COMPUTER & COMPUTATIONAL SCIENCES

Los Alamos National Laboratory

RADIANT

www.lanl.gov/radiant

RADIANT: Research & Development in Advanced Network Technology

Wu-chun Feng, Team Leader

feng@lanl.gov

Team Members: Sami Ayyorgun, Adam Englehart, Mark Gardner, Justin (Gus) Hurwitz, Eric Weigle.

Computer and Computational Sciences Division
http://www.ccs.lanl.gov
Voice: 505-665-2730
Fax: 505-665-4934

LA-UR-03-3941

NNSA

Los Alamos NATIONAL LABORATORY

What We Do

- Network Research
  - High-Performance Networking
  - Monitoring and Measurement (Systems & Networks)
  - Network Traffic Characterization

- Network-Related Research
  - Systems Support for High-Performance Computing
  - Dataflow Grid
  - Bioinformatics
  - Cyber-Security

Network Research

- High-Performance Networking
  - Hardware
    - GigE, 10GigE, Quadrics, InfiniBand, Optical / DWDM / λ-Switching.
  - Software
    - OS-Bypass Protocols / Remote Direct-Memory Access
    - High-Perf. TCP: TCP Off-Load Engines & Dynamic Right-Sizing.
- Monitoring and Measurement (Systems & Networks)
  - MAGNET+MUSE †: Software Oscilloscope for Clusters & Grids
  - TICKET *: Scalable Network Measurement w/ Commodity Parts
- Network Traffic Characterization
  - Traffic modeling to gain insight into the hardware and software design of network components.

† MAGNET: Monitoring Apparatus for General kerNel-Event Tracing
MUSE: MAGNET User-Space Environment
* TICKET: Traffic Information-Collecting Kernel with Exact Timing

Network-Related Research

- Systems Support for High-Performance Computing
  - Supercomputing in Small Spaces
    - Green Destiny: A 240-Node Cluster in One Cubic Meter (i.e., a standard computer rack).
- Dataflow Grid
  - The network *is* the computer rather than simply the fabric to interconnect computing nodes.
- Bioinformatics
  - mpiBLAST: An Open-Source Parallelization of BLAST †
- Cyber-Security
  - IRIS: Inter-Realm Infrastructure for Security

† BLAST: Basic Local Alignment Sequence Tool

High-Performance Networking: Achievements

We "own" (or are a part of) the fastest *end-to-end* networking speed records in the LAN, SAN, and WAN.

Network Environments

- LAN, i.e., Ethernet + IP + TCP + ftp
  - Throughput: 4-5 Gb/s. Latency: 20 µs.
    - Achieved at LANL in Oct. 2002. *IEEE Hot Interconnects,* 8/2003.
- SAN, i.e., Quadrics/InfiniBand + OS-bypass + src routing + MPI
  - Throughput: 6-7 Gb/s. Latency: 5 µs.
    - Quadrics: Achieved at LANL in Nov. 2000. *IEEE Micro*, 1/2002.
    - InfiniBand: Achieved at OSU in Oct. 2002. Not yet published.
- WAN, i.e., Ethernet/DWDM + IP + TCP + ftp
  - Throughput: 2-3 Gb/s. Latency: 90 ms transoceanic.
    - Achieved between California and Switzerland in Feb. 2003 and broke the previous Internet2 Land Speed Record (I2 LSR) by 125%.

High-Performance Networking: Achievements for I2 LSR

- Breaking the Internet2 Land Speed Records (I2 LSR)
  - 2.38 Gb/s with a single TCP/IP stream between Sunnyvale, California and Geneva, Switzerland on Feb. 27, 2003. (Equivalently, 23,888,060,000,000,000 meters-bits/second.) Certified Mar. 27, 2003. Awarded formally Apr. 11, 2003.
  - 2.38 Gb/s doubles as the multiple TCP/IP stream record also.
  - *First* time ever that the single-stream gigabit-per-second barrier is broken over a TCP/IP-based WAN.

High-Performance Networking: I2 LSR Partners

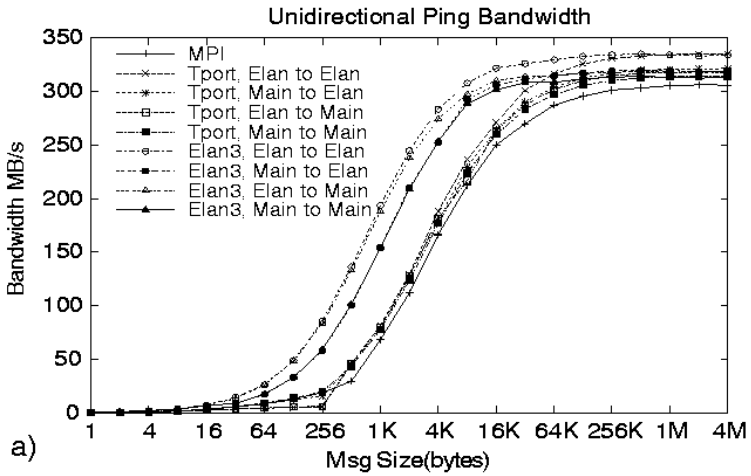
- Institutions
  - California Institute of Technology, CERN, Los Alamos National Laboratory, and Stanford Linear Accelerator Center
- Supporters and Infrastructure Enablers
  - Cisco Systems
  - DataTAG
  - Deutsche Telekom
  - Intel
  - Juniper
  - Level(3) Communications
  - Starlight
  - TeraGrid
- Funding Agencies
  - Department of Energy
  - National Science Foundation
  - European Commission

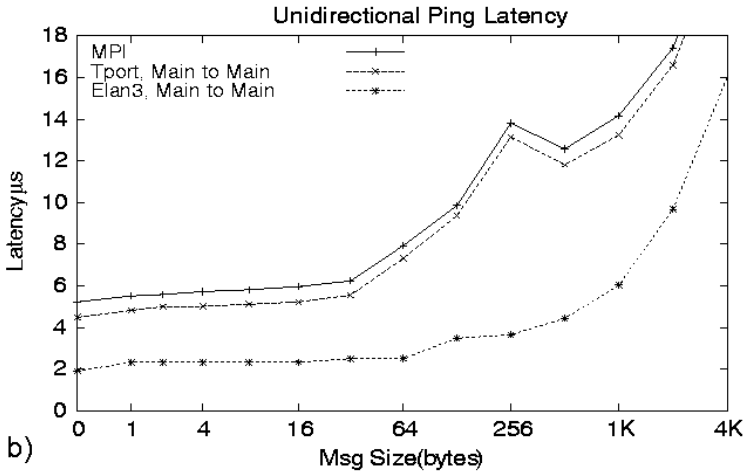
High-Performance Networking: I2 LSR Media Coverage & Publications

- Media Coverage
  - *The Register*, June 6, 2003.
    - "Data Speed Record Crushed," http://theregister.com/content/5/31085.html.
  - *TechTV*, May 20, 2003.
  - *Nature*, Mar. 27, 2003.
  - *LightReading*, Mar. 21, 2003.
  - *InfoWorld*, Mar. 17, 2003.
  - *Network World Fusion*, Mar. 17, 2003.
  - *ITWorld*, Mar. 17, 2003.
  - *IDG*, Mar. 17, 2003.
- Publications
  - "Initial End-to-End Performance Evaluation of 10-Gigabit Ethernet," To appear in *IEEE Hot Interconnects*, Aug. 2003.
  - "Optimizing 10-Gigabit Ethernet for Network of Workstations, Clusters, and Grids: A Case Study," Submitted to IEEE/ACM SC 2003, Nov. 2003.

High-Performance Networking: SANs and WANs

- Quadrics SAN. Nov. 2000.
  - MPI-to-MPI
    - Latency: 4.9 us.
    - Throughput: 2.456 Gb/s.
  - 50% better performance than Myrinet.
- Dynamic Right-Sizing in the WAN
  - As much as 30-fold improvement in TCP/IP performance while remaining TCP-friendly.
  - Early Adopters & Testers
    - NSF Web100, UCSD/SDSC, USAF.
- TCP Compatibility
  - Better performance & compatibility with TCP Vegas than with ubiquitous TCP Reno.





High-Performance Networking: Recent Publications

- "Automatic Flow-Control Adaptation for Enhancing Network Performance in Computational Grids," *Journal of Grid Computing*, 2003.
- "User-Space Auto-Tuning for TCP Flow Control in Computational Grids," *Computer Communications*, 2003.
- "Initial End-to-End Performance Evaluation of 10-Gigabit Ethernet," *IEEE Hot Interconnects*, Aug. 2003.
- "Ensuring Compatibility Between TCP Reno and TCP Vegas," *IEEE Symposium on Applications and the Internet (SAINT'03)*, Jan. 2003.
- "Dynamic Right-Sizing: An Automated, Lightweight, and Scalable Technique for Enhancing Grid Performance," *Lecture Notes in Computer Science*, 2002.
- "On the Transient Behavior of TCP Vegas," *IEEE International Conference on Computer Communications and Networks (IC3N'02)*, Oct. 2002.
- "Packet Spacing: An Enabling Mechanism for the Delivery of Multimedia Content," *Journal of Supercomputing*, Vol. 23, No. 1, Aug. 2002.
- "Dynamic Right-Sizing in FTP (drsFTP): An Automatic Technique for Enhancing Grid Performance," *IEEE Symposium on High-Performance Distributed Computing (HPDC'02)*, Jul. 2002.
- "A Comparison of TCP Automatic-Tuning Techniques for Distributed Computing," *IEEE Symposium on High-Performance Distributed Computing (HPDC'02)*, Jul. 2002.
- "Dynamic Right-Sizing in TCP: A Simulation Study," *IEEE International Conference on Computer Communications and Networks (IC3N'01)*, Oct. 2001.
- "The Future of High-Performance Networking," *Workshop on New Visions for Large-Sacle Networks: Research & Applications*, Invited Paper, Mar. 2001. (Sponsors: Federal Large-Scale Networking Working Group, DARPA, DOE, NASA, NIST, NLM, and NSF).

Monitoring & Measurement

- MAGNET + MUSE:
  Software Oscilloscope for Clusters and Grids
  - MAGNET: Monitoring Apparatus for General kerNel-Event Tracing
  - MUSE: MAGNET User-Space Environment
  - (Formerly, MAGNeT: Monitor for Application-Generated Network Traffic)

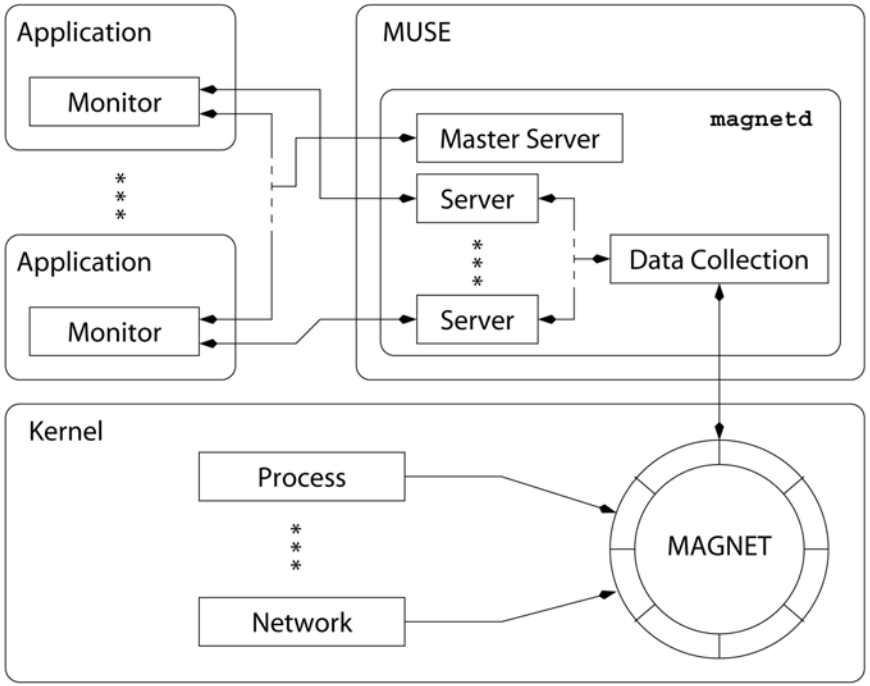
- TICKET
  - Traffic Information Collecting-Kernel with Exact Timing

Monitoring and Measurement: MAGNeT → MAGNET + MUSE

- Problem:
  - Network researchers only analyze traffic *after* it has been (adversely) modulated by the TCP/IP protocol stack.
- Solution:
  MAGNeT: Monitor for Application-Generated Network Traffic
  - Goals
    - Monitor traffic immediately after being generated by the application (i.e., unmodulated traffic) and throughout the protocol stack to see how traffic gets modulated. Is TCP/IP the obstacle to high performance?
    - Create a library of application-generated network traces to test network protocols. Why? Networking is not only FTP.
  - Feedback from *2002 Passive & Active Measurem't Workshop*
    - Why not extend monitoring to kernel events in general? That is, MAGNeT → MAGNET + MUSE. See next slide.

MAGNET + MUSE: Software Oscilloscope for Clusters & Grids

- MAGNET: Monitoring Apparatus for General kerNel-Event Tracing
  - Goals
    - Easily debug, tune, or optimize system software and parallel apps, i.e., debugging tool.
      - e.g., Identified a Linux SMP-scheduling anomaly.
    - Monitor the state of a parallel machine, i.e., diagnostics tool.



- MUSE: MAGNET User-Space Environment
  - Goal: Enable adaptive (i.e., resource-aware) apps, i.e., feedback tool, particularly for distributed grid apps.
    - Efficiently export kernel-level information to user space.
    - Provide a standard protocol by which resource-aware apps obtain access to exported kernel-space information.

Monitoring & Measurement: Recent Publications

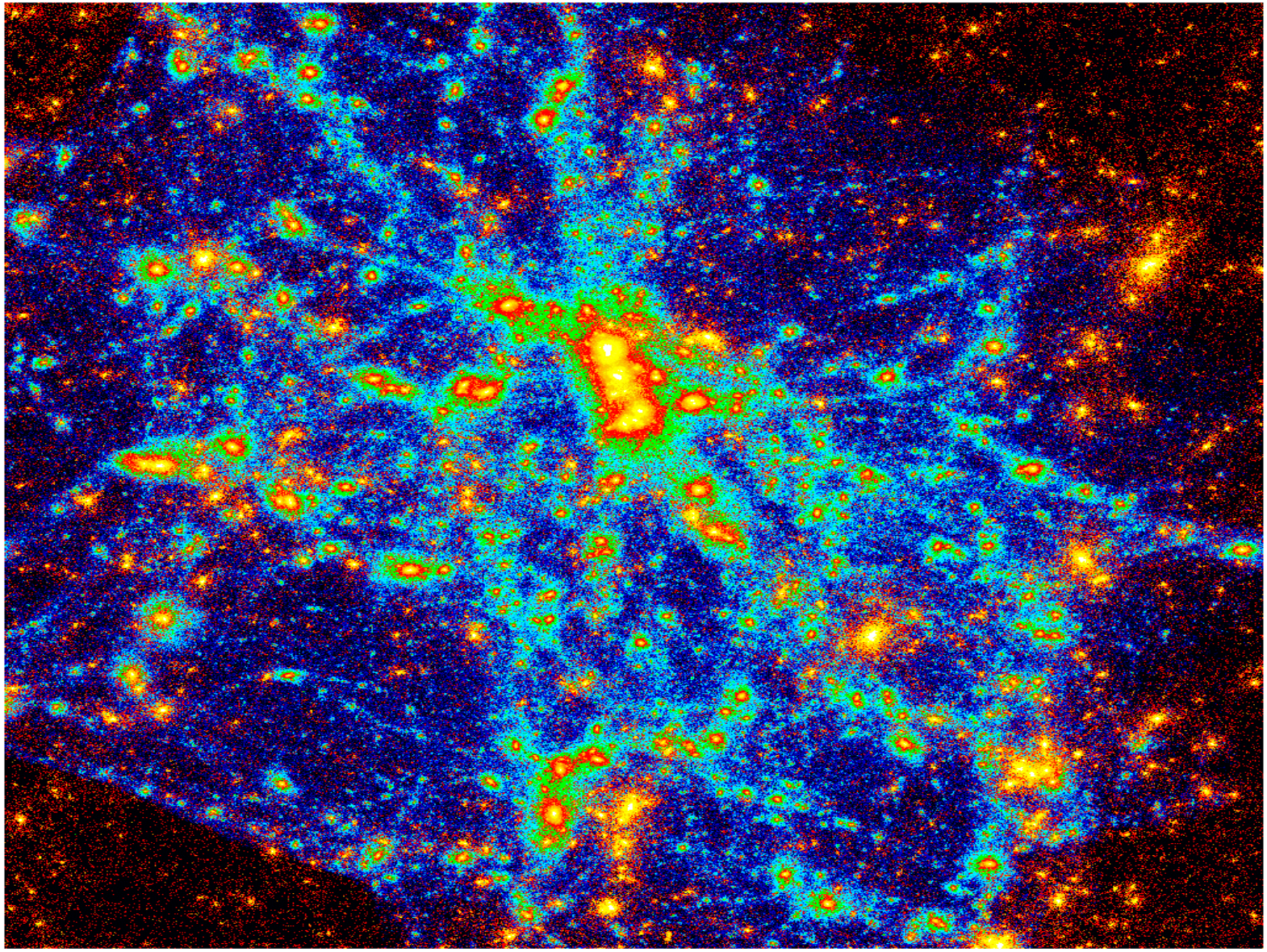
- "Online Monitoring of Computing Systems with MAGNET," *IEEE/ACM Symposium on Cluster Computing and the Grid (CCGrid'03)*, May 2003.
- "MUSE: A Software Oscilloscope for Clusters and Grids," *IEEE Parallel & Distributed Processing Symposium*, Apr. 2003.
- "The MAGNeT Toolkit: Design, Evaluation, and Implementation," *Journal of Supercomputing*, Vol. 23, No. 1, Aug. 2002.
- "Monitoring Protocol Traffic with a MAGNeT," *Passive & Active Measurement (PAM) Workshop*, Mar. 2002.
- "TICKETing High-Speed Traffic with Commodity Hardware and Software," *Passive & Active Measurement Workshop*, Mar. 2002.
- "MAGNeT: Monitor for Application-Generated Network Traffic," *IEEE Int'l Conf. on Computer Communications and Networks (IC3N'01)*, Oct. 2001.

Supercomputing in Small Spaces: "Green Destiny" Bladed Beowulf

- A 240-Node Beowulf in One Cubic Meter
- Each Node
  - 667-MHz Transmeta TM5600 CPU
    - Recently upgraded to 1-GHz Transmeta TM5800.
  - 640-MB RAM
  - 20-GB hard disk
  - 100-Mb/s Ethernet
- Overall System
  - 240 nodes
  - 150-GB RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)
  - Topology: One-level tree.

Wu-chun Feng
feng@lanl.gov

Intermediate Stage of a Gravitational N-body Simulation Performed on Green Destiny.
(10 Million Particles, 1000 timesteps, 1015 floating-point ops)



150-Million Light Years Across

Supercomputing in Small Spaces: Media Coverage

- Over 60 articles written about *Green Destiny*, e.g.,
  - "Servers on the Edge: Blades Promise Efficiency and Cost Savings," *CIO Magazine*, Mar. 15, 2003.
  - "Developments to Watch: Innovations," *BusinessWeek*, Dec. 2, 2002.
  - "Not Your Average Supercomputer," *Communications of the ACM*, Aug. 2002.
  - "Green Destiny Runs Cool," *Dr. Dobb's Journal*, Aug. 2002.
  - "Competing Visions of Supercomputing," *International Herald Tribune*, Jun. 26, 2002.
  - "At Los Alamos, Two Visions of Supercomputing," *The New York Times*, Jun. 25, 2002.
  - "Supercomputing Coming to a Closet Near You?" *HPCwire*, May 31, 2002.
  - "Smaller, Slower Supercomputers May Someday Win The Race," *HPCwire*, May 31, 2002.
  - "Supercomputing Coming to a Closet Near You?" *PCWorld.com*, May 27, 2002.
  - "Bell, Torvalds Usher Next Wave of Supercomputing," *CNN.com*, May 21, 2002.
  - "Transmeta's Low Power Finds Place in Supercomputers," *ZDNet*, May 20, 2002.

Supercomputing in Small Spaces: Publications and Talks

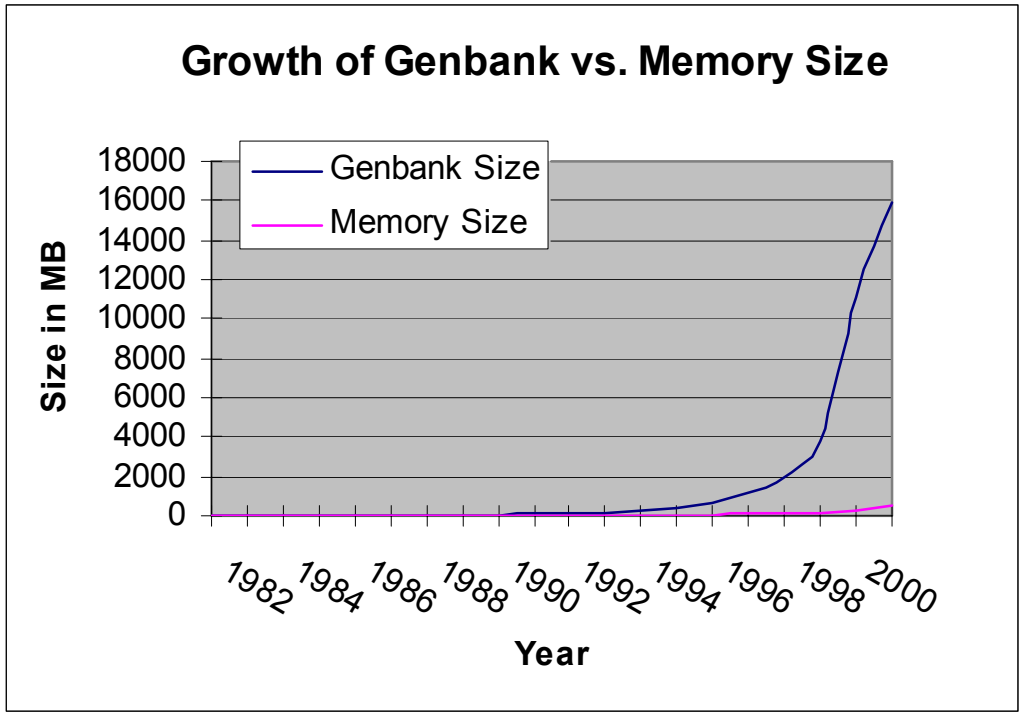
- Publications
  - "High-Density Computing: A 240-Node Beowulf in One Cubic Meter," *SC 2002: High-Performance Networking and Computing Conference*, November 2002.
  - "The Bladed Beowulf: A Cost-Effective Alternative to Traditional Beowulfs," *IEEE Cluster 2002*, September 2002.
  - "Honey, I Shrunk the Beowulf!" *31st International Conference on Parallel Processing (ICPP'02)*, August 2002.
- Invited & Keynote Talks
  - *Future Computing Conference* at the Royal United Services Institute for Defence and Security Studies, Jul. 2003.
  - *Server Blade Summit*, Mar. 2003.
  - *Rocky Mountain Institute Data Center Charrette*, Feb. 2003.
  - *15th Annual E-Source Forum*, Nov. 2002.
  - *University of Illinois at Urbana-Champaign (also broadcast over the Internet via the Access Grid)*, Oct. 2003.
  - *Eli Lilly and Company*, Sept. 2002.

Bioinformatics: Parallelizing BLAST

- Multithreading
  - Implemented in NCBI's BLAST.
- Query Segmentation
  - Divides a query into sub-queries and each sub-query is searched against a copy of the entire database on each node.
  - Many implementations exist.
- Database Segmentation
  - Fragments the database into smaller pieces where each piece fits entirely in memory. Each cluster node searches on one fragment of the database.
  - Only known open-source implementation: mpiBLAST.

Bioinformatics: Enormous Sequence Databases

Size in MB	DB name	Description
5700	nt	non-redundant nucleotide DB
2200	Human EST	Human expressed sequence tag DB
1100	Mouse EST	Human expressed sequence tag DB
510	nr	non-redundant amino acid DB



Growth Trend: Database Size vs. Memory Size

Bioinformatics: mpiBLAST Performance

BLAST Run Time for a 300kb Query against nt :

Nodes	Runtime (s)	Speedup over 1 node	Speedup / Nodes ratio
1	80774.93	1.00	1.00
4	8751.97	9.23	2.31
8	4547.83	17.76	2.22
16	2436.60	33.15	2.07
32	1349.92	59.84	1.87
64	850.75	94.95	1.48
128	473.79	170.49	1.33

Reduces search time …

From over 1346 minutes (22.4 hours) to under 8 minutes!

Bioinformatics: mpiBLAST Publications

- The Design, Implementation, and Evaluation of mpiBLAST," Best Paper Award, *ClusterWorld Conference & Expo 2003*, June 2003.
- mpiBLAST: Delivering Super-Linear Speedup with an Open-Source Parallelization of BLAST" (poster), *Pacific Symposium on Biocomputing (PSB'03)*, January 2003.