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

- 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 6429 Course
- Conclusions
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

- 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 6429 Course
- Conclusions
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 (0.5/2.0 MBytes/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 Multi-core Processors
- Accelerators (GPGPUs and Intel MIC)
- Solid State Devices (SSDs)
- Emergence of the Web Technology
- Emergence of Virtualization Technology
Network Speed Acceleration with IB and HSE

<table>
<thead>
<tr>
<th>Network Type</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet (1979 -)</td>
<td>10 Mbit/sec</td>
</tr>
<tr>
<td>Fast Ethernet (1993 -)</td>
<td>100 Mbit/sec</td>
</tr>
<tr>
<td>Gigabit Ethernet (1995 -)</td>
<td>1000 Mbit/sec</td>
</tr>
<tr>
<td>ATM (1995 -)</td>
<td>155/622/1024 Mbit/sec</td>
</tr>
<tr>
<td>Myrinet (1993 -)</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>Fibre Channel (1994 -)</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2001 -)</td>
<td>2 Gbit/sec (1X SDR)</td>
</tr>
<tr>
<td>10-Gigabit Ethernet (2001 -)</td>
<td>10 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2003 -)</td>
<td>8 Gbit/sec (4X SDR)</td>
</tr>
<tr>
<td>InfiniBand (2005 -)</td>
<td>16 Gbit/sec (4X DDR)</td>
</tr>
<tr>
<td></td>
<td>24 Gbit/sec (12X SDR)</td>
</tr>
<tr>
<td>InfiniBand (2007 -)</td>
<td>32 Gbit/sec (4X QDR)</td>
</tr>
<tr>
<td>40-Gigabit Ethernet (2010 -)</td>
<td>40 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2011 -)</td>
<td>54.6 Gbit/sec (4X FDR)</td>
</tr>
<tr>
<td>InfiniBand (2012 -)</td>
<td>2 x 54.6 Gbit/sec (4X Dual-FDR)</td>
</tr>
<tr>
<td>InfiniBand (2014?)</td>
<td>100 Gbit/sec (4X EDR)</td>
</tr>
</tbody>
</table>

50 times in the last 12 years

Powerful Multicore Processors

- Processor speed was doubling every 18 months (since early 90s to 2002/2003)
- Design and development of powerful and cost-effective uni-processor/multi-processor systems
- Speed could not be enhanced any more due to power dissipation
- Led to multi-core era
- 8/12/16-core processors are common today
- 64/128-core processors will be common in another 2-3 years
- System on Chip (SOC) like Tilera already have 72 cores in a single chip
- Up to 512/1024-core commodity processors in the next 5 years
Accelerators

- Accelerators are emerging in full-speed
- Started with nVIDIA, IBM Cell, etc.
- Intel Many-Integrated Cores (MIC) is moving ahead
- CUDA and GPGPU framework are providing software integration with the accelerators

Solid State Devices (SSDs)

- Magnetic disks have been the main component for permanent storage
- Capacity has been growing, but not speed
- Performance difference between memory and disk speed steadily increasing
- New Solid State Devices (SSDs) are built with flash memory technology
- Very fast compared to magnetic disk
- Helping to reduce the gap between memory and storage
- Some SSDs (Fusion I/O) can operate on full networking speed
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 17 years

Emergence of Virtualization Technology

- Virtualization technology to run multiple processes on the same processor
- Virtualization in networking technology (SR-IOV) is also emerging
- Lot of potential in datacenter environment to deliver performance, scalability and fault-tolerance with the lowest cost
Summary of Technology Trends

- Network has been ignored as an integral component of a computing box in the past
- Six important developments in the last decade
  - commodity and affordable networking technologies
  - powerful multi-core processors
  - Accelerators
  - Solid State Devices (SSDs)
  - emergence of the Web technology
  - emergence of Virtualization 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 6429 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/Big Data 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

Challenges in Designing Systems for Emerging 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 (multi-core and accelerators)
  – Networking
  – Storage (including SSDs)
  – Web
  – Virtualization
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 6429 Course
- Conclusions

Network-Based Computing

- powerful
- cost-effective
- commodity

Networking
- high bandwidth
- cost-effective
- commodity

Web/Grid/Cloud Technology
- powerful
- ability to integrate computation and communication

Emerging Applications
- collaborative/interactive
- wide range of computation and communication characteristics

Storage Technology (SSD)

Accelerators (GPGPUs and MIC)
High-End Computing (HEC): PetaFlop to ExaFlop

Dedicated Cluster over a SAN/LAN

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

Expected to have an Exaflop system in 2019-2022!
Trends for Commodity Computing Clusters in the Top 500 List (http://www.top500.org)

![Graph showing trends for commodity computing clusters]

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

![Diagram of clustered file systems]

Compute Node
Meta-Data Manager
I/O Server Node
I/O Server Node
I/O Server Node
Meta Data
Data
Data
Data
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 High-End Computing Environments

Enterprise Multi-tier Datacenter for Visualization and Mining
Cloud Computing Environments

Big Data Analytics with Hadoop

- **Underlying** Hadoop Distributed File System (HDFS)
- Fault-tolerance by replicating data blocks
- NameNode: stores information on data blocks
- DataNodes: store blocks and host Map-reduce computation
- JobTracker: track jobs and detect failure
- MapReduce (Distributed Computation)
- HBase (Database component)
- Model scales but high amount of communication during intermediate phases
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 6429 Course
- Conclusions

Networking and I/O Requirements

- Good System Area Networks with excellent performance (low latency, high bandwidth and low CPU utilization) for inter-processor 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 interactive applications
- RAS (Reliability, Availability, and Serviceability)
- With low cost
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
  – Designing and Developing better middleware/API/programming environments so that modern HPC and Big Data applications can be developed and implemented in a scalable fashion
  – Proposing guidelines for designing next generation networks and protocols

• Carry out research in an integrated manner (systems, networking, and applications)

Research Focus

Application Kernels/Applications

Middleware

Programming Models
  MPI, PGAS (UPC, Global Arrays, OpenSHMEM), CUDA, OpenACC, Cilk, Hadoop, MapReduce, etc.

Communication Library or Runtime for Programming Models
  Point-to-point Communication (two-sided & one-sided)
  Collective Communication
  Synchronization & Locks
  I/O & File Systems
  Fault Tolerance
  Virtualization

Networking Technologies
  (InfiniBand, 40/100GigE, Aries, BlueGene)

Multi/Many-core Architectures

Accelerators (NVIDIA and MIC)
Major Research Directions

- High Performance MPI for InfiniBand, 10GigE/iWARP and RoCE
- Programming Models (MPI+PGAS (UPC and OpenSHMEM)) for Exascale Systems
- High Performance Computing with GPGPUs and MIC
- Big Data Analytics with Hadoop
- Networking, Storage and Middleware Support for Cloud Computing and Datacenters
- Scalable and High-Performance File Systems and Storage
- Virtualization (High Performance Computing and Cloud Computing)
- WAN-IB support for Grid and Cloud Computing
- QoS-aware Designs
- Power-aware Designs
- Fault-Tolerant Designs
- More details on http://nowlab.cse.ohio-state.edu/ → Projects

MVAPICH2/MVAPICH2-X Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002
  - MAPlCH2-X (MPI + PGAS), Available since 2012
  - Support for GPGPUs (MVAPICH2-GDR) and MIC
  - Used by more than 2,100 organizations (HPC Centers, Industry and Universities) in 71 countries
  - More than 201,000 downloads from OSU site directly

- Empowering many TOP500 clusters
  - 7th ranked 462,462-core cluster (Stampede) at TACC
  - 11th ranked 74,358-core cluster (Tsubame 2.5) at Tokyo Institute of Technology
  - 16th ranked 96,192-core cluster (Pleiades) at NASA and many others

- Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu

- Partner in the U.S. NSF-TACC Stampede System
**RDMA for Apache Hadoop Project**

- High-Performance Design of Hadoop over RDMA-enabled Interconnects
  - High performance design with native InfiniBand and RoCE support at the verbs-level for HDFS, MapReduce, and RPC components
  - Easily configurable for native InfiniBand, RoCE and the traditional sockets-based support (Ethernet and InfiniBand with IPoIB)
- Current release: 0.9.8
  - Based on Apache Hadoop 1.2.1
  - Compliant with Apache Hadoop 1.2.1 APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR and FDR)
    - RoCE
    - Various multi-core platforms
    - Different file systems with disks and SSDs
  - [http://hadoop-rdma.cse.ohio-state.edu](http://hadoop-rdma.cse.ohio-state.edu)

---

**External Collaboration**

- **Industry**
  - IBM TJ Watson and IBM
  - Intel
  - Cisco
  - Cray
  - Dell
  - SUN Microsystems/Oracle
  - Apple
  - Linux Networx
  - NetApp
  - NVIDIA
  - Mellanox, Qlogic and other InfiniBand companies
  - Chelsio, NetEffect and other iWARP companies
- **Research Labs**
  - Sandia, Los Alamos, Pacific Northwest, Argonne, Lawrence Livermore National Labs
- **Supercomputing Centers**
  - Ohio Supercomputer Center (OSC)
  - Texas Advanced Computing Center (TACC)
  - San Diego Supercomputer Center (SDSC)
Research Funding

- US National Science Foundation
  - Multiple grants
- US 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
- NVIDIA
- Cisco
- Linux Networx
- Network Appliance

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 6429 Course
- Conclusions
NBCLab Clusters

Latest InfiniBand Cluster
(Deployed in August '10)

- A new research equipment funding ($900K) has enabled us to install the following cluster:
  - 1,400-core Intel quad-core Westmere cluster with 32TBytes of disks, connected with InfiniBand QDR (40 Gbps)
  - 8 GPU nodes with latest Nvidia Fermi adapters
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 6429 Course
- Conclusions

6429 Course

- Research Seminar on Network-Based Computing
- Offered every semester
- Presentation and Project-based course
- 3-credits
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 6429 Course
• Conclusions

Conclusions

• Network-Based computing is an emerging trend
• Medium to large-scale network-based computing systems are getting ready to be used
• Many open research issues need to be solved before building ultra-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