

Computing Facilities @ Argonne

Rick Stevens

Argonne National Laboratory

University of Chicago

Outline

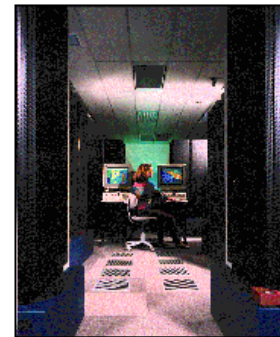
- Brief History of Computing Facilities at ANL
- Current Description [HW, SW, People]
- Usage Patterns and Current User Communities
- Facilities for Computer Science Research
- Advanced Networking and Grids @ ANL
- New User Communities and New Applications
- Trends and Strategies for the Future
- Questions!

Brief History of Argonne Computing Facilities

- Supporting both Computer Science and Applied Mathematics Research and DOE Advanced Applications
- 20 years of computing facilities innovation
 - First Unix on Vax 11/780 in DOE ~1980
 - First Denelcor HEP in DOE [first GP MIMD system]
 - First Sequent, first Encore, first BBN Butterfly + TC2000 in DOE
 - First Cydrome in DOE [VLIW]
 - First AMT DAP in DOE [SIMD]
 - First Thinking Machines Connection Machine in DOE [w/Caltech]
 - First IBM SP in DOE [installed simultaneously with Cornell]
 - First CAVE Virtual Realty System in DOE
 - First > 8 pipe Reality Monster [SGI Onyx2/Origin2000] in DOE

Overview of Argonne Computing Facilities

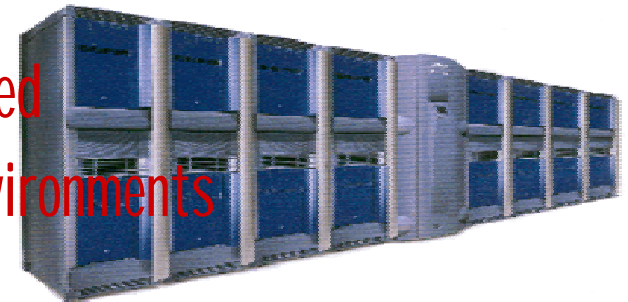
- The ACRF period [1983-1992]
 - Focus was on exploring parallel architectures
 - Developing programming models and software tools
 - Training new generation of CS researchers [> 1000]
 - Every major form of parallel computer architecture –1 [no dataflow]
 - ANL ACRF served as international center for parallel computing
- The HPCRC period [1992-1999]
 - ANL focused on production oriented parallel computing for Grand Challenges in addition to Computer Science



Argonne Computing Facilities Description



- Hardware Environment [Today]
 - IBM SP 144 nodes [ready for retirement!]
 - SGI Origin2000 [128 CPUs + 12 Infinite Reality Pipes]
 - 512 CPU IA-32 Linux Cluster + Graphics + Storage nodes
 - 120 TB IBM 3495 based tape system + 10 TB data front end
 - Special Purpose Clusters [IA-32 Linux + NT, Alpha Linux]
 - Visualization, Data Caches, Software Development, Compbio
 - Hundreds of PCs/Workstations and Servers
 - High-performance networking and QoS testbed
 - Five Access Grid nodes and development environments
 - Virtual Reality and Visualization laboratory



ANL Facilities: Available Software

- Software Environment [Hundreds of Open Source tools]
 - Comprehensive Suite of Software Development Tools
 - Compilers, libraries, debuggers, program viz, performance tools,...
 - Systems software for Parallel systems
 - Schedulers, systems management, filesystems, MP libraries,...
 - Accounting, visualization, data management,...
 - Grid Development Tools
 - Globus environment, network performance, collaboration tools,...
 - Application Code Suites
 - Climate modeling, Bioinformatics, Nuclear Physics, Astrophysics
 - Chemistry, Materials, CFD, Structural Biology, etc.

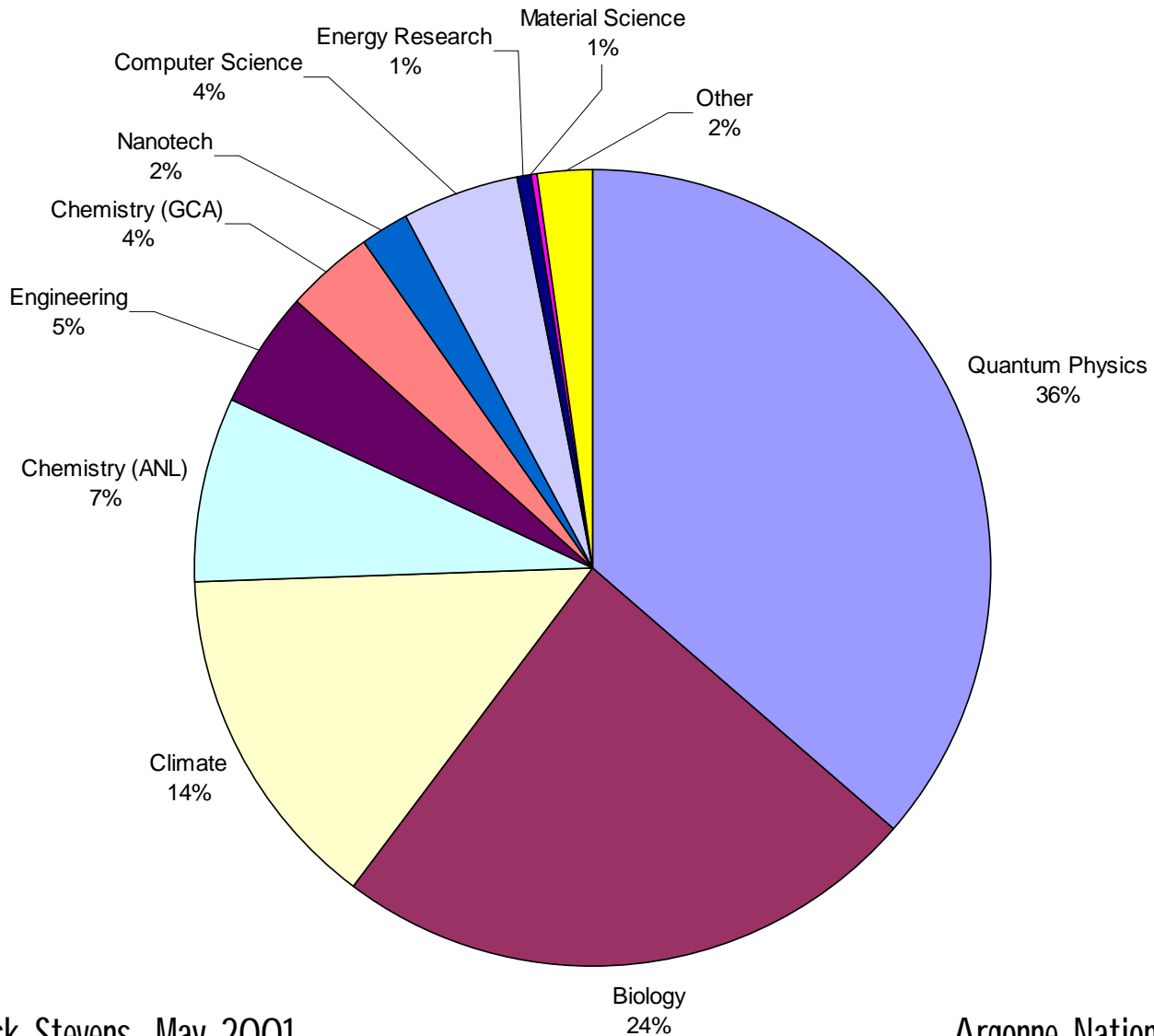
ANL Computational Facilities: Personnel

- MCS has a systems development and support staff of ~13 FTEs
 - 5 FTEs focused on Large-scale Systems
 - SP, SGI and Chiba City Clusters + Storage Systems
 - 4 FTEs focused on workstation computing environment
 - >500 + PCs/workstations, +40 servers
 - 2 FTEs on advanced network engineering and support
 - Multicast, all optical, research testbeds, security etc.
 - 2 FTEs on advanced visualization and collaboration infrastructure
 - CAVE, AccessGrid, Tiled Displays, Grid Support

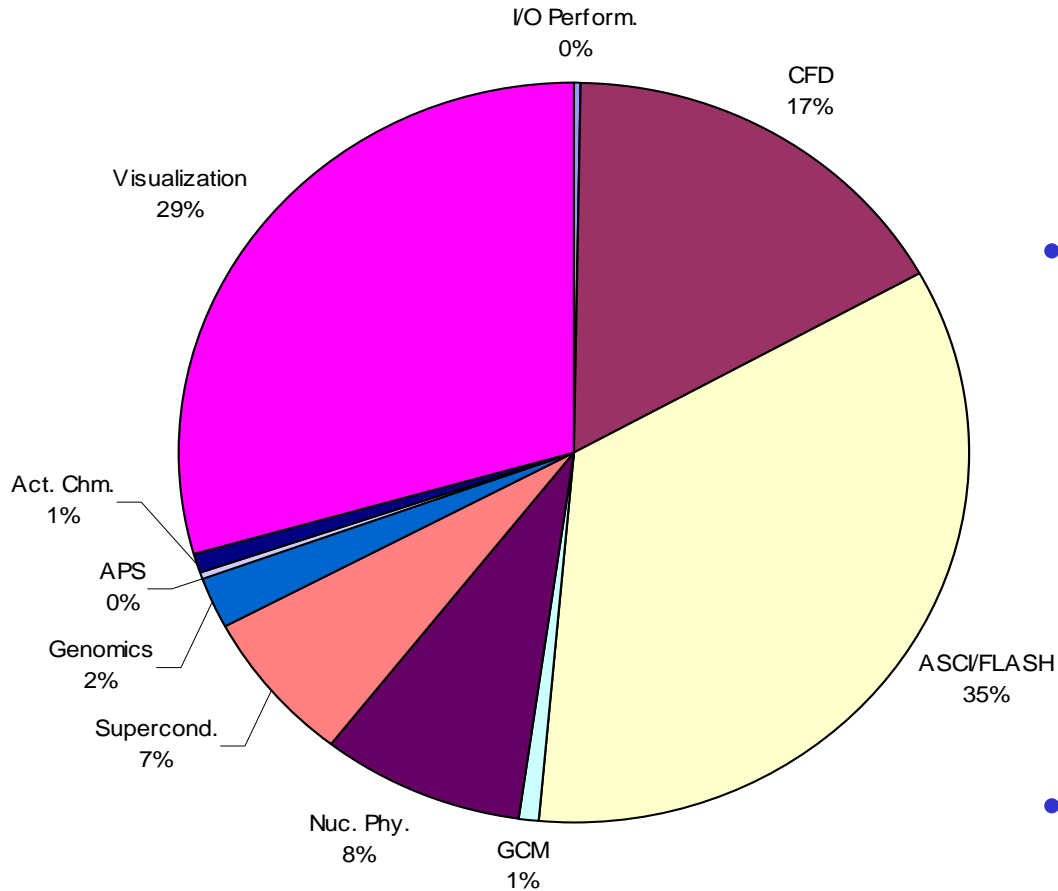
The ANL Computer Facilities User Base

- MCS Facilities support three communities
 - Computer Science and Math Researchers [internal + external]
 - Computational Science Developers [internal + external]
 - Semi-Production Computational Science Users [mostly internal]
- Total active accounts ~ 1000 users
 - Large-scale systems ~300 users
 - General MCS Division ~500 users
 - General Collaborators ~200 users

IBM SP – Usage



SGL Origin 2000 - Usage



- The SGL is split into two separate systems - Denali and Tundra.
- Denali - 96 cpus
 - Used for both CS and applications.
 - Always, always busy.
 - Challenge: scheduling.
 - Decision: emphasis on interactivity over performance. (I.e. CS over cycles.)
- Tundra: 32 cpus, 12 Irs
 - Dedicated to visualization.

What Facility Support Does Computer Science Need?

- Interactivity
 - Edit, Compile, Run, Debug/Run, Repeat.
 - In many cases those runs are very short and very wide.
- Flexible Systems Software
 - A specific OS, kernel, or a specific set of libraries and compilers.
 - [Which frequently conflict with some other user's needs]
 - Re-configurable hardware, Access to hardware counters
 - Permission to crash the machine, In some cases, root access
- Ability to test at Scale
- Non-requirements
 - Exclusivity. "Performance" is an issue, but typically only on timing runs.

Scalability – an Unrecognized Crisis

- Scalability is hard:
 - Complexity of solutions + Fault tolerance
 - Understanding program behavior is hard – huge amounts of data
 - Lack of available scalability testbeds
- Scalability research is important:
 - Improve real performance of applications
 - The size of the average system is growing. We need better scalability now. Everyone will need it in 3-5 years.
- To have a scalable system, we need scalable...
 - ... algorithms
 - ... development tools
 - ... Systems software and middleware
 - ... systems administration methods and tools

Chiba City - the Argonne Scalable Cluster

256 computing nodes.

512 PIII CPUs.

32 visualization nodes.

8 storage nodes.
4TB of disk.

Myrinet
interconnect.

Mission: Scalability
and open source
software testbed.

1 of 2 rows of Chiba City:



<http://www.mcs.anl.gov/chiba/>

Chiba City Timeline

- November 1998: Started thinking seriously about it
- October 1999: Installation
- November - February: Development
 - Development of the management software, debugging all kinds of things.
- March - June: Early users
 - Physics simulation code, weather code, communications libraries, ...
- August - Present: Production support
 - Available for research partners in computer science and computational science
 - Primary objective: reliable full-scale application runs
- June 2001: Scalable System Software Developers
 - Available to other system software projects that require a scalable testbed
 - Outreach to CS departments, and systems software developers and researchers

Chiba City Utilization — Availability is Important

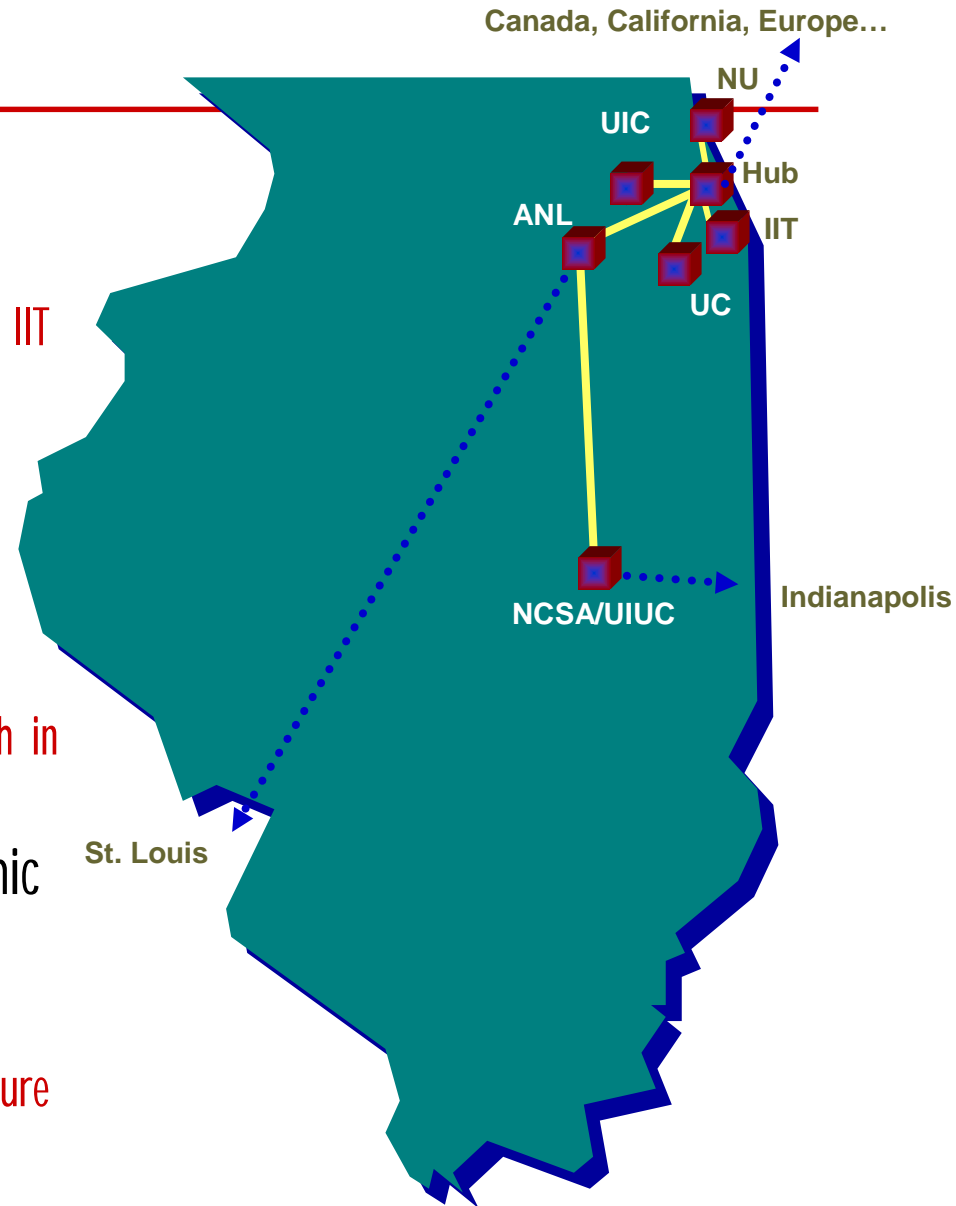
| Week | Node Hrs | % Utilization | Total |
|-------------|-----------------|----------------------|--------------|
| 2001/02/19 | 12,836.9 | 34.1 | 37,632.00 |
| 2001/02/26 | 12,678.2 | 33.7 | 37,632.00 |
| 2001/03/05 | 11,289.2 | 30.0 | 37,632.00 |
| 2001/03/12 | 15,984.5 | 42.5 | 37,632.00 |
| 2001/03/19 | 17,302.2 | 46.0 | 37,632.00 |
| 2001/03/26 | 15,500.4 | 41.2 | 37,632.00 |
| 2001/04/02 | 17,971.1 | 47.8 | 37,632.00 |
| 2001/04/09 | 23,362.2 | 62.1 | 37,632.00 |
| 2001/04/16 | 19,436.9 | 51.7 | 37,632.00 |
| 2001/04/23 | 23,212.3 | 61.7 | 37,632.00 |

Networking and Grids

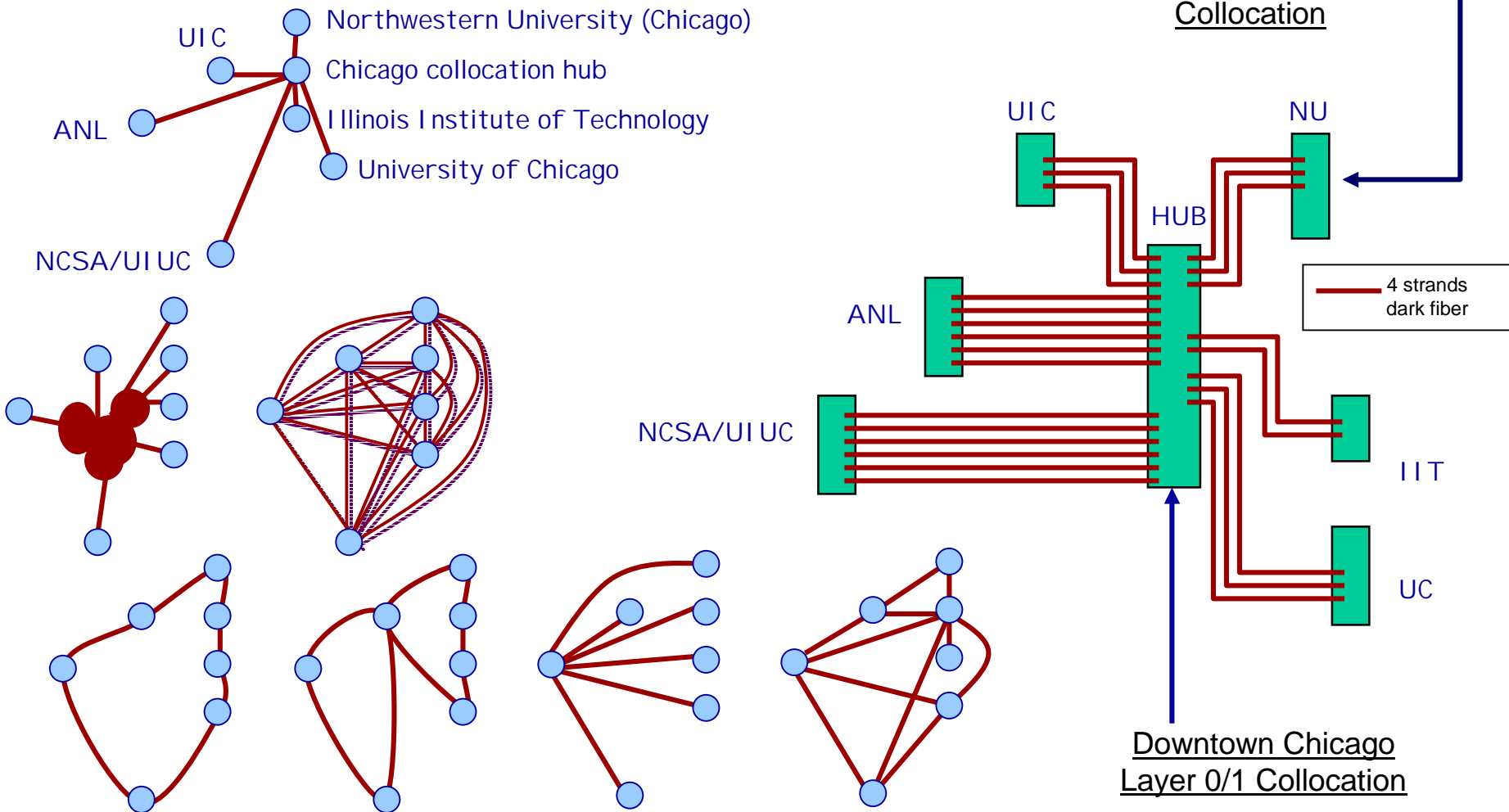
- Driven by dispersion of researchers [people remain the most critical resource, even in the 21st century]
- Connecting computational resources: the biggest machines, data storage, the [local] user interface to computation
- Collaborative technology [Access Grids]
- Existing large-scale Grid building efforts
 - The distributed terascale NSF proposal: ANL, NCSA, SDSC, Caltech
 - The Consortium for Cell Signaling: U. of Texas, 20 others
 - The GriPhyN Project: Grid tech for CMS, ATLAS, LIGO, SDSS
- Grids are not useful without BANDWIDTH

I-WIRE

- Leverages Longstanding Research Partnerships
 - ANL/UC, NCSA/UIUC, UIC, Northwestern, IIT
- Addresses Need for Network Research Program
 - Essential to DOE science, mathematics, computer science research
 - Recognition of rapid ongoing evolution in network technologies and exponential growth in demand.
- Vision to create joint industry / academic / laboratory partnerships
 - Accelerate concept-to-reality
 - Invent future applications by simulating future technology environments



I-WIRE Proposed Topology

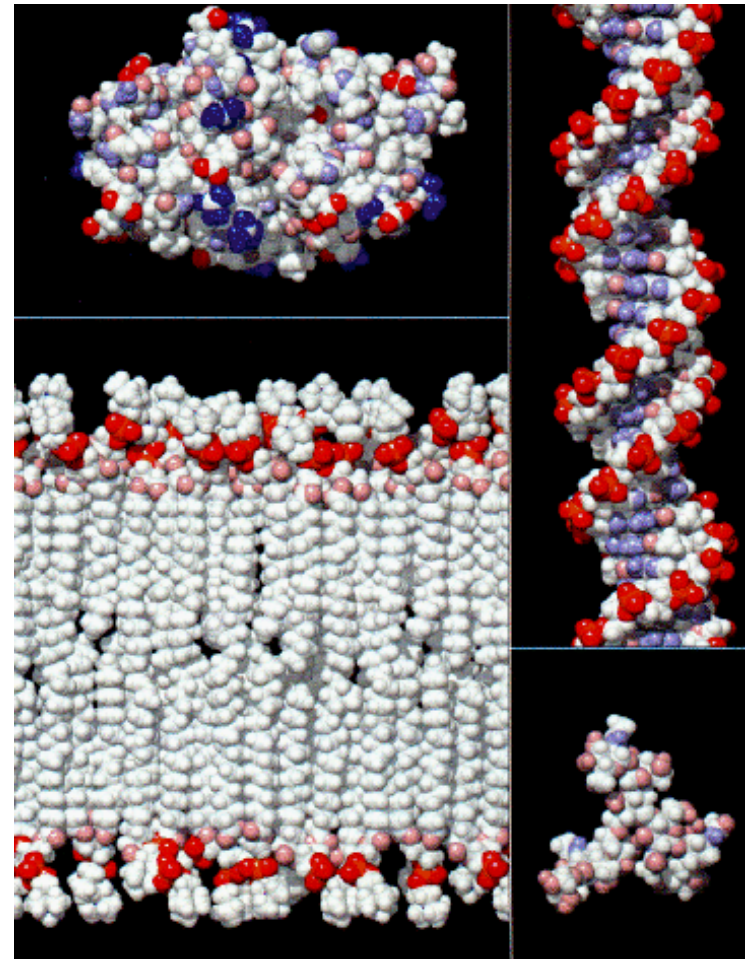


The New User Communities @ ANL

- Biology and Nanoscience are our strategic focus for Growth
 - Genome-to-life and National Nanotechnology Initiatives
- Continuing support for Climate, HEP/NP, Nuclear Engineering and Energy Systems [but don't see much strategic growth]
- We believe "Grids" will emerge as the dominate mechanism for application delivery [compute + data + people + tools]
- *DOE computing facilities must become Grid enabled*
 - And part of the national and international fabric of compute and data resources available to scientists through advanced resource sharing mechanisms

Computational Biology and Bioinformatics Areas

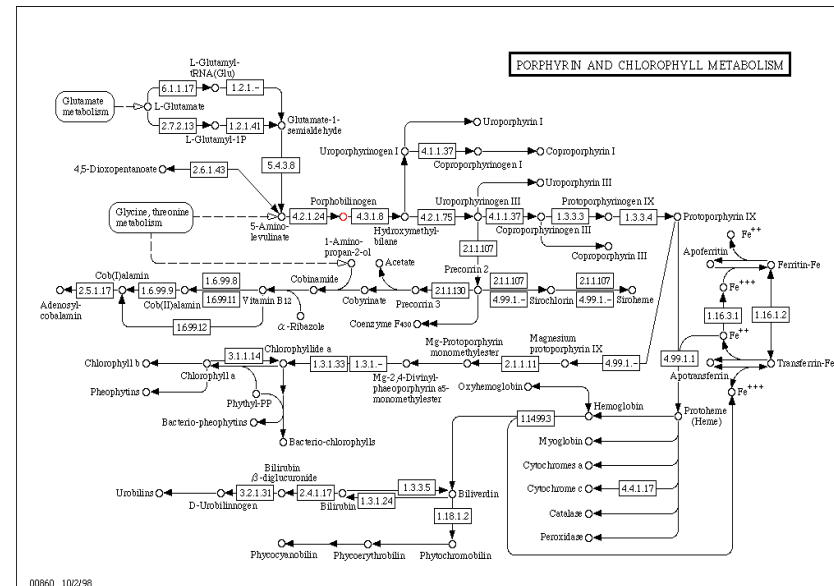
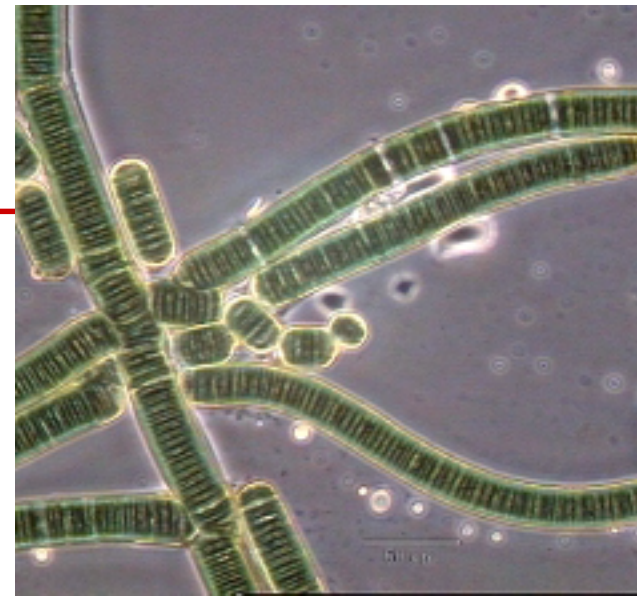
- Sequence analysis, interpretation and annotation
- Structural biology modeling and simulation
- Computational approaches to proteomics
- Analysis of gene expression array data
- Computer models of cells and cellular processes
- Computer Aided Biological Design (CABD)



Biological CAD: Tools for Design in Life Science

- Understanding biological systems from an information systems standpoint [e.g., organization, communication, transformation]
- Modeling biological systems: genes, molecules, pathways, organelles, cells, tissues, organs and organisms, communities
- Designing new biological structures and systems:

- New biochemical pathways
- Engineered microorganisms
- Computer Aided Genome Design
- Synthetic Model Organisms



Compare Your Facility to Other Leading HPC Centers

- Historically our focus has been primarily on enabling advanced CS research and supporting a limited set of applications...
 - We have less FLOPS and BYTES than NERSC, NCSA or SDSC
 - We have more computer scientists than most Centers
 - We do more software development than most HPC centers
 - We have substantially fewer support staff than most HPC centers
 - We don't see ourselves competing with most HPC Centers but rather complementing them and collaborating on enabling technologies
 - We believe we can take on more risky technologies than HPC Centers primarily focused on production computing for applications
 - We believe our facilities quality is world class, and enabling world class CS

The Most Important Trend in HPC for the Next Decade?

- Complete dominance of commodity technologies and architectures
 - Transition from PCs to embedded systems as best price/performance
 - How to leverage this trend is the key HPC systems architecture issue...
- Increasing availability of bandwidth and storage capacities
 - All optical networking will drive bandwidth prices steeply down
- Continuing Struggle with Scalability in Software
 - By 2010 we will be building systems \gg 1 M CPUs
- Grid oriented infrastructure will be incredibly important
 - Data intensive science communities will demand Grid-like facilities
- Hopefully something unexpected will make future HPC more interesting

Advanced Computing Facilities

- Immediate: 1–2yrs
 - Develop $O[10^3]$ Node Software Scalability Research System
 - Deploy Teraflops Applications Science Compute Engine [SciDAC + GTL]
- Near Term: 3–5yrs
 - Upgrade Scalability Testbed and Applications Engine [10–30 TF]
 - Prototype Petaflops Capable Systems Technologies and Micro Architectures
- Medium Term: 5–10yrs
 - Deploy Targeted Petaflops System for Biology and Nanoscience Apps
 - Prototypes for Exaflops System Technologies and Micro Architectures
- Long Term: 10–20yrs
 - Testbed Facility for Alternative Programming Model Development
 - Deploy Exaflops Applications Computing Environment

Networking Infrastructure and Facilities

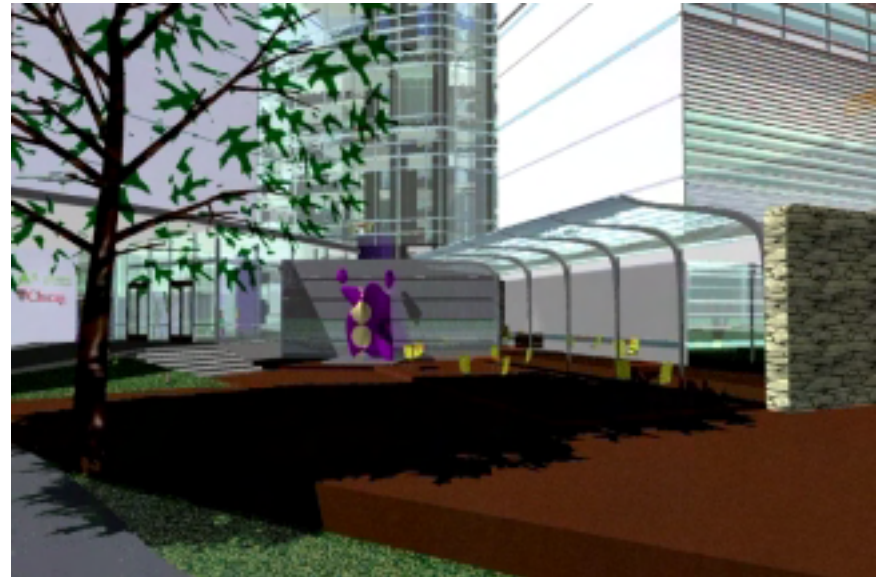
- Immediate: 1–2yrs
 - I-WIRE Testbed and Grid Collaboration and Computing Infrastructure
 - Design New Conventional Facilities for Computation Related Activities
- Near Term: 3–5yrs
 - Upgrade of ANL Wide Networking and Distributed Storage Infrastructure
 - Occupancy of New Conventional Facilities [e.g. MCS, ACS, CI, Etc.]
- Medium Term: 5–10yrs
 - Production Use of National/International All Optical Research Network
 - Development of Laboratories for Alternative Computing Model Research
- Long Term: 10–20yrs
 - Establish Argonne Constellation Science Centers
 - Construction of Computing Facilities for Exaflops Systems

Where Does Moore's Law Take Us?

| Technology Area | Annual Growth Rate | 1 | 5 | 10 | 20 | 30 |
|------------------------------|---------------------|-------------|-------------|-------------|-------------|-------------|
| OPS/sec/\$ | 60% | 10.0E+6 | 104.9E+6 | 1.1E+9 | 120.9E+9 | 13.3E+12 |
| FLOPS/sec/\$ | 60% | 5.0E+6 | 52.4E+6 | 549.8E+6 | 60.4E+9 | 6.6E+12 |
| Bytes/\$ (RAM) | 70% | 1.0E+6 | 14.2E+6 | 201.6E+6 | 40.6E+9 | 8.2E+12 |
| Bytes (Low)/\$ (disk) | 60% | 200.0E+6 | 2.1E+9 | 22.0E+9 | 2.4E+12 | 265.8E+12 |
| Bytes (High)/\$ (disk) | 40% | 50.0E+6 | 268.9E+6 | 1.4E+9 | 41.8E+9 | 1.2E+12 |
| Network NIC (bits/sec/\$) | 50% | 100.0E+6 | 759.4E+6 | 5.8E+9 | 332.5E+9 | 19.2E+12 |
| Clock Frequency (GH) | 35% | 1.4E+9 | 6.3E+9 | 28.1E+9 | 566.0E+9 | 11.4E+12 |
| \$1,000 Systems | Fraction TSC | 2000 | 2005 | 2010 | 2020 | 2030 |
| FLOPS | 20% | 1.0E+9 | 10.5E+9 | 110.0E+9 | 12.1E+12 | 1.3E+15 |
| RAM (Bytes) | 20% | 200.0E+6 | 2.8E+9 | 40.3E+9 | 8.1E+12 | 1.6E+15 |
| Storage (Bytes) | 30% | 60.0E+9 | 629.1E+9 | 6.6E+12 | 725.4E+12 | 79.8E+15 |
| BW (bits/sec) | 20% | 10.0E+9 | 53.8E+9 | 289.3E+9 | 8.4E+12 | 242.0E+12 |
| \$1,000,000 Systems | Fraction TSC | 2000 | 2005 | 2010 | 2020 | 2030 |
| FLOPS | 10% | 500.0E+9 | 5.2E+12 | 55.0E+12 | 6.0E+15 | 664.6E+15 |
| RAM (Bytes) | 20% | 200.0E+9 | 2.8E+12 | 40.3E+12 | 8.1E+15 | 1.6E+18 |
| Storage (Bytes) | 30% | 60.0E+12 | 629.1E+12 | 6.6E+15 | 725.4E+15 | 79.8E+18 |
| BW (bits/sec) | 20% | 10.0E+12 | 53.8E+12 | 289.3E+12 | 8.4E+15 | 242.0E+15 |
| \$10,000,000 Systems | Fraction TSC | 2000 | 2005 | 2010 | 2020 | 2030 |
| FLOPS | 10% | 5.0E+12 | 52.4E+12 | 549.8E+12 | 60.4E+15 | 6.6E+18 |
| RAM (Bytes) | 20% | 2.0E+12 | 28.4E+12 | 403.2E+12 | 81.3E+15 | 16.4E+18 |
| Storage (Bytes) | 30% | 600.0E+12 | 6.3E+15 | 66.0E+15 | 7.3E+18 | 797.5E+18 |
| BW (bits/sec) | 20% | 100.0E+12 | 537.8E+12 | 2.9E+15 | 83.7E+15 | 2.4E+18 |
| \$100,000,000 Systems | Fraction TSC | 2000 | 2005 | 2010 | 2020 | 2030 |
| FLOPS | 10% | 50.0E+12 | 524.3E+12 | 5.5E+15 | 604.5E+15 | 66.5E+18 |
| RAM (Bytes) | 20% | 20.0E+12 | 284.0E+12 | 4.0E+15 | 812.8E+15 | 163.9E+18 |
| Storage (Bytes) | 30% | 6.0E+15 | 62.9E+15 | 659.7E+15 | 72.5E+18 | 8.0E+21 |
| BW (bits/sec) | 20% | 1.0E+15 | 5.4E+15 | 28.9E+15 | 836.7E+15 | 24.2E+18 |

Building Critical Mass – Argonne Theory Institute

- Organized to spread advanced computational culture throughout Lab computational and theoretical activities
- ~ 400 people
 - Theorists, computer scientists, computational scientists
 - Lab joint appointments with ANL divisions
 - University joint appointments with UofC departments
- Advanced Computing Environment
 - Petaflops-scale computing resources
 - Research and development systems
 - Storage for data intensive science
 - Advanced visualization and imaging
 - Terabit network access to the Grid
 - Integrated collaborative spaces



Endi

Slides after this are Backups

Computer Science and Math Research

- Immediate: 1–2yrs
 - Scalable Scientific Computing Software for Commodity Based Systems
 - Networking, Distributed Computing and Collaboration Technologies*
- Near Term: 3–5yrs
 - New Directions for Applied Mathematics and Scientific Software
 - Integration of Scientific Computing, Sensors and Experimental Systems*
- Medium Term: 5–10yrs
 - Exploitation of New HPC Architectures (e.g. Grids, Bg, BI++, Pim, Ic)
 - New Classes of Distributed Scientific Applications Environments*
- Long Term: 10–20yrs
 - Automated Problem Solving and Rapid Software Development
 - Alternative Models for Computation (e.g. QC, BioComp)

Computational Science Applications

- Immediate: 1–2yrs
 - Launch Internal Activities for Computational Biology and Nanoscience
 - Form Partnerships Programs for Climate Modeling and CFD
- Near Term: 3–5yrs
 - Establish Centers for LS Code Development in Bio and Nano
 - Form Partnerships for Energy Systems and Nuclear Engineering
- Medium Term: 5–10yrs
 - Establish Large-scale Computational Science Centers in Bio and Nano
 - Establish Virtual Design/problem Solving Centers for Partnership Apps
- Long Term: 10–20yrs
 - Fully Integrate Computation With Theory and Experimental Programs
 - Applications of New Computing Models (e.g. QC, BioComp)

Example Application Accomplishments

- Computational Biology
 - Computational modeling of time-resolved protein dynamics [GC w/Harvard]
 - Computational modeling of MDH catalysis [GC w/Harvard]
- Computational Chemistry
 - Parallel ab initio methods for MPPs [Grand Challenge with PNNL]
- Materials Science
 - Million Particle Molecular Dynamics Simulations of Porous Glasses:
- Astrophysics
 - Astrophysical turbulence, multigrid and higher-order methods
- Climate Modeling
 - 650-year run of Fully Coupled Climate Model

Example Application Accomplishments

- Physics

- Nuclear Forces: Conducted first ab initio computations of 10-body nuclei on the Linux cluster
- Multidimensional Deformation Space: Four-dimensional shape space calculations disproved the presence of rotational bands that experimentalists had claimed existed in association with hyper deformed shapes of certain isotopes.

- Advanced Photon Source

- Crystallography - Improved optimization and numerical techniques into LaueView.
- Microtomography - Won the Third Annual Global II Award for use of the GUSTO testbed for real-time, collaborative analysis of data from a microtomographic beamline

Example Computer Science Accomplishments

- ANL Computing facilities have enabled researchers to explore new techniques for portable, scalable parallel programming
 - Linear Algebra –LAPACK and the level-3 BLAS,...
 - Programming Models - Monmacs, and p4, precursors to MPI
 - Scientific toolkits - PETSc has enabled several Gordon Bell awards
 - Parallel I/O - ROMIO has been adopted by several vendors and is being used on all three ASCI machines as part of the DOE Accelerated Strategic Computing Initiative [ROMIO has been a reference model for MPI-IO]

Example Computer Science Accomplishments

- Parallel Filesystem – PVFS has achieved > 3 GB/s on Chiba City
- Mesh generation, partitioning, and refinement – Software developed in conjunction with Argonne research on parallel unstructured meshes won the 1992 Gordon Bell prize for scalable performance
- Optimization software – Nug30 optimization problem was attacked with resources at ANL and elsewhere
- First interactive supercomputer in the loop CAVE application was developed at ANL in 1994 [IBM SP to SGI Onyx]

Learning Experiences

- Barnraising:
 - Building these things by hand with lots of volunteers is fun - our scarcest resource was space.
 - Detailed instructions are critical.
 - Our rate: 128 nodes/day.
- Configuration
 - The hierarchical, database-driven approach has worked very well.
 - Remote power and remote console are awesome.
- Pain:
 - Replacing all the memory in the cluster.
 - Upgrading the BIOS on every node.
 - We stress hardware far more than vendors do - AGP lossage, memory
- Scalability Challenges
 - Myrinet
 - Took a little while for us to get all of the nodes using myrinet happily (early driver releases, mapper, ...)
 - Very small error rates can kill in the large.
 - RSH doesn't scale.
 - RSH is used by default to launch jobs, but can only invoke 256. Boom.
 - Network gear
 - Gets very confused when 32 nodes all try to boot through them at once.
 - PBS uses UDP for internal communication. UDP loses badly in big, congested

A Few Computer Science Updates

- PVFS - a parallel file system
 - From Clemson / ANL
 - Using 48 I/O Nodes, a single 20GB file, and reading on 112 nodes, Rob is seeing:
 - 3.1 GB/sec writes
 - 2.9 GB/sec reads
 - Very, very soon in production on Chiba. Next week.
- Distributed Tiled Displays
 - Running the Future Lab's Active Mural, a 15-panel, 4096x1995 pixel display, for many kinds of visualization apps.
- MPD, the MPICH multi-purpose daemon
 - An experiment into architecture for job management
 - Can currently launch 100 processes/second
 - 2.6 seconds from pressing return on a front-end node until all processes have started
 - Have tested up to 2000 processes.
- Initial "Practical Scalability Tests"
 - Cluster reboot and rebuild times
 - NFS: How far can you stretch an NFS server?

The Msys and City Toolkits

Msys

A toolkit of system administration programs, including:

- `cfg` - centralized management of configuration files
- `sanity` - a tool for automatic configuration checking and fixing
- `pkg` - tools and mechanisms for flexible software installation
- `softenv` - tools for setting up the user's environment that adapt to software changes
- `hostbase` - a database of hostname information and scripts for driving all name-related services
- `clan`, `whatami` - utilities
- `anlpasswd` - a `passwd` replacement that catches guessable passwords

Rick Stevens, May 2001

City

Cluster-specific tools that build on top of Msys.

- `chex` - the node console management system
- `citydb` - the database used to manage `cfgs`
- `chex` - the node management system
- `city_transit` - file distribution scripts
- filesystem images
- power utilities
- account management

Msys and City are both open source.

Both toolkits are available at:

<http://www.mcs.anl.gov/systems/software/>

Both were supported by ANL LDRD and DOE funds.

Argonne National Laboratory + University of Chicago

Chiba City – Open Issues

- Systems Issues
 - Reliably running 512-cpu jobs for a long time is difficult. This is our primary priority.
 - PBS is not a reliable scheduler at our scale. It crashes a lot.
 - Focused development time.
- Computer Science Usability
 - Some researchers still use the farms-of-workstations approach because it's just a bit easier.
 - Being able to schedule non-exclusive jobs.
- Our Own Success
 - The machine is oversubscribed.
 - We need to move the SP users somewhere.
 - Growth of computational science across ANL.
- An ideal scenario for 2001.
 - 256-node Chiba for our computational science partners.
 - 256-node Chiba for ANL.
 - 1024-node Chiba for scalable computer science and computational science.
 - A development team.