Netlib, NHSE and other Sources

Jack Dongarra

Computer Science Department
University of Tennessee

and

Mathematical Sciences Section
Oak Ridge National Laboratory

(http://www.netlib.org/utk/people/JackDongarra.html)

NHSE

- National HPCC Software Exchange
- NASA funded CRPC project
- Center for Research on Parallel Computation (CRPC)
  - Argonne National Laboratory
  - California Institute of Technology
  - Rice University
  - Syracuse University
  - University of Tennessee
- Uniform interface to distributed HPCC software repositories
- Facilitation of cross-agency and interdisciplinary software reuse
- Material from ASTA, HPCS, and IITA components of the HPCC program
- http://www.netlib.org/nhse/

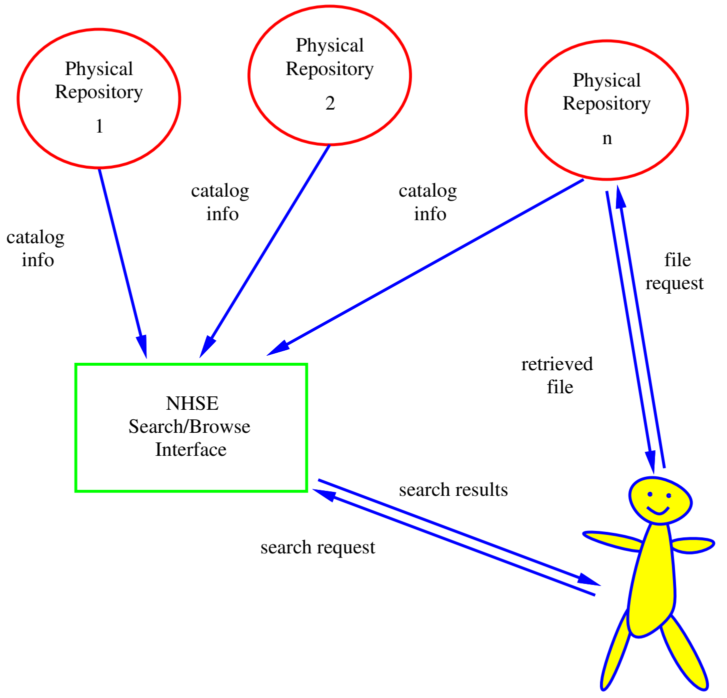
Goals:

- To facilitate an active exchange program for HPCC software and enabling technologies via the National Information Infrastructure.
- To promote contributions and use by Grand Challenge teams, as well as other members of the high performance computing community.

Software here includes algorithms, specifications, designs, documentations, reports...

NHSE Components

- Outreach and technology transition
  - To the HPCC user community and industry
- Distribution via the WWW
- Discipline oriented repositories
- Uniform user interface
- Software review
  - Simple submission
  - On going review
- Measurement
- Hypertext road map
- Publishing tools
  - Repository-in-a-box
- Naming & authentication architecture
- Selective capitalization of emerging technologies



Virtual Repository Architecture

Benefits

1. Faster development of high-quality software so that scientists can spend less time writing and debugging programs and more time on research problems.
2. Less duplication of software development effort by sharing of software modules.
3. Less time and effort spent in locating relevant software and information through the use of appropriate indexing and search mechanisms and domain-specific expert help systems.
4. Reducing information overload through the use of filters and automatic search mechanisms.

Intended Audience

- HPCC application and computer science community
  - Source of material for NHSE
- Users of NASA, NSF, DOE and other supercomputer centers
  - Good targets for NHSE
  - Natural support organization: supercomputer center staff
- Other users of high performance computers
  - Current and potential industrial users
  - No natural support organization
- Applicable to other domains

NHSE

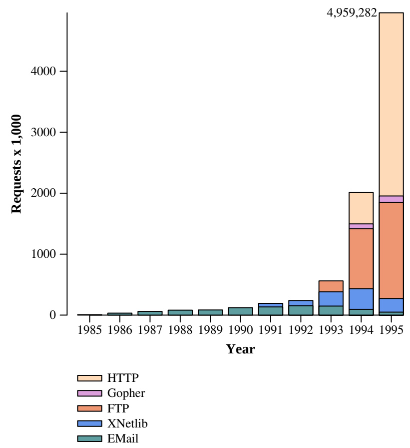
- Based on Existing Technologies
  - WWW Browser (Mosaic / Netscape / etc)
    - Distributed / Scalable
    - URL: http://www.netlib.org/nhse/
  - Netlib
    - Repository for math software since 1985
- Repositories Currently Available
  - Netlib, Softlib, CITlib
  - ASSET - (Asset Source for SW Engineering Tech.)
  - CARDS - (Comprehensive Approach to Reusable Defense SW)
  - ELSA - (Electronic Library Services and Appl.)
  - GAMS (Virtual Software Repository)
  - STARS - (SW Technology for Adaptable, Reliable Systems)
  - Many examples related to GC problems
- Currently Available Information
  - NHSE currently points to 200+ modules
    - software catalog
    - tech reports and papers
    - parallel processing tools
    - libraries of reuseable components
    - Grand Challenge prototype codes

Netlib - Network Access to Mathematical Software and Data

- Began in 1985
  - JD and Eric Grosse, AT&T Bell Labs
- Motivated by the need for cost-effective, timely distribution of high-quality mathematical software to the community.
- Designed to send, by return electronic mail, requested items.
- Automatic mechanism for the disseminate of public domain software.
  - Still in use and growing
  - Mirrored at a number of sites
    * netlib2.cs.utk.edu
    * netlib1.epm.ornl.gov
    * research.att.com
    * netlib.no
    * unix.hensa.ac.uk
    * ftp.zip-berlin.de

Requests Made to the Netlib Repositories at the Univ. of Tennessee & ORNL

8,345,418 total requests to these repositories as of Dec 20, 1995



Data as of - Dec 20, 1995 at 02:21:52

Breakdown of requests to each Netlib library

(an alphabetical listing is also available.)
Data as of 12/01/95 at 02:42:11

Library Name	Number of accesses
lapack	469,634
pvm3	375,582
linpack	254,992
slatec	245,724
blas	176,606
clapack	128,081
linalg	126,834
eispack	125,404
slatec/src	117,374
toms	115,825
f2c	96,255
c++	96,182
benchmark	85,217
master	69,587
f2c/src	65,913
minpack	60,158
fn	59,266
fftpack	58,083
na-digest	50,750
port	48,820
slatec/lin	46,232
hence	45,103
confdb	42,631
c++/answerbook	37,298
slatec/chk	37,203

The following types of software are being made available:

- Systems software and software tools.
  - compilers
  - message-passing communication subsystems
  - parallel monitors and debuggers.
- Basic building blocks for accomplishing common computational and communication tasks.
  - Building blocks are meant to be used by Grand Challenge teams
- Research codes that have been developed to solve difficult computational problems.
  - Many have been developed to solve specific problems
  - Serve as proofs of concept
  - Models for developing general-purpose reusable software

Review Procedure

- Treat the review and submission as independent processes.
- Submissions are "subject" to (ongoing) review.
- Review status abstract for each submission.
  - Based on author comments
  - Package documentation
  - Our independent reviewer testing
  - Comments from users.

• NHSE Software Catalog

- Benchmark and example programs (4)
- Data analysis and visualization (22)
- Numerical libraries and routines (57)
  - Computational geometry (7)
  - Linear algebra (18)
  - Optimization (4)
  - Partial differential equations (3)
  - Other (25)
- Parallel processing tools
  - Communication libraries (25)
  - Execution and performance analyzers (31)
  - Parallel I/O systems (5)
  - Parallel programming environments (12)
  - Parallel programming languages and compilers (26)
  - Parallel runtime systems (10)
  - Source code analyzers and restructurers (7)
  - Miscellaneous (16)
- Scientific and engineering applications (66)

NHSE Software Submission

Goals

- Ensure *fixity of publication*
  (file fingerprints, unique name)
- Prevent impersonation and unauthorized changes
  (digital signatures)
- Exercise quality control
  (review classification)
- Promote interoperability
  (RIG Basic Interoperability Data Model)

Technical and Political Issues

- How will the naive user find the right software?
  - Answer: Via the NHSE search/browser interface and the Road Map
- How will authentication, integrity, and version control be implemented?
  - Answer: By a publishing system that includes unique naming & digital signatures.
- Will the NHSE support distribution of software that is not free or cannot be freely distributed?
  - Answer: Yes, but...
    - only provide classification, review, and access
    - use of encryption and separate key distribution
    - NHSE will not have an accounting department
- Will the NHSE be responsible for support of software?
  - Answer: NO!
    - Any support will be by author (or appropriate agent)

User Profiles

- **Currently processed manually**
  - **Click on Submit: User Profile on NHSE home page**
  - **Fill in email address, software needs, information needs**
  - **NHSE Librarian sends you search results and keeps you posted on future items of interest enditemiza**
  - **Future - automate processing**
    - **Handle larger volume more quickly**
    - **Intelligent information agents**
    - **Semantic filtering**

Affiliations, Collaborations, and Intimacies

- **W3 Consortium affiliate member (W^3C)**
- **Reuse Library Interoperablity Group (RIG) Member**
- **Active within IETF**
- **Interaction with Corporation for National Research Initiatives (CNRI)**

Road Map

- **Structured Knowledge Base on Software Components**
  - **Similar to a hypertext encyclopedia**
  - **Assembled with the help of panels of experts**
- **various-perspective catalog**
  - **Glossaries**
  - **Algorithms**
  - **Applications categories**
  - **Enabling software**
  - **Benchmarks**
  - **Hardware**
  - **Education and training**
  - **Links to other sites**
- **Prototype Currently Available**
  - **Topic: available applications codes and their component HPCC technologies**

Different disciplines will maintain their own software repositories

- **Users should not need to access each of these repositories separately**
- **NHSE will provide a uniform interface to a virtual HPCC software repository which will be built on top of the distributed set of discipline-oriented repositories.**
- **The interface will assist the user in locating relevant resources and in retrieving these resources.**
- **A combined browse/search interface will allow the user to explore the various HPCC areas and become familiar with the available resources.**
- **A longer term goal of the NHSE is to provide users with domain-specific expert help in locating and understanding relevant resources.**

Research Issues

- **Repository in a Box**
  - **Everything required for a site to setup a repository and get connected to the NHSE**
    - **publishing tools**
    - **ftp, http, gopher, email servers**
    - **logging and statistical tools**
    - **integration with replication service**
    - **guidance and support**
- **LIFN**
  - **assigned to a particular sequence of bytes,**
  - **once the assignment has been made, the same LIFN cannot subsequently be used to name any other sequence of bytes.**
- **Indexing and Searching**
  - **Currently limited to syntatic keyword matching**
  - **Harvest system provides a set of customizable tools**

Needs for Browsing and Searching

- **Current Web interfaces are difficult and frustrating for the user who is attempting to locate specific information.**
- **Browsing by following hypertext links is slow and can be disorienting.**
- **Keyword searching suffers from the vocabulary mismatch problem and is unsuitable for users with imprecise information and software needs.**

Capitalization of Emerging Technologies

- **Three Levels of Software Development (Pasadena I Workshop)**
  - **Research prototype**
  - **Advanced development prototype**
  - **Commercial product**
- **Focus:**
  - **Move from research to advanced development prototype**
- **Strategy**
  - **Convene outside review panel**
  - **Select as many projects as budget will permit**
  - **Impose requirements for inclusion at Level 2, Reviewed Level**

Combined Searching and Browsing Interface

- **Augment a users mental perception and knowledge of specific domains,**
- **Improve a users understanding of the problem to be solved, and**
- **Form successively more focused and better articulated queries.**

The interface will be in the form of a thesaurus-based roadmap.

HPCC Thesaurus

- **The NHSE will define the top levels of an HPCC thesaurus, drawing on an existing HPCC glossary and on the current NHSE contents to generate thesaurus terms.**
- **Subject area specialists will be called upon to refine the lower levels.**
- **The thesaurus, along with a high-level classification scheme, will form the basis of a hypertext roadmap.**
- **The roadmap will include scope notes and annotations to familiarize users with various HPCC areas and will serve as a springboard for thesaurus-assisted searches.**

HPCC Cataloging

- **To enable searching, cataloging information must be made available for NHSE assets.**
- **Each physical repository will be responsible for maintaining a network-accessible file containing such cataloging information.**
- **These files will be retrieved and indexed by an NHSE indexer on a regular basis, and the resulting searchable index will be replicated for reliability.**
- **The NHSE will use the Harvest system from University of Colorado to do the collection, indexing, and index replication.**

Naming and Authentication

- **The main advantage of distributing the repository is to allow the software to be maintained by those in the best position to keep it up-to-date.**
- **Copies of popular software packages may be transparently mirrored to increase availability, improve response time, and prevent bottlenecks.**

Research Issues

- **Assignment of unique identifiers to retrievable assets**
- **Collection and merging of cataloging information, so that, verification of the authenticity, integrity of retrieved assets, and tracking can take place**

NHSE's approach to these issues

- **Implement a location-independent naming architecture that unambiguously associates a unique name with the byte contents of a published asset and that includes mechanisms for authentication and integrity checking.**
- **Publishing tools will be made available to assist publishers with naming and cryptographically signing published assets, and with exporting asset descriptions to an NHSE search service.**
- **Authentication of assets will be handled by an asymmetric public-private key encryption system.**

Quality Control

- **Users of the NHSE need to have confidence that the software they obtain is high-quality and well-tested.**
- **Quality control will be impossible to automate completely because it requires human judgement and evaluation.**
- **Tools and procedures need to be developed to facilitate the testing and classification or labeling of software with respect to its quality and reliability.**
- **Research is needed to determine what quality control information is most useful and can reasonably be obtained, how to acquire this information, and in what format it should be stored for easy access.**
- **Academic journal model**

Interoperation

- The NHSE will catalog and provide access to software and software-related artifacts from all the HPCC software repositories.
- Assets accessible from other existing software repositories, such as ASSET, CARDS, DSRS, and ESLA, to name just a few, may also be of interest to NHSE users.
- The NHSE will be participating in a small-scale interoperability experiment with the above repositories to help define requirements for further interoperation efforts.
- The NHSE will also be working with the Reuse Library Interoperability Group (RIG) on establishing standards for unique naming, asset description and classification, and asset evaluation.
- In the future, the NHSE will interoperate with these other repositories so that software from them may be retrieved directly from the NHSE interface.

Measurement

- Keep statistics on downloads from NHSE pages
  - Including userid of request source
    - not always possible
    - invasion of privacy
- Statistical Survey of Unreviewed and Partially Reviewed Software
  - Identification of candidates for full review
  - Email inquiries to determine usage pattern
- Systematic Survey of Reviewed Software
  - Application
  - Usage pattern
  - Industry versus academia

Software Survey

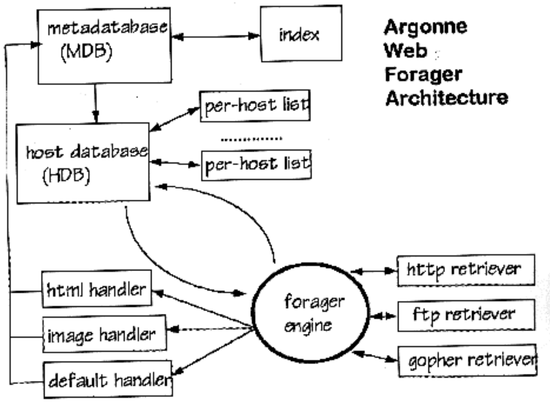
- 200+ software items from NHSE collection which have been abstracted and keyword indexed
- testbed for search strategies
- software review candidates

Reliable File Mirrowing

- Current mirroring systems inadequate
- Replication service to provide:
  - Fast selective file transfer
  - Independent of machine architecture and operating system dependenices
- Integrated with location lookup mechanisms

NHSE at ANL

- Modular web robot
- Parallel web indexing engine
- DNS/geographical mapping
- Autonomous agents for collecting information on the web



Naming of Distributed Resources

- Need single name to refer to an object
- May have multiple locations
- May change over time
- 2 kinds
  - `urn:netlib:Welcome` (Dynamic)
  - `lifn:netlib:Welcome_940328` (Fixed)
- Fixed reference to a published object

Overall Strategy for the NHSE:

- Effectiveness of the NHSE will depend on discipline-oriented groups and Grand Challenge teams having ownership of the discipline-oriented software repositories.
- The information and software residing in these repositories will be best maintained and kept up-to-date by the individual disciplines, rather than by centralized administration.
- Central administration will be used instead to handle interoperation and meet common needs.
- Although the various disciplines will have ownership of the repositories, they should not be expected to develop the software and tools for building, managing, and interfacing to their repositories.
- Much useful information retrieval (IR) software is currently available, both in the form of client and server programs (e.g., http servers and WWW browsers), as and this software should be incorporated into the NHSE.

Summary

- **Initial implementation built on existing technologies**
  - **WWW**
    - **Distributed**
    - **Scalable**
  - **Netlib, etc**
  - **Rapid deployment**
- **Multilevel review and classification scheme**
  - **Unreviewed**
  - **Partially reviewed**
  - **Reviewed**
- **Road Map**

Summary (continued)

- **Measurement and evaluation**
  - **Statistics for unreviewed and partially reviewed**
  - **Evaluations for reviewed**
- **Outreach and technology transition**
  - **Educational activities aimed at the user community**
  - **Fostering technology development by industry**
- **Capitalization of emerging technologies**
  - **Foster transition from research to advanced development prototype**
- **Working on standardization within WWW community**
  - **member of RIG, IETF, IESG, WWW consort**

Some Url's Related to Scientific Computing

- **National HPCC Software Exchange**
  http://www.netlib.org/nhse/
- **NHSE Software Catalog**
  http://www.netlib.org/nhse/sw_catalog/index.html
- **Netlib Repository**
  http://www.netlib.org/
- **Computational Science Education**
  http://www.netlib.org/nhse/cse_edu.html
- **Books, Course Materials, and Tutorials**
  http://www.netlib.org/nhse/cse_edu.html#books
- **CS267 - Applications of Parallel Computers**
  U.C. Berkeley CS267: Spring 1995 Jim Demmel
  http://www.icsi.berkeley.edu/cs267/
- **18.337 Parallel Scientific Computing**
  MIT Spring, 1995 Alan Edelman
  http://web.mit.edu/18.337/WWW/home.html
- **North Carolina State University**
  Visualizations in Materials Science
  http://vims.ncsu.edu/cgi/index.acgi
- **Computational Science Education Project**
  http://csep1.phy.ornl.gov/csep.html