

NASA CIPA Cluster Computing Project



Douglas E. Salane

Mathematics Dept.

John Jay College of Criminal Justice
The City University of New York



NASA Research Summit
July 18, 2003

Project Title

Computer Clusters to Support Curricular Improvements in Computer Networking and Parallel/Distributed Computing

NASA UNCFSP
Curriculum Improvement Partnership Award

Institutional Background John Jay College/CUNY

- **College:** Specialized Liberal Arts College within CUNY (12,500 students including 1300 graduate students).
- **Degrees:** Law and Police Science, Public Management, Fire Science, Security, Forensic Science, Computer Information Systems, M.S. in Forensic Computing (2004), Ph.D. in Criminal Justice.
- **Mission:** Advance the practice of criminal justice and public administration through research and providing a professional workforce.

CIS Major at John Jay College

- **Mathematics foundation:** Calculus, Discrete Mathematics and Operations Research.
- **Core courses:** Basic computer science (programming fundamentals, algorithms and complexity, operating systems).
- **Capstone courses:** Systems areas of computing and applications of interest to Criminal Justice and Law Enforcement Agencies (computer networking, database systems, distributed systems, security, forensic analysis).

CIS Students

- 513 CIS Majors
- 44% Female
- 68% Day
- 32% Evening
- 30% African American
- 32% Hispanic
- 12% Asian
- 12% White
- 14% Other

Cluster Computing Project

- **Goal:** Ensure computer majors are prepared for challenging careers and/or graduate studies in computer networking and parallel/distributed computing.
- **Prerequisite:** Faculty/staff, computing facilities and curriculum development.
- **Mechanism:** NASA developed research and technologies in high performance computing and computer networking (GISS, GSFC, NAS at AMES, JPL), especially in cluster computing.

Computer Networking and Distributed Computing at John Jay College

- Law enforcement and criminal justice agencies require effective distributed systems
- Investigate abuse and misuse of computers and computer networks
- Security of distributed information systems
- Data mining techniques in criminal investigations and criminal justice research
- Highly available, high throughput database systems for emergency response
- Expertise in cryptography and cryptographic security protocols to thwart and prosecute computer and cyber crimes
- Protect national information infrastructure and privacy of individuals

High Performance Computing at John Jay College

- Standards and codes for buildings (Diffusion-Convection models, domain decomposition)
- Analysis of Large Data Sets (Principal Component Analysis – SVD)
- Toxicology (molecular modeling – Gauss)
- Modeling transportation systems
- Aircraft control systems (Parallel Schur Decomposition)
- Research and course work in parallel algorithms
- Computational Number Theory

Development Areas

- **Computer Networking** – Involves communication protocols at or below the Transport Level (TCP, IP, Data/Link layer protocols, IPSec)
- **Distributed Computing** – Involves protocols above the transport level (HTTP, SOAP, XML, CORBA, Grid Technologies, SSL, Web/Database Technologies)
- **Parallel Computing** – A type of distributed computing, goal is solution of computationally intensive problems (High speed interconnects, MPI, PVM, OpenMP, ScaLAPACK).

Importance of Cluster Computing

- High throughput and availability (e.g. Google Search Engine, storage systems, database systems).
- Supercomputer performance for large-scale computation (e.g. CFD, molecular modeling, graphics and animation).
- Use latest developments in computer networking and parallel/distributed computing (Infiniband).
- Rely on developments in OS (e.g. SSI, MOSIX, PBS, MAUI, GFS, Linux), Discrete Mathematics (e.g. graph partitioning).
- Lab environment for parallel computing and distributed algorithms.

NASA Technologies and Research

- **Commodity Cluster Computing:** Sterling and Becker at GSFC ('95), The original work that spawned an industry. NAS at AMES, JPL
- **Portable Batch Systems (PBS):** A scheduler for clusters. NAS at AMES
- **Information Power Grid (Globus Toolkit):** Integrate a widely distributed and diverse set of computing resources into a secure, useful system. NAS at AMES, Argonne National Labs
- **Climate Modeling:** Open MP in clusters. GISS.

Topics added or coverage expanded

- Amdahl's Law
- TCP/IP
- cache coherence
- Enterprise-level database systems
- client/server
- distributed File Systems
- SIMD/MIMD models
- graph theory, Graph Algorithms
- multithreading
- multiprogramming
- availability/throughput
- SMP scheduling
- distributed shared memory
- computer clusters
- web services and technologies
- queuing analysis
- IPC & synchronization
- packet vs circuit switching
- parallel algorithms (searching)
- Distributed algorithms
- load balancing
- high performance linear algebra software
- security & cryptography

Curriculum Development

- **New courses:** Artificial Intelligence, Computer Networking, Discrete Mathematics, Graphics and Systems Analysis.
- **New mathematics requirement:** Includes Discrete Mathematics and advanced OR courses (graph theory and network queuing analysis).
- **Revised OS curriculum:** Provides broader, deeper, hands-on coverage of IPC, synchronization, message passing and distributed file systems.

Curriculum Development (continued)

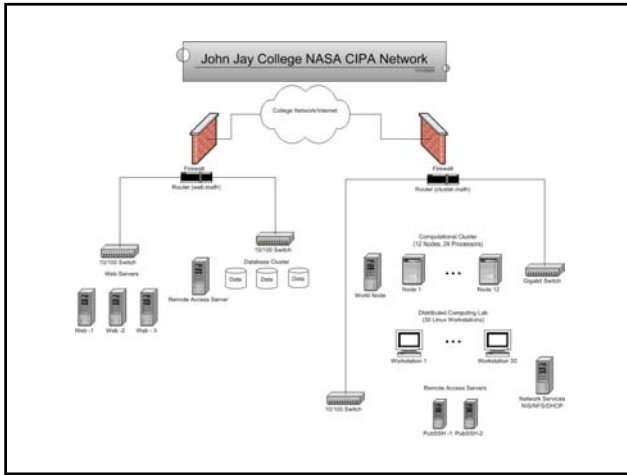
- **Revised senior level database system course:** includes data-mining, enterprise level database, and web-based information systems.
- **Developed six graduate courses:** Networks and Distributed Processing, Network Forensics and Security, and Architecture of Operating Systems.

Facilities Development

- **Computational Cluster (Beowulf Cluster):** 12 nodes, 24 processors, production cluster (MPICH, MPE, PBS, HPL Benchmark, ScaLAPACK, Super LU). [The Computational Cluster](#)
- **Database Cluster:** 4 nodes with remote access server, web server, Microsoft SQL and Oracle 9i.
- **Distributed Computing Laboratory:** Computing Laboratory with 30 Linux Workstations (partnership with Science [Distributed Computing Lab](#))

Facilities Development (Continued)

- **PubSSH:** Two remote login computers for students to access CIPA facilities and accounts. Uses SSH and load balancing.
- **CIPA Network:** Designed and implemented high speed network and supporting services to interconnect computer clusters, distributed computing laboratory, and college network.



Faculty/Staff/Student Development

- **Computer Clusters:** Expertise designing, building and managing computer clusters and networks (Linux OS configuration, PBS, NetPerf, Netpipe, LMBench, NFS, GFS).
- **Cluster Middleware/Applications:** Expertise in software and applications for parallel computing: (MPICH, MPE, BLAS, BLACS, PBLAS, LAPACK, ScaLAPACK).
- **Computational Number Theory:** Developed the Large Number Library for computational clusters and parallel implementation of Miller-Rabin algorithm for primality testing.

Faculty/Staff/Student Development

- **NIBRS:** FBI's National Incident Based Reporting System (4 gigabyte database)
 - Converted NIBRS flat files to relational database
 - Web based interface
- **NIBRS Crime Studies**
 - Crime and Age
 - Incident Time vs. Offense, Population and Crimes
- **NIBRS Work Underway**
 - Adding all years up to 2001 (20 gigabyte database)
 - Portal on College web site
 - Database Cluster for high availability and throughput

Faculty/Staff/Student Development

- **Parallel Algorithms:** Parallel computation of real Schur form of a matrix and visualization techniques.
- **File systems:** Comparison of file systems for use in clusters (NFS, AFS, GFS, PVS).
- **Web Information Systems:** Expense Database and Literature Reference Database. [LRD View](#)
- **Distributed Algorithms:** Distributed Bellman Ford Algorithm in Computational Cluster.
- **Cluster Computing Colloquium Series:** lectures, tutorials and vendor presentations (20).

Impact on CIS Students

- Provided over 350 CIS majors access to NASA cluster and high performance computing technologies.
- 50 graduates: 42 took advanced computer networking course, 50 advanced database course
- 15 NASA/CIPA students: 12 have graduated, all are pursuing M.S. in Computer Science or MIS.

Impact on the Department/College

- Provided the institutional capability to launch an M.S. program in Forensic Computing (distributed information systems and network security).
- Center of expertise in computer networking, distributed computing, high performance computing and database systems.
- Hired two recent Ph.Ds with backgrounds in computer networking and parallel computing.

Future Directions

- Fine tune cluster, add applications & tools: super LU for sparse matrices, pvoke.
- Web portal for NIBRS, distributed databases.
- Utilize problem solving capability to establish additional partnerships.
- Attract additional qualified students to the CIS major.
- Implement M.S. in Forensic Computing

- Douglas E. Salane
 - NASA CIPA Cluster Computing Project
 - dsalane@jjay.cuny.edu
 - web.math.jjay.cuny.edu
- Faculty Positions Available
 - Ph.D in Computer Science or Math required
 - Send e-mail or see web site

Credits

- NASA UNCFSP Curriculum Partnership Improvement Award
- Graduate Research and Technology Initiative of CUNY (01,02,03)
- Open Source and freely available software (Linux, GNU compilers and languages, Apache, PHP, MySQL)

NASA CIPA Players at John Jay

- Doug Salane, PI, computer networks, cluster computing, large-scale computation, distributed computing
- Peter Shenkin, CO-PI, database systems, web services, computer security
- Mythili Mantharam, CO-PI, parallel algorithms, numerical linear algebra
- Boris Bonderenko, Research Assistant
- **Curriculum Development Team**
 - Konstantinos Georgatos, Thurai Kugan, Mythili Mantharam,
 - Doug Salane, Peter Shenkin, Ruli Ye
- **NASA CIPA Students**

– Nankumar Itwar	John Young	Atiqal Mondal
– Uttam Singha	Richard Cumberbatch	
– Gerardo Vasquez	Samra Vlasnovec	
– Kraffins Villicin	Anton Lebedev	
– Tien Ngyuen	Raul Cabrera	
– Ke Tang	Sheik Shamimullah	

Linpack Benchmark Results

	R	Site	CPUs	R _{max}	R _{peak}
Earth Simulator	1	Japan	5,120	35,860	40,960
Linux Network X	3	LLNL	2,304	7,634	11,060
HP Alpha Server	20	NASA GSFC	1392	2,164	2,184
SGI Origin 3000	78	NASA AMES	1,024	852.9	1,209.6
JJ Cluster		John Jay	24	12.4	28.8

R – rank in Top 500 Super Computers list
 R_{max} – Linpack Benchmark (GFLOPS)
 R_{peak} – Theoretical Highest Performance (GFLOPS)