Millennium: Cluster Technology for Computational Science and Engineering

David Culler

E. Brewer, J. Canny, J. Demmel,
A. Joseph, J. Landay, S. McCanne
A. Neureuther, C. Papadimitrou, K. Yelick

EECS, U.C. Berkeley

ILP

March 12, 1998

Project Goals

- Enable major advances in Computational Science and Engineering
  – Simulation, Modeling, and Information Processing becoming ubiquitous
- Explore novel design techniques for large, complex systems
  – Fundamental Computer Science problems ahead are problems of scale
- Develop fundamentally better ways of assimilating and interacting with large volumes of information
  – and with each other
- Explore emerging technologies
  – networking, OS, devices

Topics Today

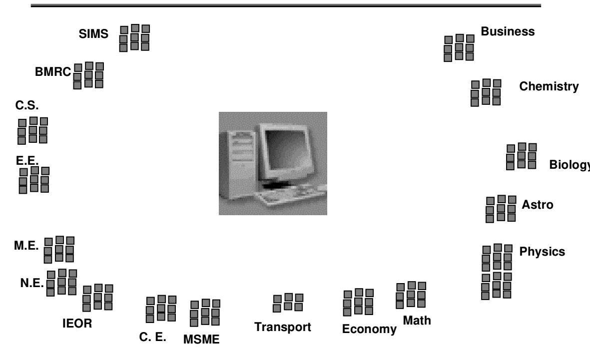
- Millennium Test bed
- Cluster-base High Performance Computing
- Towards a Computational Economy

The Millennium Test Bed

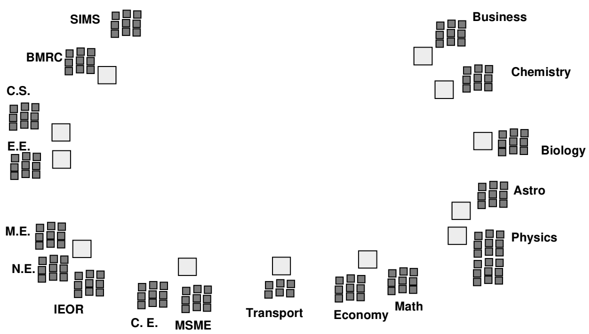
17 Campus Units and NERSC

School of Info. Mgmt and Sys.
Business
BMRC
Chemistry
Computer Science
Electrical Eng.
Biology
Astro
Mechanical Eng.
Physics
Nuclear Eng.
IEOR
Civil Eng.
MSME
Inst. Of Transport
Economy
Math

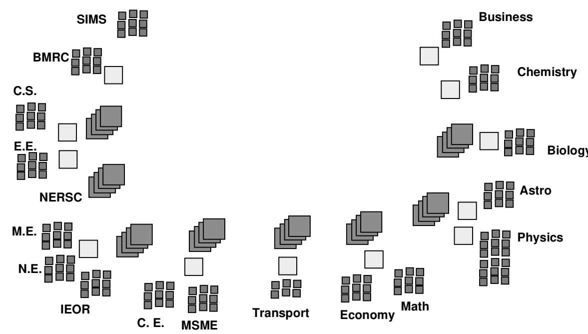
Sci. & Eng. NT Workstation



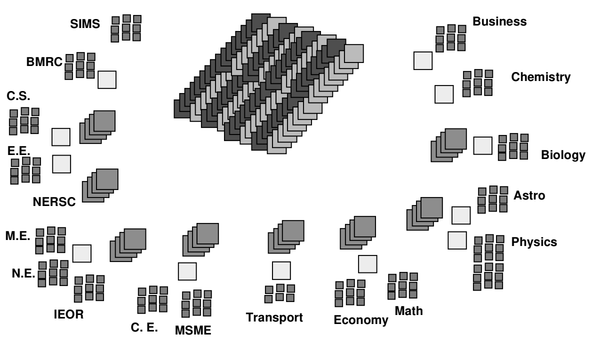
SMP server => storage, small-scale parallelism



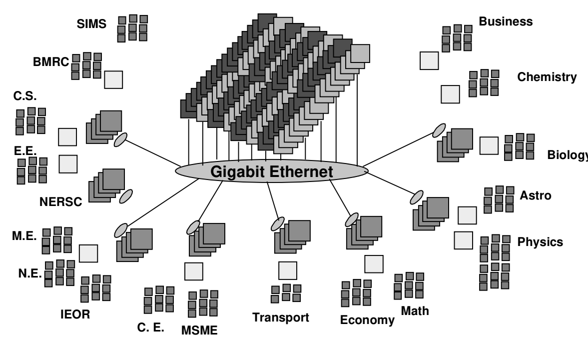
Group Cluster of SMPs => Parallelism



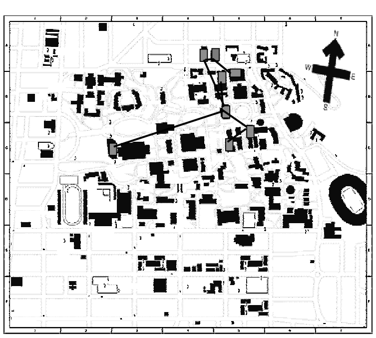
Campus Cluster => large-scale Parallelism



Gigabit Ethernet Connectivity



Physical Connectivity



Visualization and Novel User Interfaces

Industrial / Academic Collaboration

- Computers via Intel Technology 2000 grant
  - 200 NT desktops
  - 16 department 4-way SMPs
  - 8 5x4 Group Clusters,
  - 1 ~100x4 Campus Cluster
  - PPro => Pentium II => Merced
- Additional storage via IBM SUR grant
  - 0.5 TB this year => 4 TB
- NT tools via Microsoft grant
- Solaris x86 tools via SMCC grant
- Bay Networks discounts the gigabit Ethernet
- Campus provides Technical staff
- Research provides the prog. and system support

200 Gflop/s
150 GB memory
8 TB disk

Sample Applications

- Astrophysical Simulations
  - Star formation
  - Turbulence in geophysical flows
  - Data-mining Cosmic Microwave Background Radiation
- CEE Pacific Earthquake Eng. Research Center
  - Finite element modeling of earthquake impact
- Technology CAD
  - Simulation of E-beam and Optical Lithography
- National Aerospace System Emulation
- Phylogenetic History of Life

The CS Research Agenda

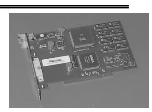
- High Performance Cluster Computing Environment
  - Fast communication on Clusters of SMPs
  - Compiler Techniques for Performance and Ease of use
    - » see K. Yelick's ILP talk
  - Numerical Techniques and Solvers
    - » Particles, FFT, AMR, Multigrid, Sparse and Dense Lin. Alg.
- Novel System Design Techniques
  - clusters of clusters
  - Computational Economy
- Novel modes of interacting with large amounts of data

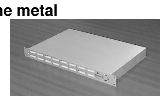
Design of a Large Cluster for S&E

- Classic Architecture Problem "in the large"
  - Given fixed budget, what is the best partitioning of node, group and campus cluster resources?
- Basic node has several degrees of freedom
  - processors per node (4, 2, 1) - Disks
  - memory capacity - Space, Volume
  - PCI busses - Power
- Clustering adds additional degrees of freedom
  - network, network interfaces
- Cost is well-defined (Intel)
- Workload is defined by real applications
- Design against technology change
  - Quad PPro, Dual P*II*, P*II*, … *Merced*

Cluster Interconnect Design

- Proposed design based on MyriNet
  - 16+8 port switch in fat-tree variant
  - today offers best latency, BW, simplicity, flexibility, and cost
    - » source-based packet routing, open to the metal
  - link-by-link flow control with cut-through routing
  - almost reliable





- System Area Network (SAN) revolution
  - Tandem/Compaq ServerNet

Gigabit Ethernet

Communication Interface Revolution

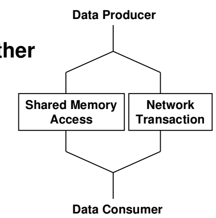
- Low Overhead Communication "Happens"
- Academic Research put it on the map
  - Active Messages (AM), FM, PM, …Unet
  - Memory Messaging (Get/Put, Reflective, VMMC, Mem. Chan.)
- Intel / Microsoft / Compaq recognized it

VIA

  - Virtual Interface Architecture 1.0 released 12/16/97
- Apply UCB virtual networks to VIA

Multiprotocol Communication

- Hardware has two fundamental protocols
- Communication may involve either
- At what level is this exposed?
  - Who must cope with it?
- Uniform Programming model
  - Message Passing (MPI)
    - » multiprotocol run-time
  - Shared address space
    - » shared virtual memory
    - » multiprotocol code-generation
- Hybrid Programming model
  - MPI + threads = performance * complexity



Example: Multiprotocol AM

- Careful shared-memory programming to get BW within SMP
  - cache alignment, special copy routine
- Novel Concurrent Access Algorithm for shared message queue object
  - lock-free techniques borrowed from non-blocking literature
  - depends on synchronization operations of instruction set and system timing
- Attention to network protocol impacts memory protocol
  - adaptive fractional polling

Inter-Cluster Networking

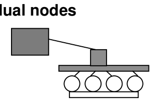
- Gigabit Ethernet - what was the question?
  - ATM, FiberChannels, HPPI, Serial HPPI, HPPI 6400, SCI, P1394, … fading fast
  - standard due in April
- Not the Ethernet you remember
  - switched, full duplex - multiframe bursts
  - broadcast, multicast trees - level 3 switching
  - flow control - QoS support
- Fast Network Interfaces
- Switches clean and fast
- Clearly the Storage and Video Transport
- Is it also the Cluster solution?
  - VIA/IP

Inter-Cluster Research Agenda

- Vastly expands the scope of systems challenge
  - integrate well-connected resources according application needs, rather than physical packaging
  - resource allocation, management, and administration
- Network bandwidth matches display BW
  - Protocols and run-time sys. for visualization, media transport, interaction, and collaboration.
- Community can share non-trivial resources while preserving sense of ownership
  - Bandwidth translates into efficiency of exchange
  - Data can be anywhere
- Important networking technology in its own right.
  - Layer 3 switching, QoS, VLan

Remote Execution

- NOW lessons
  - UNIX syscall / command interface does not virtualize well
    - » inter-positioning helps
  - Global support more error prone than individual nodes
    - » good design helps
    - » watch-dogs and fast restart help
  - Explicit coordination tends to be very fragile
  - Complex system interactions
  - No allocation policy pleases all

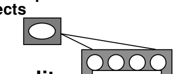


=> Need looser, more robust design techniques

- Key developments
  - Smart Clients: decision making close to the user
  - Implicit Co-ordination: use naturally occurring events to schedule resources
  - Virtual Networks: fast communication with multiprogramming

Millennium "Smart Client"

- Adopt the NT "everything is two-tier, at least"
  - UI stays on the desktop and interacts with computation "in the cluster of clusters" via distributed objects
  - Single-system image provided by wrapper



- Client can provide complete functionality
  - resource discovery, load balancing
  - request remote execution service
- Higher level services 3-tier optimization
  - directory service, membership, parallel startup

Computational Economy Approach

- System has a supply of various resources
- Demand on resources revealed in price
  - distinct from the cost of acquiring the resources
- User has unique assessment of value
- Client agent negotiates for system resources on user's behalf
  - submits requests, receives bids or participates in auctions
  - selects resources of highest value at least cost

Advantages of the Approach

- Decentralized load balancing
  - according to user's perception of what is important, not system's own metric
  - adapts to system and workload changes
- Creates Incentive to adopt efficient modes of use
  - exploit under-utilized resources
  - maximize flexibility (e.g., migratable, restartable applications)
- Establishes user-to-user feedback on resource usage
  - basis for exchange rate across resources
- Powerful framework for system design
  - Natural for client to be watchful, proactive, and wary
  - Generalizes from resources to services
- Rich body of theory ready for application

Millennium Resource Allocation

- Property rights establish "fair share" currency
  - each brings resources to the system
- Price determined by competition for the resource
- User (agent) determines value
- Provide enabling technology for Evolution of markets
  - bilateral trade
  - multilateral trade
  - standardized contracts
  - markets for resources and services
- Monitor how it progresses
- Elevate useful applications into Services

Millennium Target Community

- Sophisticated computational scientists
  - willing to author new applications
  - push the state of the art
- Researchers that want easy-to-use tools
  - some computer knowledge, but just a tool
  - Example: maximum likelyhood calculations for evolutionary biology
- Students

Approach: Focus on Services

- Most users use services (only)
  - such users don't need accounts on all systems
  - easier to use, output is graphs/visualization
  - enables easy student/class usage
  - services solve specific problems
    - » protein folding, SVD, simulations, ...
- Some users will still log in, write apps
- Easy conversion of apps to services

Service Economics

- Services make economic models simpler!
  - Services simplify resource tracking over time
  - Build models for each service
    - » can tie resource needs to service inputs
    - » can bid well based on history
  - Services are well defined => pay per use
- Services abstract resources
  - enables high availability
  - enables varying resources over time
- Current Demonstration: TACC transformational services
  - transcend, wingman

Primitives for a Comp. Economy

- Server side
  - Monitoring of resource usage, enforcement of contracts
  - major challenge in Unix
    - » build parallel thread structure and interpose on calls
    - » fundamentally same machinery for redirection
  - support in NT 5.0
- Client side
  - agents, protocols, UI
- Bidding, negotiation, brokering
  - requests for quotes, auctions, value functions
- Banking
  - digicash

System Administration

- Uniformity is key
- Clusters evolve and are constantly changing over time
- Administrative domains tend to diverge

=> create incentive to simplify administration
  - more uniform, higher value

- Build automated system providing weakly consistent database of the state of system health and inference rules
  - apply expert system diagnosis technology

Systems of Systems Design

- It is about making things work at large scale
  - things change, things break, demands extreme
- Make all components wary, reactive, and self-tuning
- Use implicit information whenever possible
- User behavior is critical to closing the loop
  - when there is personal responsibility

- Millennium is a good model of large scale systems challenges

What is Millennium About?

- An experiment in large-scale system design
- Advance the state of computational science and engineering
- Exploring novel design techniques
- Exploring important new technologies