

Information Technology for the Life Sciences

Dr. Joseph M. Jasinski
IBM Research



*Thomas J. Watson Research Center
PO Box 218
Yorktown Heights, NY 10598*

September, 2001

IBM Research Worldwide



Innovations



1944: Mark I



1948: SSEC



1956: RAMAC

Typical mathematical formula:
 $D = B^2 - 4AC$
Equivalent FORTRAN statement:
 $D = B**2 - 4*A*C$

1957: FORTRAN



1966: One-Device Memory Cell



1967: Fractals



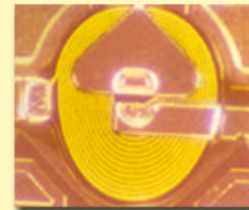
1970: Relational Database



1971: Speech Recognition



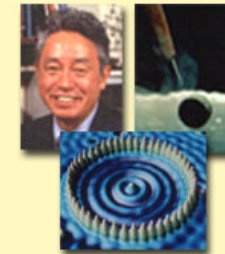
1973: Winchester Disk



1979: Thin Film Recording Heads



1980: RISC



Nobel Prizes



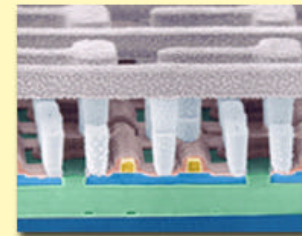
1994: SiGe



1993: RS/6000 SP
1996,97: Deep Blue



1997: Copper Interconnect Wiring



1998: Silicon-on-Insulator



1998: Microdrive

What Has Changed

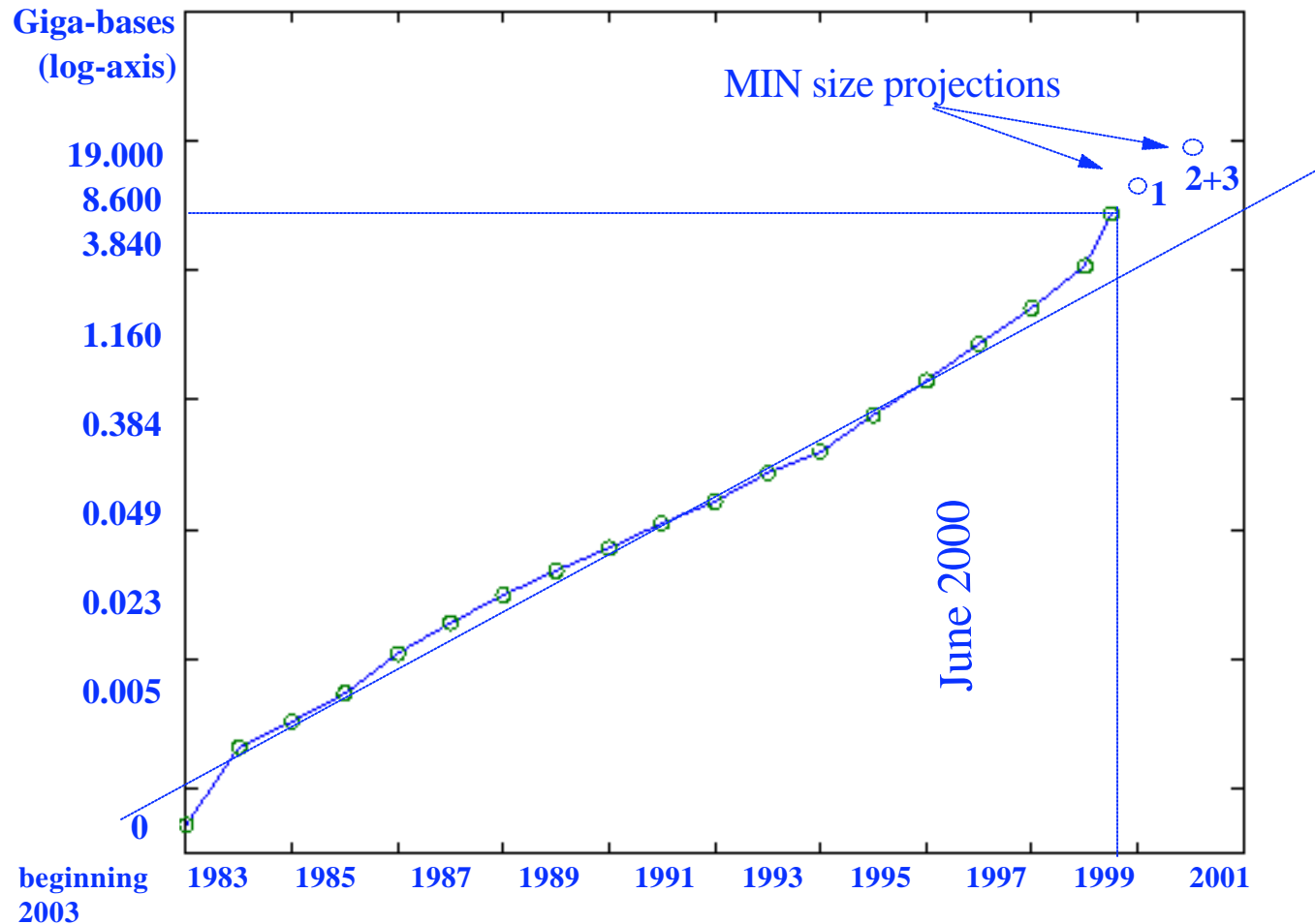
Recent developments in laboratory automation combined with powerful computational and algorithmic capabilities have created a bioinformatics industry.

Genomic and proteomic data will be readily available in massive amounts. This will revolutionize agriscience, drug discovery, biotechnology and ultimately human healthcare.

This revolution is computationally and data intensive.



GenBank ^(TM) Size (bases)

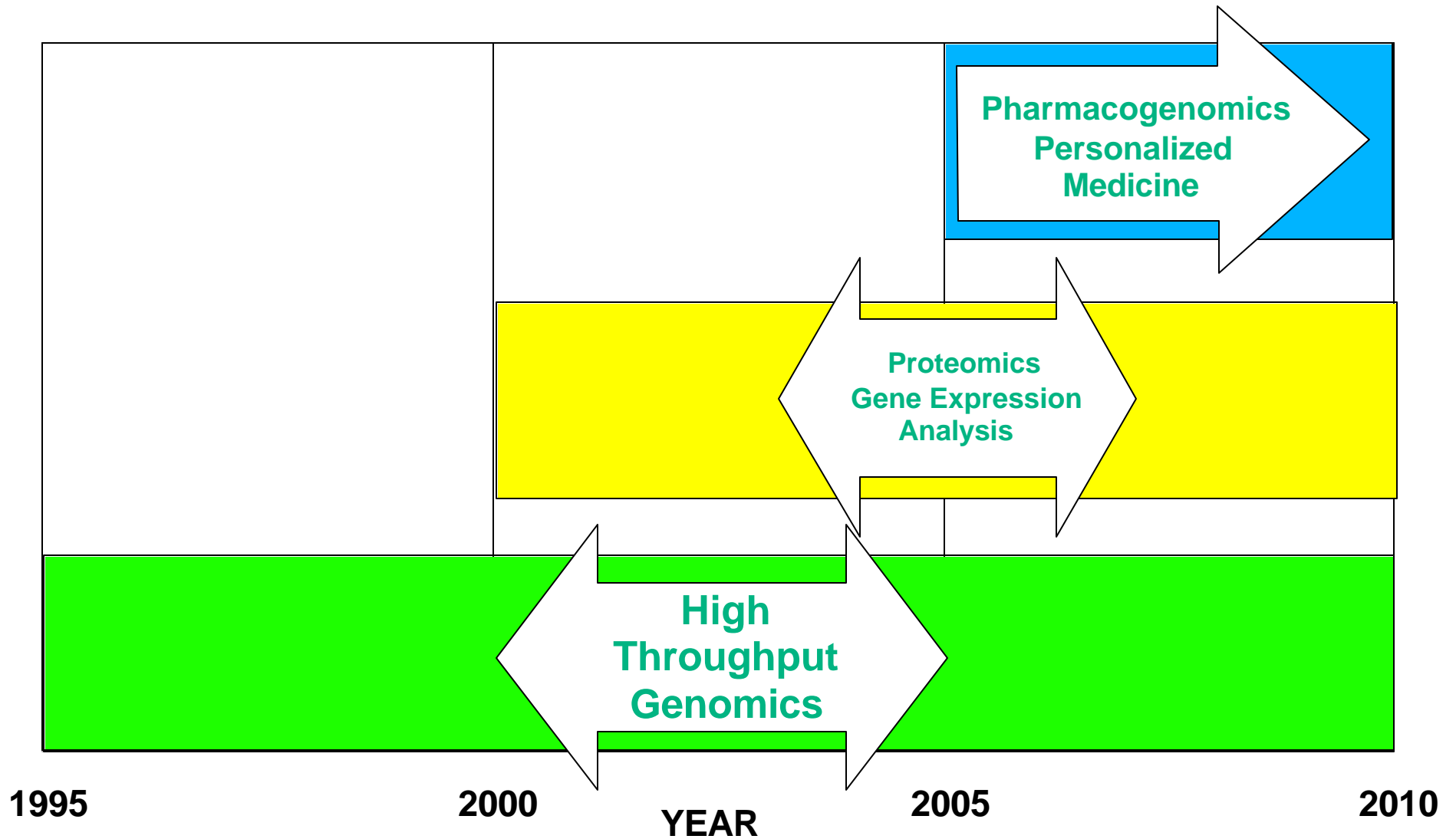


Major addition of year

- 1 = Homo sapiens (+3.5 Gb)
- 2 = Mus musculus (+ 3.5 Gb)
- 3 = Felis catus (+ ~3.0 Gb)

Size of end-dbase = # bases x 10 bytes

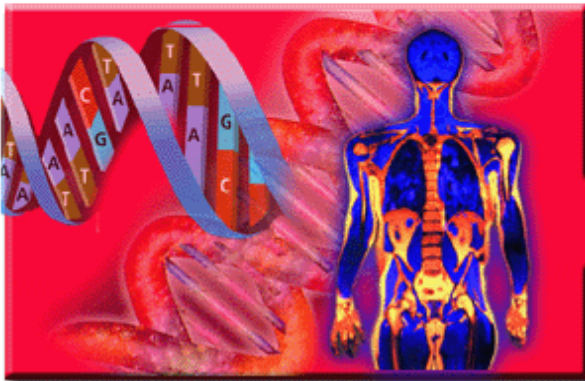
Towards Personalized Molecular Medicine



Convergence Creates New Models

Scientific discovery
New drugs and treatments
Revolution in healthcare

Life Sciences



**Information
Technology**



IT enablers for Life Sciences

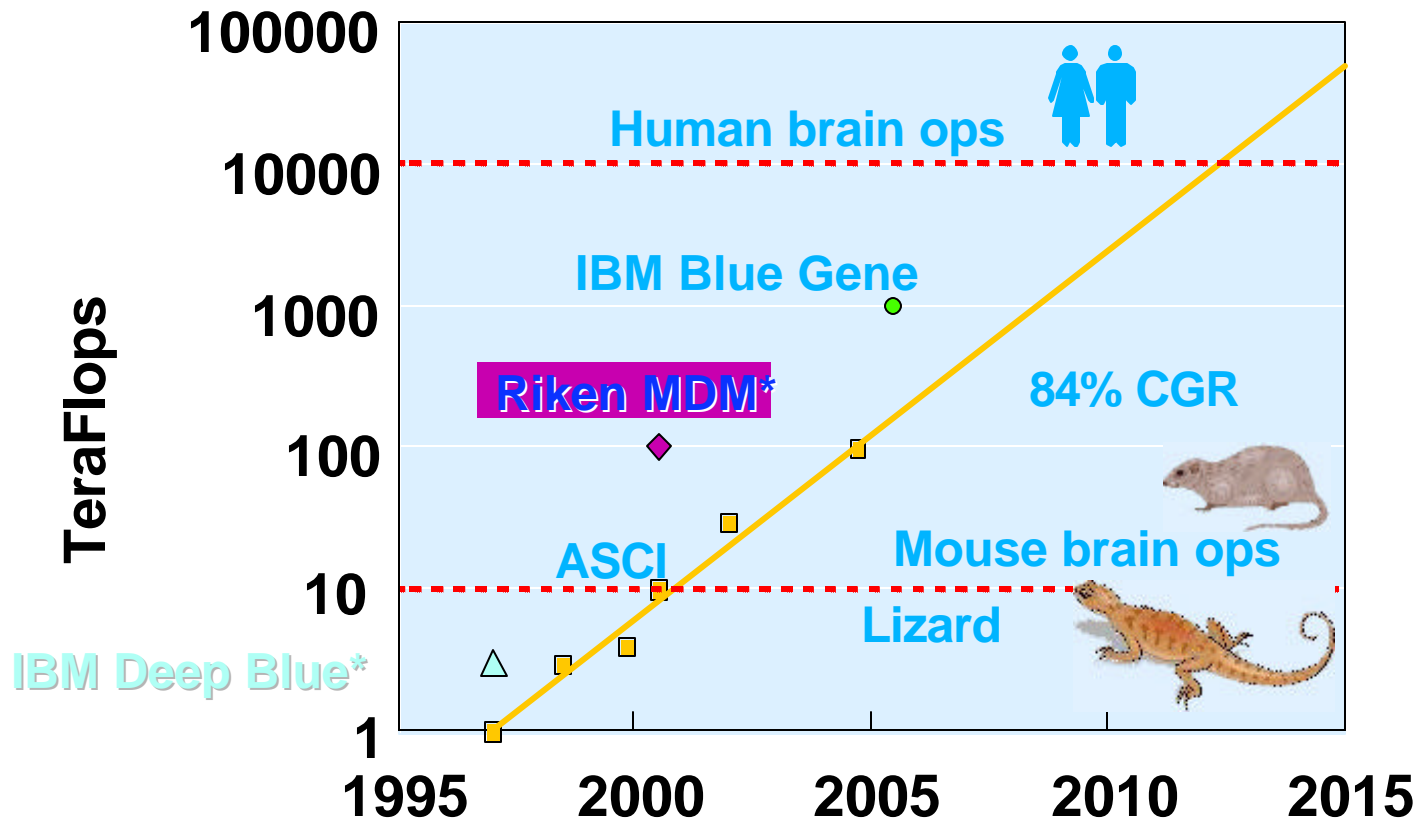
- **Scaling: Large S&TC infrastructures for complex environments**
- **Integration: Data integration and knowledge management solutions**
- **e-Business: New models for web-based scientific and healthcare business**



IBM in the Life Sciences

- Leader in supercomputing and high performance storage
 - Data integration and scalable database
 - Leader in KM solutions
 - Leader in e-business, security and privacy
 - + \$200M investment in IBM Life Sciences Solutions
 - + \$100M investments in Research
-
-

Supercomputing Roadmap

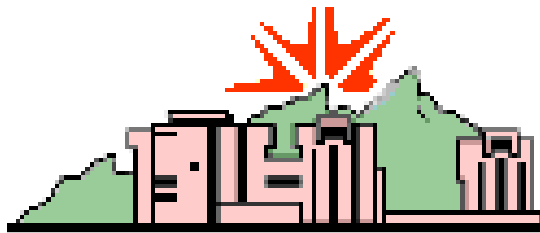


Source: ASCI Roadmap www.llnl.gov/asci, IBM

Brain ops/sec: Kurzweil 1999, [The Age of Spiritual Machines](#)

Moravec 1998, www.transhumanist.com/volume1/moravec.htm

IBM Terascale Systems

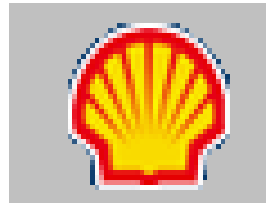


National Center for Atmospheric Research (NCAR)



NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

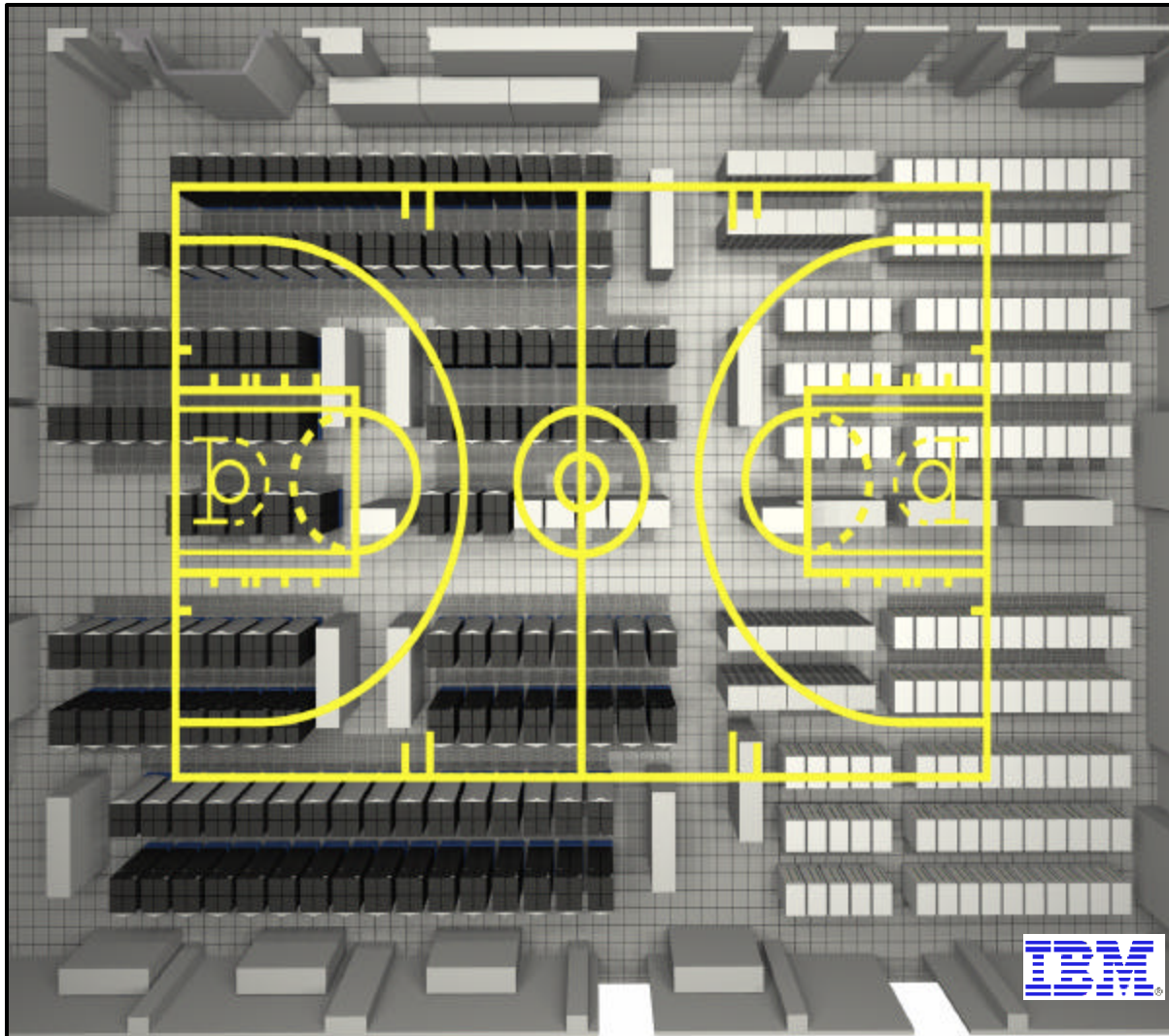
Advancing Computational Science of Scale—
Producing Real Results



Oak Ridge National Laboratory
Bringing Science to Life



This summer at LLNL, IBM's "ASCI White" will become the world's most capable supercomputer (surpassing IBM's Blue Pacific, also at LLNL)



IBM's ASCI White

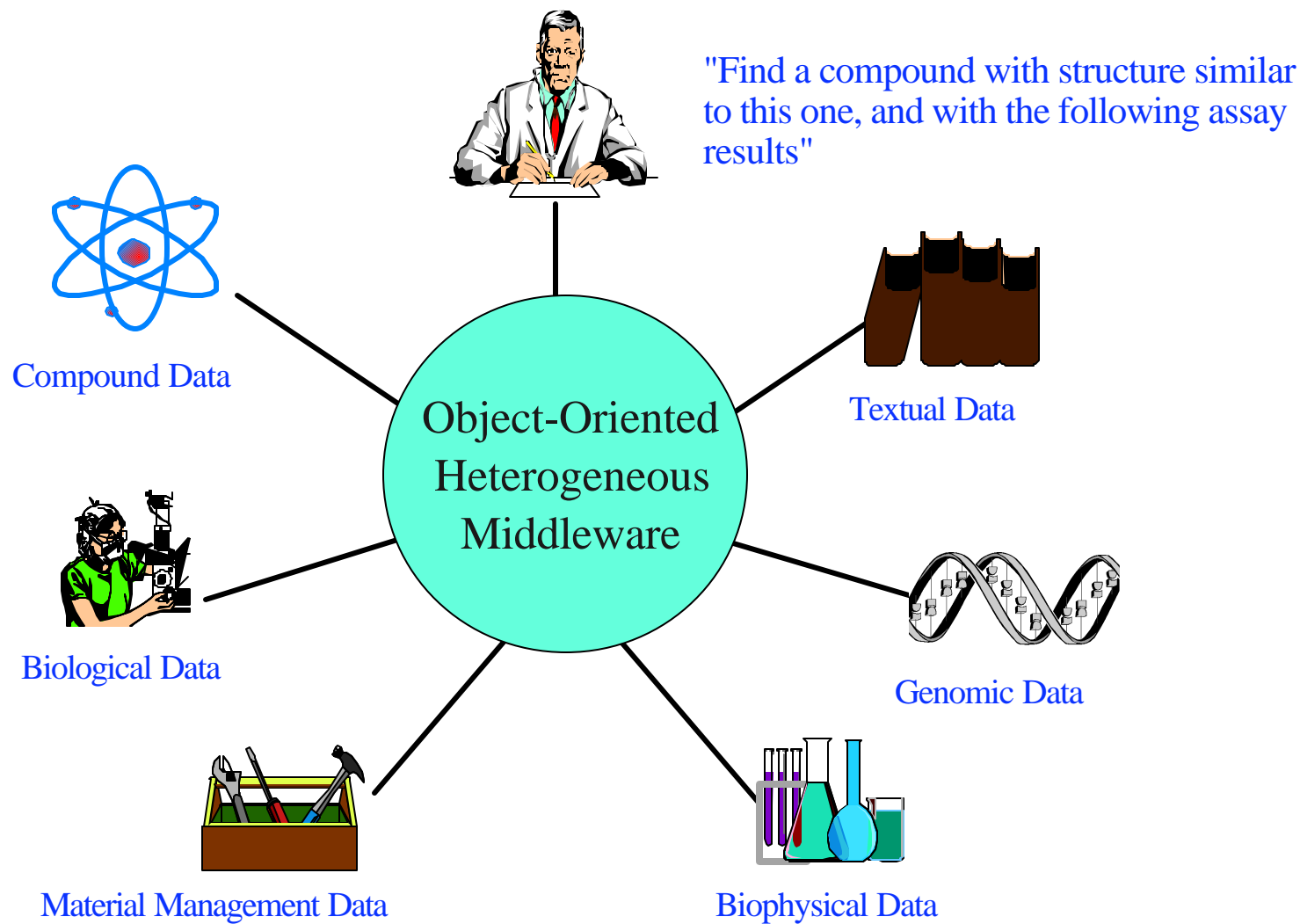
- 8,192 copper processors
- Six trillion bytes of memory
- More than 160 trillion bytes of disk storage capacity
- Will draw 6.2 MW for cooling and power
- Will be delivered from Poughkeepsie NY to LLNL in 28 tractor-trailer trucks

This level of computing has never been achieved anywhere.

-- Dr. David Cooper, LLNL's AD for Computation

This is the second time IBM has exceeded contract requirements by achieving a peak rate of 12.3 trillion floating-point operations per second.

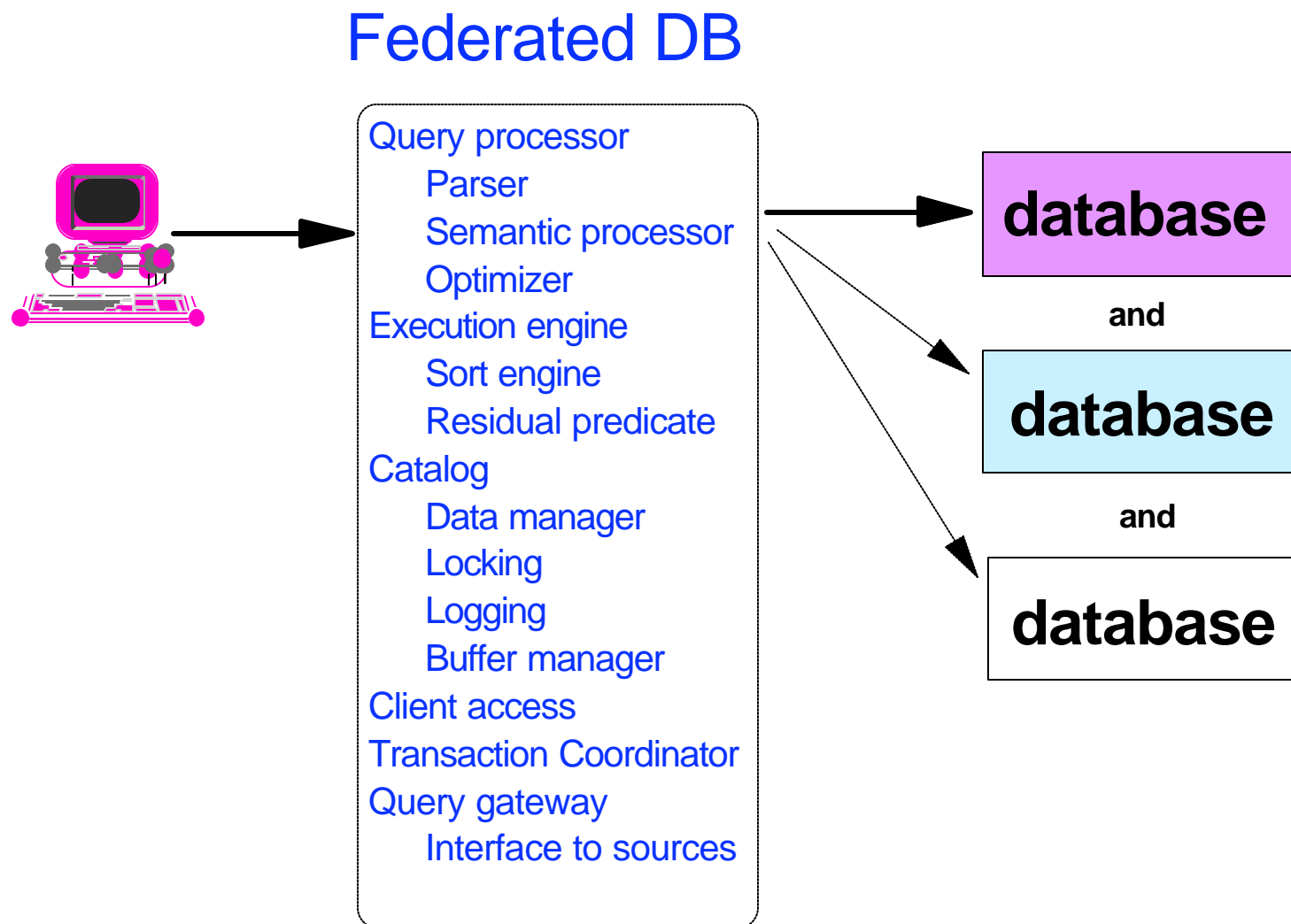
The Challenge: Access to Heterogeneous Data



DiscoveryLink

- **Provide a single "virtual database" to applications**
 - Appears to be one data source
 - Supports a high level query language
 - **No perturbation of existing data, sources**
 - **Exploit capabilities of existing sources**
 - To search for and manipulate data
 - Lose no functionality
 - **Integrate data from different data sources**
 - Diverse types of data
 - Diverse sources
 - One query can combine data from multiple sources
-
-

Federated database needs a full database engine!



Computational Biology at IBM Research



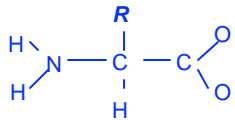
*Thomas J. Watson Research Center
O Box 218
Yorktown Heights, NY 10598*

Computational Biology Center Projects

- **Bioinformatics Algorithms**
- **Functional Genomics and Modelling**
- **Structural Biology**
- **Protein Dynamics and Blue Gene**
- **Data Management and Integration (Discovery Link)**

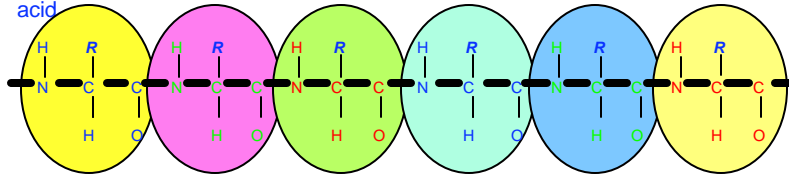
What is a Protein?

Sequence



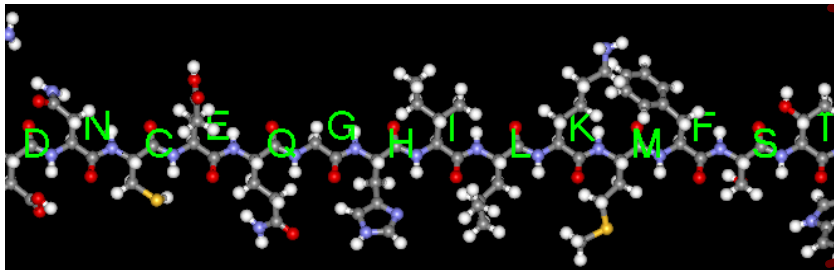
There are 20 natural amino acids with different physicochemical properties, such as: shape, volume, flexibility, hydrophobic, hydrophilic, charge

Amino acid



"Beads on a string"

Structure



Function

Structural: keratin (skin, hair, nail), collagen (tendon), fibrin (clot)

Motive: actomyosin (muscle)

Transport: Hemoglobin (blood)

FIBRINOGEN GAMMA CHAIN

QIHDITGKDCQDIANKGAKQSGLYFIKPLKANQQFLVYCEIDG
SGNGWTVFQKRLDGSVDFKKNWIQYKEGFGHLSPTGTTEFWLG
NEKIHLISTQSAIPYALRVELEDWNGRTSTADYAMFKVGP
KYRLTYAYFAGGDAGDAFDGFDGDDPSDKFFTSHNGMQFSTW
DNDNDKFEGNCAEODGSGWWMNKCHAGHLNGVYYOGGTYSKAS
T



Precursor of fibrin. Fibrin polymerizes to form blood clots. Conversion of fibrinogen to fibrin is regulated via a cascade of factors to control blood clotting.

Protein Structure and Folding

Fundamental Questions

- ▶ What is the structure of this protein?
 - ▶ Can be experimentally determined, today we know the structure of ~15,000 proteins
 - ▶ Can be predicted for some proteins, usually in ~1 day on today's computers

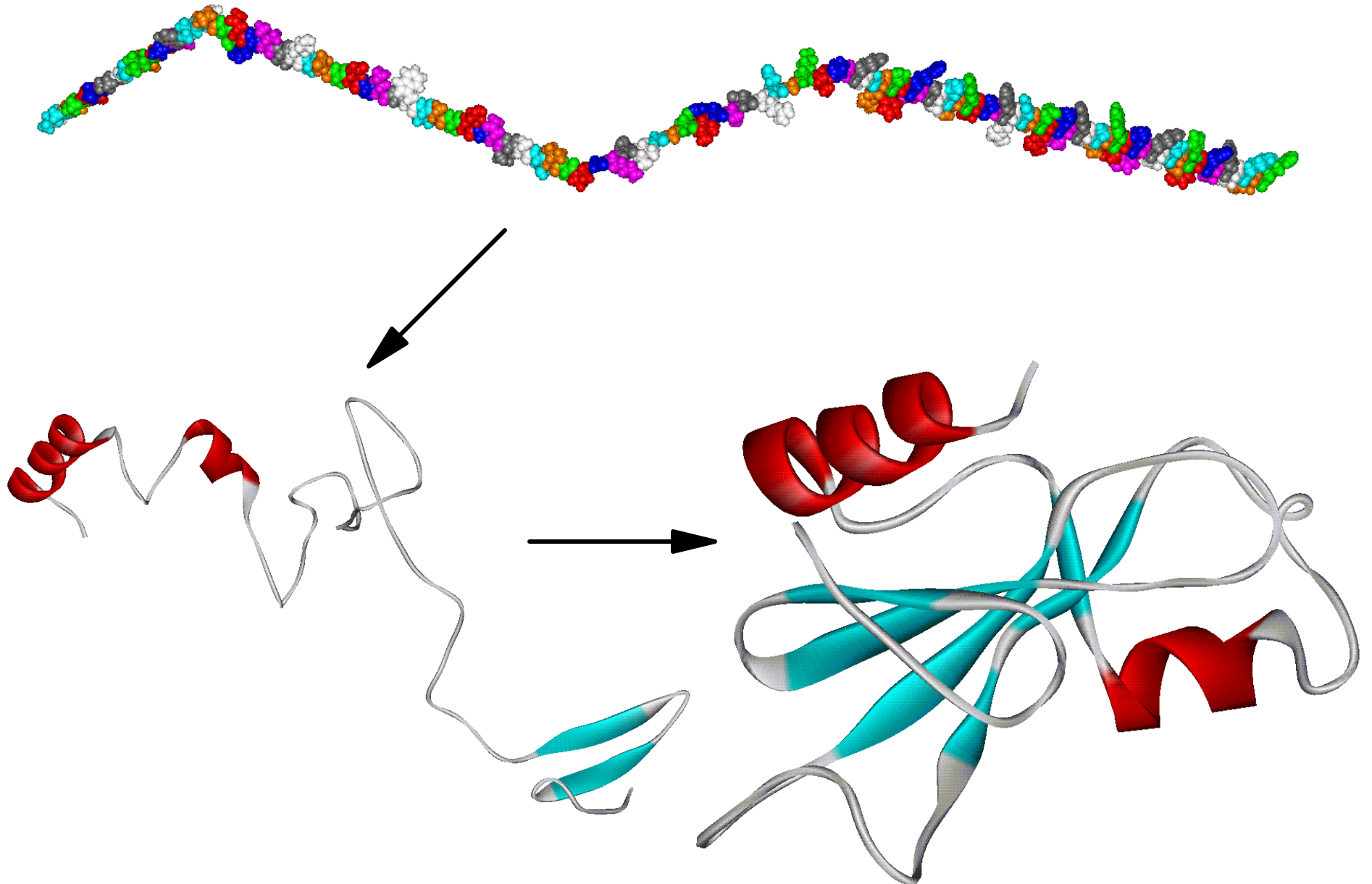
Protein Structure Prediction

- ▶ How does this protein form this structure?

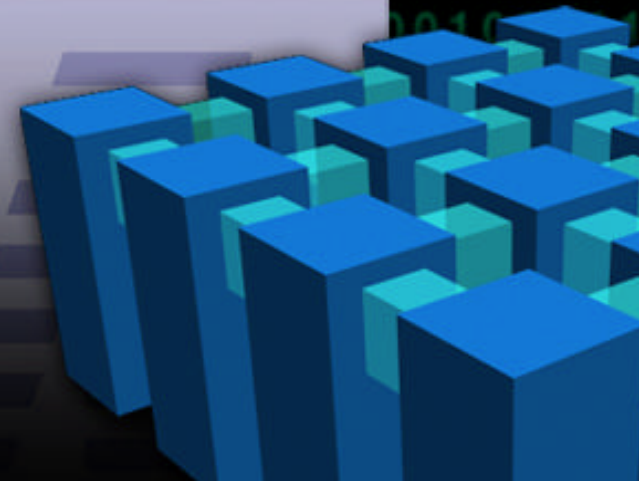
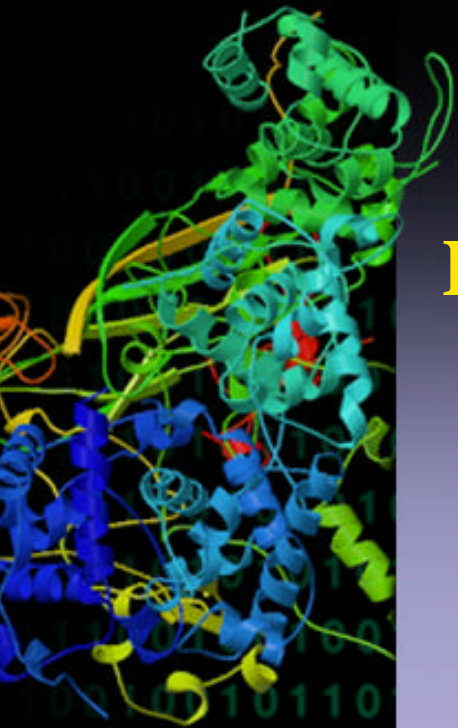
Protein Folding

- ▶ The process or mechanism of folding
- ▶ Limited experimental characterization
- ▶ Why does this protein form **this** structure?
 - ▶ Why not some other fold?
 - ▶ **Levinthal's Paradox:** As there are an astronomical number of conformations possible, an unbiased search would take too long for a protein to fold. Yet most proteins fold in seconds!

Protein Folding



Blue Gene: Petaflop Computing for Large Scale Biomolecular Simulation



Background and Goals for Blue Gene

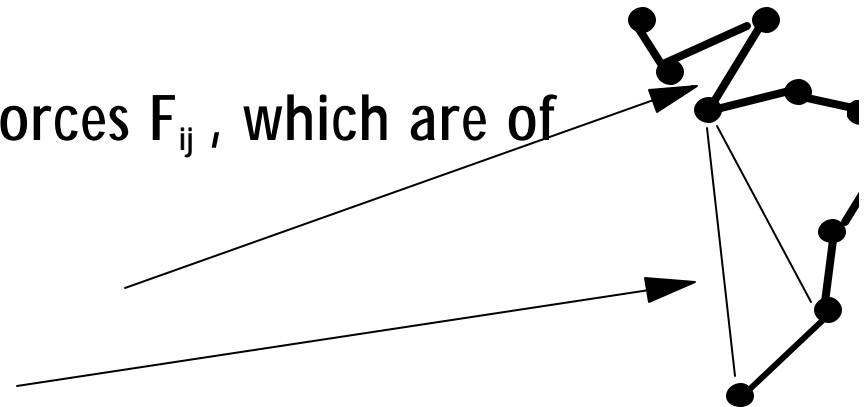
- In December 1999, IBM Research announced an ambitious project to build a large supercomputer with which to attack problems such as protein folding.
- The Blue Gene project has two basic goals:
 - Advance the state of the art of biomolecular simulation.
 - Advance the state of the art in computer design and software for extremely large scale systems.

Molecular Dynamics

- The classical equation of motion for atoms:
 - is solved in time steps of typically 1-10 fs. Index i goes over all N atoms in system.

$$m_i \frac{d^2 x_i}{dt^2} = \sum_j F_{ij}$$

- Comp. time is dominated by calculation of forces F_{ij} , which are of 2 key types:
 - bonded: stretch, bend and twist of bonds
 - non-bonded: van der Waals and Coulomb
- The Long range force calculation is dominant, because the index j goes over nearly n atoms, (globally $O(n^2)$, with a direct approach).
- In contrast the short range forces are order 1 per atom, (globally $O(n)$), and are not dominant despite a large prefactor.



Large Scale Dynamical Simulations: Molecular Dynamics

- Solve the classical equations of motion for the molecules
 - What does 1 petaflop-year buy in terms of simulation?
 - 32,000 atom simulation (implies $10^{* * 9}$ pairwise force evaluations/time step)
 - 150 FLOP/pairwise force evaluation
 - $1.5 \times 10^{* * 11}$ FLOPs/time step
 - $2 \times 10^{* * 11}$ time-steps
 - Typical simulation time step size for long-range forces is ~5 femtoseconds
 - Simulations that represent significant fractions of the entire folding time are enabled
-
-

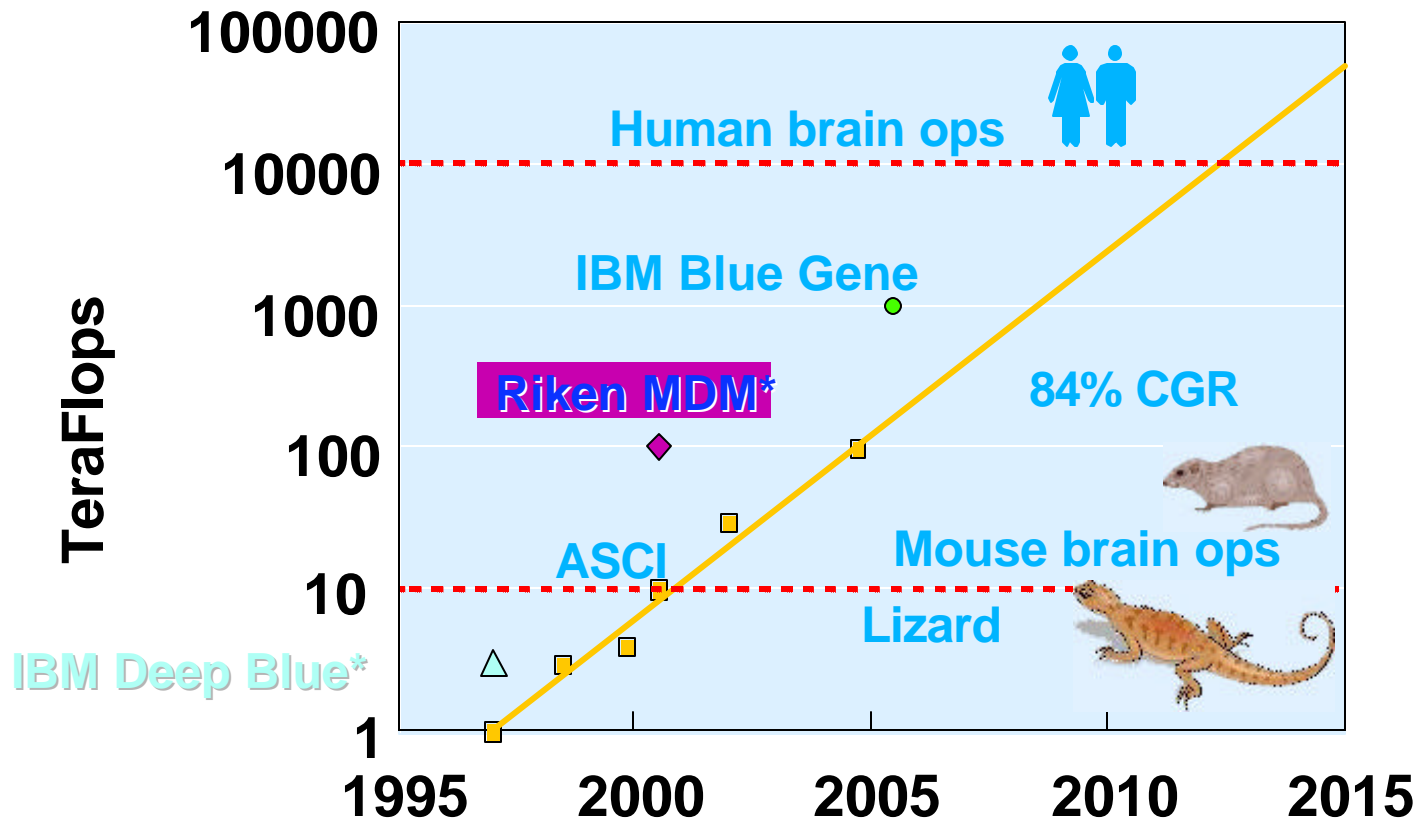
Why build large machines?

- **Because we have significant scientific problems that**
 - can be solved via computer simulation
 - have a large enough extrinsic benefit to justify a significant effort for a 10-15 year acceleration
 - **Because we gain useful insight in computer design by pushing the envelope**
 - especially when technology is at a new threshold
 - complete single chip systems
 - massively parallel system (at affordable size)
-
-

The Road to Petaflops



Supercomputing Roadmap



Source: ASCI Roadmap www.llnl.gov/asci, IBM

Brain ops/sec: Kurzweil 1999, [The Age of Spiritual Machines](#)

Moravec 1998, www.transhumanist.com/volume1/moravec.htm

Cellular Architecture

