Network Based Computing: Trends, Issues and Challenges

Dhabaleswar K. (DK) Panda
Department of Computer Science and Engineering
The Ohio State University

E-mail: panda@cse.ohio-state.edu
http://nowlab.cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Presentation Outline

Dominant Computing System and Paradigm

- Continues to change over the years
- Objective: design cost-effective systems
- Depends on two primary factors
  - technological trends (processor, memory, disk, interconnection)
    - past
    - current and future
  - computational demand of applications
    - past
    - current and emerging

Past and Current Systems

- Single Board Computer, Mainframe, Personal Computer, Workstation, Multiprocessors, Client-Server, Massively Parallel Processor
- Have been dominant computing systems during some years
- Where are we headed next?

Basic Resources for Designing Computing Systems

- Computing cycles
  - capacity (total amount of CPU cycles)
  - rate (MHz)
- Memory and Disk
  - capacity (MBytes)
  - bandwidth (MBytes/sec)
  - latency (nano/micro/milli sec)
- Interconnection/network between processor-memory-disk
  - bandwidth (MBytes/sec)
  - latency (microsec)

Presentation Outline

Technological Trends in the Past

- Processor
  - Speed: 1.35x per year (before mid 1980s), 1.58x per year (afterwards)
  - Cost ($ per MIPS): reduces by a factor of 0.5 per year (83-90)
- Memory (DRAM)
  - Capacity: 4x-every-three-years (1.6x per year)
  - Cycle time: 1/3 in 10 years
  - Cost ($ per MByte): reduces by a factor of 2.7 per year
- Magnetic Disk
  - Capacity: 1.25x per year (prior to 1990), 1.50x per year (afterwards)
  - Access time: 1/2 in 10 years (80-90)
  - Cost ($ per MByte): reduces by a factor of 1.3/year (86-92), 2.0/year (92-95)

Technological Trends in the Past (Contd.)

- Interconnection schemes
  - System Bus: 10.0 MBytes/sec (Multibus-I) to 1024 MBytes/sec (Sun SMP system)
    - 100 times improvement
  - Interconnection Network: 1.0 MByte/sec (CM-2) to 300 MBytes/sec (Cray T3D/T3E)
    - 300 times improvement
  - Local Area Network (LAN):
    - Ethernet (1974-): 10 Megabits/sec (1.25 MBytes/sec)
    - Token Ring (1980-): 4/16 Megabits/sec (0.5/2.0MBytes/sec)
    - FDDI (1987-): 100 Megabits/sec (12.5 MBytes/sec)
    - Fast Ethernet (1993-): 100 Megabits/sec (12.5 MBytes/sec)
    - 10 times improvement in 20 years

Impact on Designing Past and Current Computing Systems

- Computing systems have been designed by ignoring LANs
  - Mainframes
  - Symmetric MultiProcessors (SMPs)
  - Massively Parallel Systems (MPPs)
  - Personal Computers
  - Desktop Workstations
- Objective: Design cost-effective systems by taking advantage of on-going advancements in processor, memory, disk, bus, and interconnection network technologies.
- LANs/WANs have been primarily used to
  - provide connectivity among computing systems
  - support file transfers between computing systems (client-server) and resources (disks, printers, scanners, …)
  - support transactions between system

Technological Trends - Current and Future

- Networking Technology
- Powerful Uniprocessor/SMP Systems
- Emergence of the Web Technology

Growth in Commodity Network Technology

Representative commodity networks; their entries into the market

Ethernet (1979 -)	10 Mbit/sec
Fast Ethernet (1993 -)	100 Mbit/sec
Gigabit Ethernet (1995 -)	1000 Mbit /sec
ATM (1995 -)	155/622/1024 Mbit/sec
Myrinet (1993 -)	1 Gbit/sec
Fibre Channel (1994 -)	1 Gbit/sec
InfiniBand (2001 -)	2.5 Gbit/sec (1X) → 2 Gbit/sec
10-Gigabit Ethernet (2001 -)	10 Gbit/sec
InfiniBand (2003 -)	10 Gbit/sec (4X) → 8 Gbit/sec
InfiniBand (2005 -)	20 Gbit/sec (4X DDR) → 16 Gbit/sec
	30 Gbit/sec (12X SDR) → 24 Gbit/sec
InfiniBand (2007 -)	40 Gbit/sec (4X QDR) → 32 Gbit/sec

16 times in the last 6 years

Growth in Powerful Computing Systems

- Processor speed is doubling every 18 months
- Design and development of powerful and cost-effective uni-processor/multi-processor systems
- Use of commodity components (processor, memory, and disk) and economy of scale keep on pushing down the prices of these boxes
- Allowing people to have powerful computing boxes on their desks

Emergence of the Web Technology

- Web technology to share information and data
- CGI scripts and Java support
  - to integrate computation with communication
  - to blur distinction between heterogeneous computing systems (architectures, operating systems, instruction sets)
- Rapid growth in the last ten years

Summary of Technology Trends

- Network has been ignored as an integral component of a computing box in the past
- Three important developments in 90's
  - commodity and affordable networking technologies
  - powerful uniprocessor/SMP systems
  - emergence of the Web technology

Presentation Outline

- Motivation
- Technological Trends
- Applications (Past, Current, and Emerging)
- Defining Network-Based Computing Systems
- Challenges and Research Issues
- Research Challenges and Projects
- Available Experimental Testbed
- Related 888 Course
- Conclusions

Past Applications

- Sequential Batch Jobs
- Grand Challenge (Parallel) Applications
- Transactions-based Applications
  - airline/hotel/car reservation
  - databases

Trend for Computational Demand

- Continuous increase in demand
  - multiple design choices
  - larger data set
  - finer granularity of computation
  - simulation with finer time step
  - low-latency/high-throughput transaction, ……
- Expectation changes with the availability of better computing systems

Emerging Applications

- High Performance and High Throughput Computing Applications
  - Weather forecasting, physical modeling and simulations (aircraft, engine), drug designs, …
- Database applications
  - data-mining, data ware-housing, enterprise computing
- Financial
  - e-commerce, on-line banking, on-line stock trading
- Digital Library
  - library of audio/video, global library
- Collaborative computing and visualization
  - shared virtual environment
- Telemedicine
  - content-based image retrieval, collaborative visualization/diagnosis
- Virtual Reality, Education, and Entertainment

High Performance Computing

- Large number of parallel applications in scientific and engineering disciplines
- Grand Challenge Applications
  - weather forecasting
  - manufacturing (computer-aided design, modeling, and simulation)
  - designing and analyzing VLSI circuits
  - analyzing chemical formulae for new medicines
  - simulation and visualization of physical phenomena (astronomy, deforestation, global warming, ...)
  - simulating nuclear reactions (US Department of Energy's ASCI program)

High Throughput Computing

- Simulation experiments for large number of design choices
- Common requirement for scientists, engineers, physicists, chemists, biomedical engineers, ...
- Need results in a short period of time
- Example: design of an aircraft wing
  - 5 degrees of freedom (major parameters)
  - 5 different values for each parameter
    - 25 different simulation experiments (independent)
    - each simulation experiment takes 2 hours
- Availability of 25 computers over a network will allow design decisions to be made in 2 hours (instead of 50 hours using a single computer)

Grid Computing

- Applications require computing systems with enormous computing power at affordable cost
- Some applications require sharing of computational cycles of a set of systems over network
  - to provide meta-computing/grid computing environment to users
  - to distribute parallel jobs over a set of systems
- Some applications require high-speed connectivity between systems to support cooperative parallel computation
- Example: Weather Forecasting, Virtual Radio Astronomy Observatory, ..

Scalable Database Servers

- Continuous increase in demand for organizing data, searching them, and making decisions
- Designing database servers which
  - deliver high performance and throughput and scalable
  - support a range of thick and thin clients over network
- Data-mining and knowledge discovery
  - establishing association/correlation between data
  - deriving patterns from data
- Requires enormous amount of computing power
  - to integrate data generated on a daily/weekly/monthly basis
  - to search all possible association/correlation/pattern between data

Web/Multimedia Server

- Web technology is providing a new dimension
  - to share information and data (text, audio, video, image, ..) across geographically distributed systems
  - to search for information
  - to allow collaboration
- Number of URLs keep increasing in an exponential manner
- Requires scalable servers which
  - provide fast search/access
  - allow fast transfer of data between servers and clients
  - support both push and pull technologies

Enterprise Computing

- Computing for decision making within a corporation/organization
  - using data from sales, inventory, purchase, production, ... -
  - spreaded over multiple heterogeneous computers of independent vendors/suppliers/organizations
- Example: Moving an army for a battle at a distant location (Iraq War) within 7 days
- Requires coordination between
  - soldiers over different base units
  - a number of countries and their defense units
  - a number of government agencies/departments
  - a large number of suppliers
  - transportation units

Digital Library

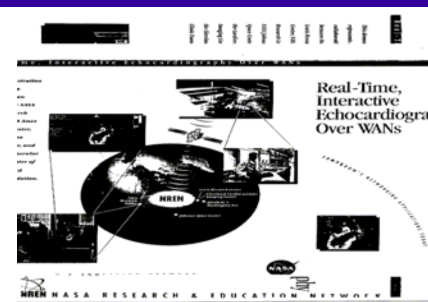
- Connecting databases of libraries within a city/campus/state/nation for easy search of titles
- Global digital library (Federation of digital libraries)
- Full search capabilities on articles (not keyword search on titles or abstracts)
- Library of audio (radio programs)
- Library of video (TV programs)
- Library of medicine (big initiative by US National Institute of Health)
  - visual human body
  - database of brain tumors
  - database of cancer

Collaborative Interaction

- Video Teleconferencing
  - real-time processing and transfer of audio clips and video images
  - significant advantage for organizations to cut down travel time and cost and enhance productivity
- Shared Virtual Environment
  - shared chalkboards/whiteboards
  - group editors
  - computer-supported cooperative work (solving a big software problem using experts from geographically distributed areas)

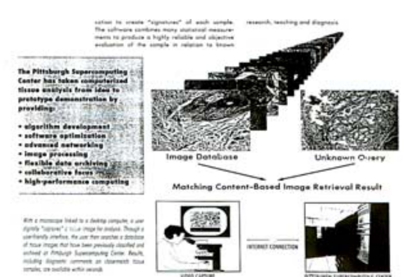
Telemedicine

- Biggest push in the current and next decade to provide quality health-care with affordable cost
- Remote Diagnosis
  - connecting smaller hospitals and clinics with bigger ones
  - connecting remote places (flight shuttles, air planes, space stations)
  - ability to transfer x-ray/MRI images/ECGs in real-time
  - ability to have remote feeling/touch using haptic devices
  - video teleconferencing capabilities
- Sharing of patient records
  - sharing of text, X-ray, MRI image, Ultra-sound, ECG, ...
  - among doctors across several hospitals/clinics to minimize tests
  - sharing of records (without identification) with researchers and medical students for medical research and education



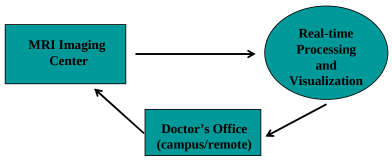
Telemedicine (Contd.)

- Content-Based Image Retrieval
  - database of images (tissues, cancer cells, tumors, ...) at a center (state/nation level)
  - pathologists at hospitals try to identify new images
  - can initiate a query to the center for identifying complex images
  - speedy diagnosis



Telemedicine (Contd.)

- On-line Processing of MRI Images and Diagnosis
  - currently done in an off-line manner
  - takes a few hours to develop the complete image
  - a patient may be asked to come for MRI several times



Virtual Reality

- Simple Virtual Reality (2D, one-to-one)
  - virtual tourism
  - virtual hiking/biking ......
  - flight simulation/navigation
- CAVE Immersive Virtual Environment (3D, one-to-one)
  - Cave Automatic Virtual Environment (CAVE)
  - volumetric visualization and interaction
  - exploration of data in a visual manner
  - discovery of knowledge

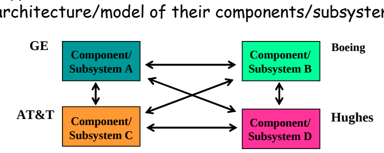
Figure 2.
CAVE immersive virtual environment

Electronic Commerce

- On-line banking
- On-line shopping
  - virtual distributors
  - virtual warehouse
  - real-time interaction between
    - consumer
    - merchant
    - consumer's bank
    - merchant's bank
- On-line buying and selling stock
  - real-time interaction between
    - consumer
    - broker
    - stock exchange

Manufacturing

- Component-based simulation of a product
  - parts of an aircraft
  - parts of a car
- Integrated simulation of a product
- Integrated simulation of a product with components from different vendors
  - different vendors need not disclose the architecture/model of their components/subsystems



Education

- Distribution and sharing of course materials
  - class notes and lecture (audio and video)
- Homework submission and grading (electronically)
- Virtual Classroom
  - distance education
  - faculty teaching in his/her office
  - students listening to it from dorms
- Virtual Campus/University
  - shared virtual environment for interaction between
    - student-student
    - student-faculty
    - faculty-faculty

Entertainment

- Video-on-Demand
  - 24-hour access to video stores
- Digital Leisure (Integration of Computer and TV)
  - video chatroom
  - interactive neighborhood game
  - virtual window shopping (with friends)
  - virtual family room (with distant family members)

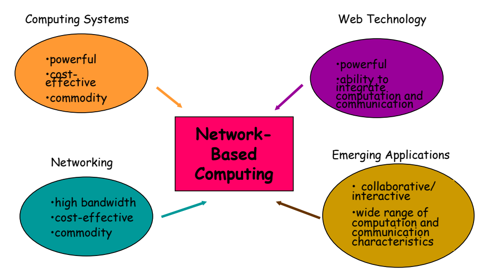
Summary of Applications

- Continuous increase in computational demand
- Wide range of applications with different computation and communication characteristics
- Many applications are collaborative/interactive in nature
- How to build computing systems to support modern applications while taking advantage of advancement in technologies?
  - computing
  - networking
  - web

Presentation Outline

- Motivation
- Technological Trends
- Applications (Past, Current, and Emerging)
- Defining Network-Based Computing Systems
- Challenges and Research Issues
- Research Challenges and Projects
- Available Experimental Testbed
- Related 888 Course
- Conclusions

Network-Based Computing



A closer look at the applications

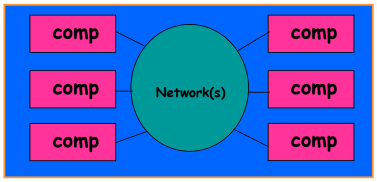
- Characteristics
  - tasks
    - parallel
    - independent sequential
  - computation
    - non-interactive
    - interactive
      - user(s)/client(s) over the LAN
      - user(s)/client(s) over the WAN
  - data
    - centralized
    - distributed

Different Types of NBC Systems

- Five major categories
  - Dedicated Cluster over a LAN
  - Clusters with Interactive clients
  - Globally-Interconnected Systems
  - Interconnected Clusters over WAN
  - Multi-tier Data Centers

Dedicated Cluster over a SAN/LAN

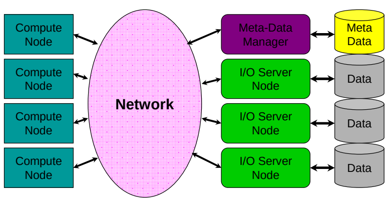
- Packaged inside a rack/room interconnected with SAN/LAN



Trends for Computing Clusters in the Top 500 List

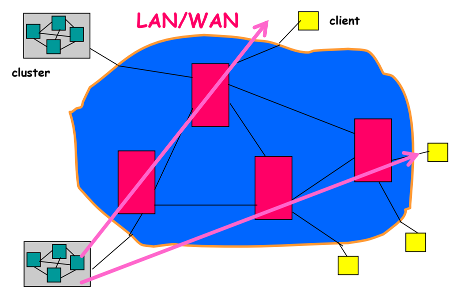
- Top 500 list of Supercomputers (www.top500.org)
  - June 2001: 33/500 (6.6%)
  - Nov 2001: 43/500 (8.6%)
  - June 2002: 80/500 (16%)
  - Nov 2002: 93/500 (18.6%)
  - June 2003: 149/500 (29.8%)
  - Nov 2003: 208/500 (41.6%)
  - June 2004: 291/500 (58.2%)
  - Nov 2004: 294/500 (58.8%)
  - June 2005: 304/500 (60.8%)
  - Nov 2005: 360/500 (72.0%)
  - June 2006: 364/500 (72.8%)
  - Nov 2006: 361/500 (72.2%)
  - June 2007: 373/500 (74.6%)
  - Nov 2007: 406/500 (81.2%)

Clustered File Systems: Parallel Virtual File System (PVFS)

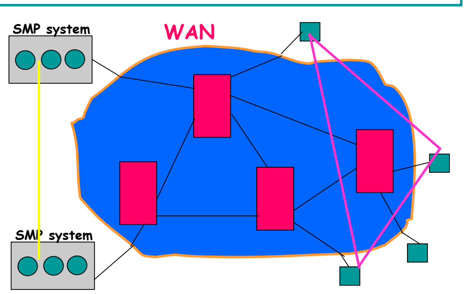


- *Relies on Striping of data across different nodes*
- *Tries to aggregate I/O bandwidth from multiple nodes*
- *Utilizes the local file system on the I/O Server nodes*

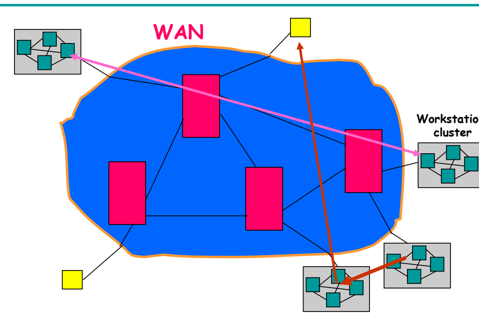
Cluster with Interactive Clients



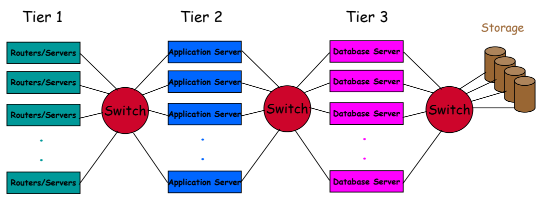
Globally-Interconnected Systems



Interconnected Clusters over WAN

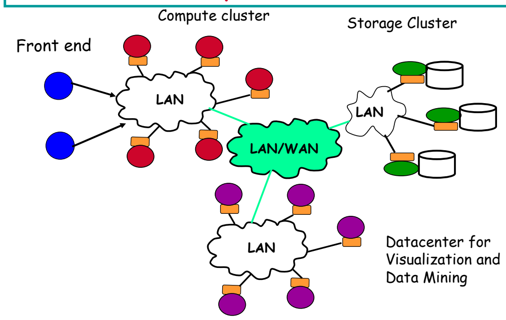


Generic Three-Tier Model for Data Centers



- All major search engines and e-commerce companies are using clusters for multi-tier datacenter
- Google, Amazon, Financial institutions, …..

Integrated Environment with Multiple Clusters



Presentation Outline

- Motivation
- Technological Trends
- Applications (Past, Current, and Emerging)
- Defining Network-Based Computing Systems
- Challenges and Research Issues
- Research Challenges and Projects
- Available Experimental Testbed
- Related 888 Course
- Conclusions

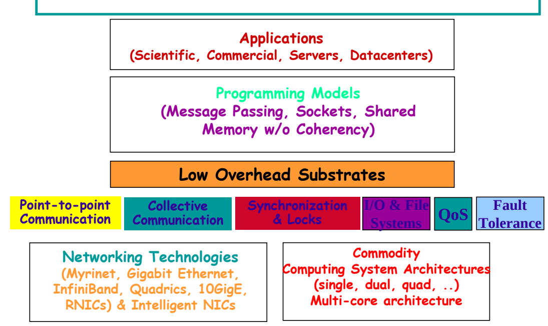
Requirements for Modern Clusters, Cluster-based Servers, and Datacenters

- Requires
  - good Systems Area Network with excellent performance (low latency and high bandwidth) for interprocessor communication (IPC) and I/O
  - good Storage Area Networks high performance I/O
  - good WAN connectivity in addition to intra-cluster SAN/LAN connectivity
  - Quality of Service (QoS) for supporting interactive applications
  - RAS (Reliability, Availability, and Serviceability)

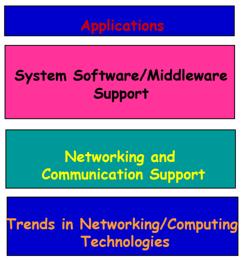
Our Vision

- Network-Based Computing Group to take a lead in
  - Proposing new designs for high performance NBC systems by taking advantages of modern networking technologies and computing systems
  - Developing better middleware/API/programming environments so that modern NBC applications can be developed and implemented in a scalable fashion
- Carry out research in an integrated manner (systems, networking, and applications)

Research Challenges



Research Challenges and Their Dependencies



Major Research Directions

- System Software/Middleware
  - High Performance MPI on InfiniBand Cluster
  - Clustered Storage and File Systems
  - Solaris NFS over RDMA
  - iWARP and its Benefits to High Performance Computing
  - Efficient Shared Memory on High-Speed Interconnects
  - High Performance Computing with Virtual Machines (Xen-IB)
  - Design of Scalable Data-Centers with InfiniBand
- Networking and Communication Support
  - High Performance Networking for TCP-based Applications
  - NIC-level Support for Collective Communication and Synchronization
  - NIC-level Support for Quality of Service (QoS)
  - Micro-Benchmarks and Performance Comparison of High-Speed Interconnects
- More details on http://nowlab.cse.ohio-state.edu/ → Projects

Group Members

- Currently 10 PhD students and 2 MS students
  - PhD: Lei Chai, Wei Huang, Matthew Koop, Ping Lai, Amith Mamidala, Sundeep Narravul, Ranjit Noronha, Xiangyong Ouyang, Gopal Santhanaraman and Karthik Vaidyanathan
  - MS: Rahul Kumar and Karthik Gopalakrishnan
- Two students graduated during Summer/Autumn
  - Sayantan Sur (PhD, IBM TJ Watson)
  - Abhinav Vishnu (PhD, IBM)
- All current students are supported as RAs
- Several new students in the group
- System administrator and Programmer
  - Jonathan Perkins
  - Brian Curtis

External Collaboration

- Industry
  - IBM TJ Watson and IBM
  - Intel
  - Cisco
  - Dell
  - SUN Microsystems
  - Apple
  - Linux Networx
  - NetApp
  - Mellanox, Qlogic and other InfiniBand companies
  - Chelsio, NetEffect and other iWARP companies
- Research Labs
  - Sandia National Lab
  - Los Alamos National Lab
  - Pacific Northwest National Lab
  - Argonne National Lab
- Other units in campus
  - Bio-Medical Informatics (BMI)
  - Ohio Supercomputer Center (OSC)

Research Funding

- National Science Foundation
  - Multiple grants
- Department of Energy
  - Multiple grants
- Sandia National Lab
- Los Alamos National Lab
- Pacific Northwest National Lab
- Ohio Board of Regents
- IBM Research
- Intel
- SUN
- Mellanox
- Cisco
- Linux Networx
- Network Appliance

Research Funding (Cont'd)

- Three Major Large-Scale Grants
  - Started in Sept '06
  - DOE Grant for Next Generation Programming Models
    - $10M for five years
    - $1M for OSU/NBC Group
    - Collaboration with Argonne National Laboratory, Rice University, Univ. of Berkeley, and Pacific Northwest National Laboratory
  - DOE Grant for Scalable Fault Tolerance in Large-Scale Clusters
    - $10M for five years
    - $1M for OSU/NBC Group
    - Collaboration with Argonne National Laboratory, Indiana University, Lawrence Berkeley National Laboratory, and Oakridge National Laboratory
  - AVETEC Grant for Center for Performance Evaluation of Cluster Interconnects
    - $750K for three years for OSU/NBC Group
    - Collaboration with major DOE Labs and NASA
- Several continuation grants from other funding agencies
- Several other large-scale grants are under review

Presentation Outline

- Motivation
- Technological Trends
- Applications (Past, Current, and Emerging)
- Defining Network-Based Computing Systems
- Challenges and Research Issues
- Research Challenges and Projects
- Available Experimental Testbed
- Related 888 Course
- Conclusions

NBCLab Infrastructure

- Multiple Clusters (around 500 processors in total)
  - 128-processor
    - 32 dual Intel EM64T processor
    - 32 dual AMD Opteron
    - InfiniBand DDR on all nodes (32 nodes with dual rail)
  - 128-processor (Donation by Intel for 99 years)
    - 64 dual Intel Pentium 2.6 GHz processor
    - InfiniBand SDR on all nodes
  - 32-processor (Donation by Appro, Mellanox and Intel)
    - 16 dual Intel EM64T blades
    - InfiniBand SDR and DDR on all nodes
  - 28-processor (Donation by Sun)
    - 12 dual AMD Opteron
    - 1 quad AMD Opteron
  - 32-processor
    - 16 dual-Pentium 300 MHz nodes
    - Myrinet, Gigabit Ethernet, and GigaNet cLAN
  - 96-processor
    - 16 quad-Pentium 700 MHz nodes
    - 16 dual-Pentium 1.0 GHz nodes
    - Myrinet and GigaNet cLAN
  - 32-processor
    - 8 dual-Xeon 2.4 GHz nodes
    - 8 dual-Xeon 3.0 GHz nodes
    - InfiniBand, Quadrics, Ammasso (1GigE), Level 5(1GigE), Chelsio (10GigE), Neteffect (10GigE)
  - Many other donated systems from Intel, AMD, Apple, IBM, SUN, Microway and QLogic
- Other test systems and desktops
  - 18 dual GHz nodes

NBCLab Clusters

High-End Computing and Networking Research Testbed for Next Generation Data Driven, Interactive Applications

PI: D. K. Panda
Co-PIs: G. Agrawal, P. Sadayappan, J. Saltz and H.-W. Shen

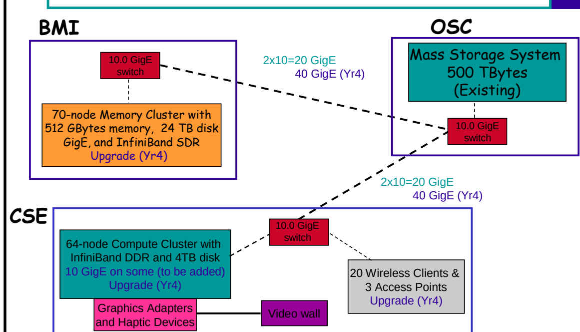
Other Investigators: S. Ahalt, U. Catalyurek, H. Ferhatosmanoglu, H.-W. Jin, T. Kurc, M. Lauria, D. Lee, R. Machiraju, S. Parthasarathy, P. Sinha, D. Stredney, A. E. Stutz, and P. Wyckoff

Dept. of Computer Science and Engineering, Dept. of Biomedical Informatics, and Ohio Supercomputer Center

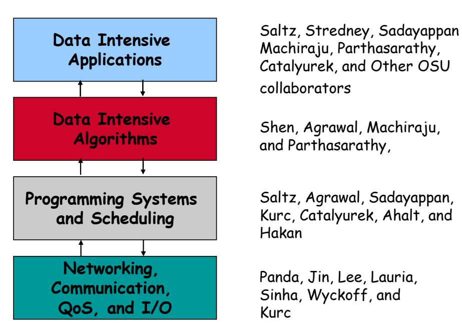
The Ohio State University

Total funding: $3.01M ($1.53M from NSF + $1.48 from Ohio State Board of Regents and Various units in OSU)

Experimental Testbed Installed



Collaboration among the Components and Investigators



Wright Center for Innovation (WCI)

- A new research equipment funding ($600K) has enabled us to install the following cluster:
  - 64 computing nodes with Intel dual quad-core processors (512 cores)
  - 8 storage nodes with Intel dual quad-core processors (64 cores)
  - A total of 576 cores
  - Storage nodes with 32 TBytes of space
  - Connected with InfiniBand DDR (20 Gbps)
- Focuses on Research on Advanced Data Management

Presentation Outline

- Motivation
- Technological Trends
- Applications (Past, Current, and Emerging)
- Defining Network-Based Computing Systems
- Challenges and Research Issues
- Research Challenges and Projects
- Available Experimental Testbed
- Related 888 Course
- Conclusions

888.08P course

- Research Seminar on Network-Based Computing
- Offered every quarter
- This quarter (Wi '08) offering
- Presentation and Project-based course
- 3-credits
- Two-quarter rule

Presentation Outline

- Motivation
- Technological Trends
- Applications (Past, Current, and Emerging)
- Defining Network-Based Computing Systems
- Challenges and Research Issues
- Research Challenges and Projects
- Available Experimental Testbed
- Related 888 Course
- Conclusions

Conclusions

- Network-Based computing is an emerging trend
- Small to medium scale network-based computing systems are getting ready to be used
- Many open research issues need to be solved before building large-scale network-based computing systems
- Will be able to provide affordable, portable, scalable, and ..able computing paradigm for the next 10-15 years
- Exciting area of research

Web Pointers

NBC **home page**

http://www.cse.ohio-state.edu/~panda/
http://nowlab.cse.ohio-state.edu/

E-mail: panda@cse.ohio-state.edu

Acknowledgements

Our research is supported by the following organizations

- Current Funding support by
- Current Equipment support by