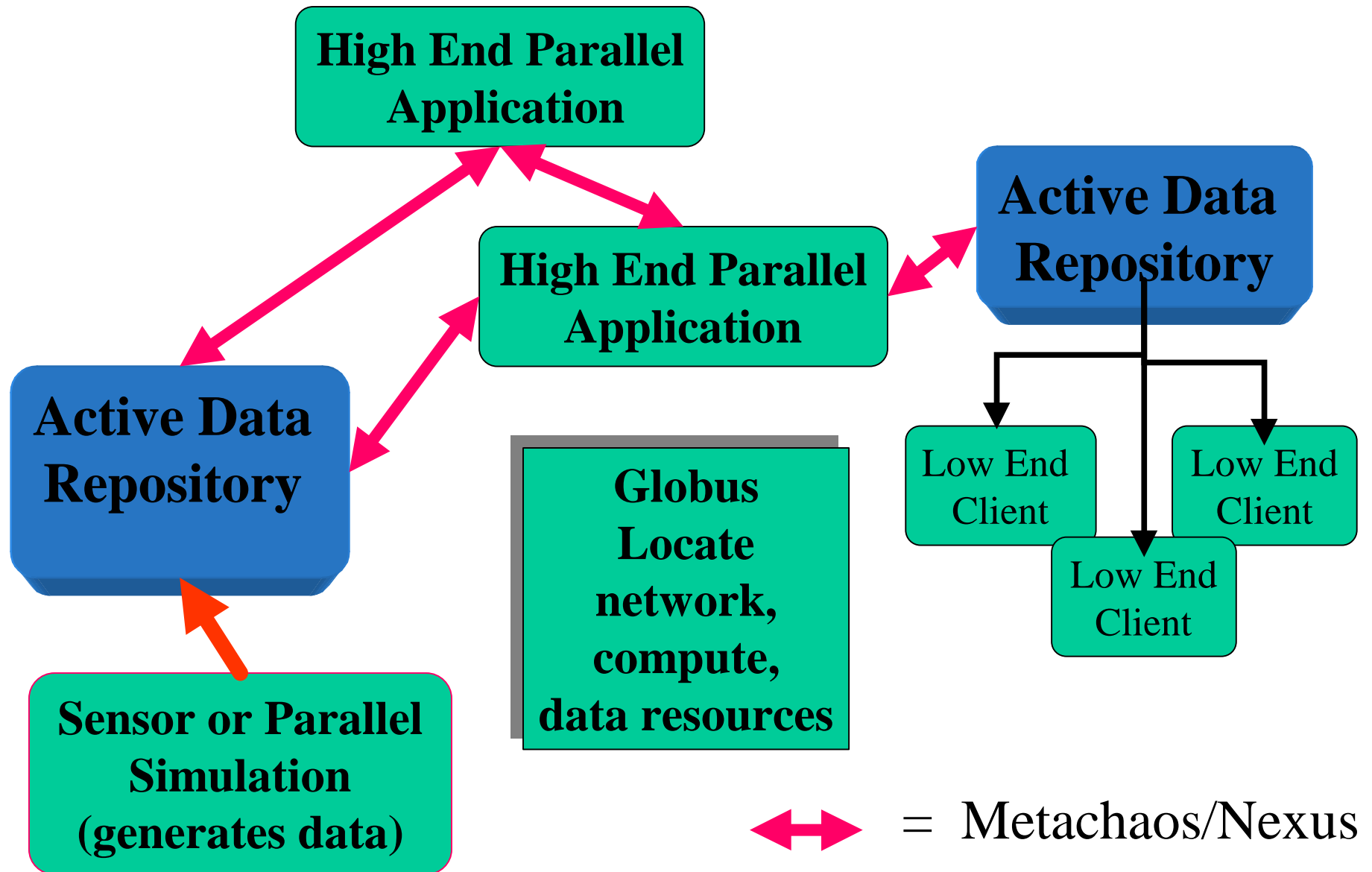
# Examples of Services

- Software that links a datastructure's coordinate system to particular processors and particular disks
  - e.g. virtual microscope -- stage position, magnification and focus
  - satellite sensor processing -- set of latitudes/longitudes
  - Data may be sparse or dense
    - example of sparse data -- virtual microscope -- not all data is always available at every power
- Software that supports mappings between different datastructures

# Networked Based ADR Computing



# Network Based ADR Scenarios



# Conclusion

- Dramatic advances in computers are enabling new classes of applications in science, engineering and medicine
- Many seemingly different applications have common computational characteristics
- Tools can be developed that optimize performance of broad applications classes



# Collaborators

## University of Maryland

Asmara Afework

Mike Beynon

Charlie Chang

John Davis

Renato Ferreira

Bongki Moon

Kilian Stoffel

Alan Sussman

Mustafa Uysal

## Johns Hopkins Medical Institutions

Angelo Demarzo

Jim Dick

Bill Merz

Robert Miller

Jerry Rottman

Mark Silberman

Kilian Stoffel

Merwyn Taylor