@xxx.lanl.gov

first steps toward electronic research communication

Paul H. Ginsparg

hep-th@xxx.lanl.gov is the e-mail address for the first of a series of automated archives for electronic communication of research information. This "e-print archive" went on-line in August, 1991. It began as an experimental means of circumventing recognized inadequacies of research journals, but unexpectedly became within a very short period the primary means of communicating ongoing research information in formal areas of high energy particle theory. Its rapid acceptance within this community depended critically on both recent technological advances and particular behavioral aspects of the community. There are now more than 3600 regular users of hep-th worldwide.

The archiving software has been expanded to serve a number of other research disciplines (see Figure 1). The extended, automatically maintained database and distribution system currently serves over 20,000 users from more than 60 countries and processes over 30,000 messages per day. It is already one of the largest and most active

databases on the internet. This system may be a paradigm for worldwide, discipline-wide scientific-information exchange when the next generation of "electronic-data highways" begins to provide more universal access to high-speed computer networks.

Background

The rapid acceptance of electronic communication of research information in my own community of high-energy theoretical physics was facilitated by a pre-existing "preprint culture," in which the irrelevance of refereed journals to ongoing research has long been recognized. Since the mid-1970s the primary means of communicating new research ideas and results has been a preprint-distribution system in which printed copies of papers were sent through the ordinary mail to large distribution lists at the same time that they were submitted to journals for publication. (The larger high-energy physics groups typically spent between $15,000 and $20,000 per year on copying, postage, and labor costs for their preprint distribution.) Typically, it takes six months to a year for a paper to appear in a journal. Members of our community have therefore learned to determine from the title and abstract (and occasionally the authors) whether we wish to read a paper as well as to verify results ourselves rather than rely on the alleged verification of overworked or otherwise careless referees. The small amount of filtering provided by refereed journals plays no effective role in our research.

Taking advantage of advances in computer software and hardware, many of us had already begun using highly informal mechanisms of electronic-information exchange by the mid-1980s. The first such advance was a program called TeX, which was written by computer scientist Donald E. Knuth of Stanford. TeX was soon adopted as our standard scientific wordprocessor, and for the first time we could produce for ourselves a printed version equal or superior in quality to the published version. TeX has the additional virtue of being based on ASCII, so transmitting TeX files between different computer systems is straightforward. Collaboration at a distance became extraordinarily efficient, since we no longer had to express-mail versions of a paper back and forth and could instead see one another's revisions essentially in real time. Figures and technical illustrations can also be generated within a TeX-oriented picture environment or, more generally, can be transmitted as standardized Postscript files produced by a variety of graphics programs.

A second technological advance achieved during the same period was the exponential increase in computer network connectivity. By the end of the 1980s, virtually all researchers in this community were plugged into one or another of the interconnected worldwide networks and were using e-mail on a daily basis. Finally, the development of large on-line archives of research papers has been enabled by the widespread availability of low-cost, high-powered workstations with high-capacity storage media. After compression, storing an average paper with figures requires 40 kilobytes of memory. Thus one of the current generation of

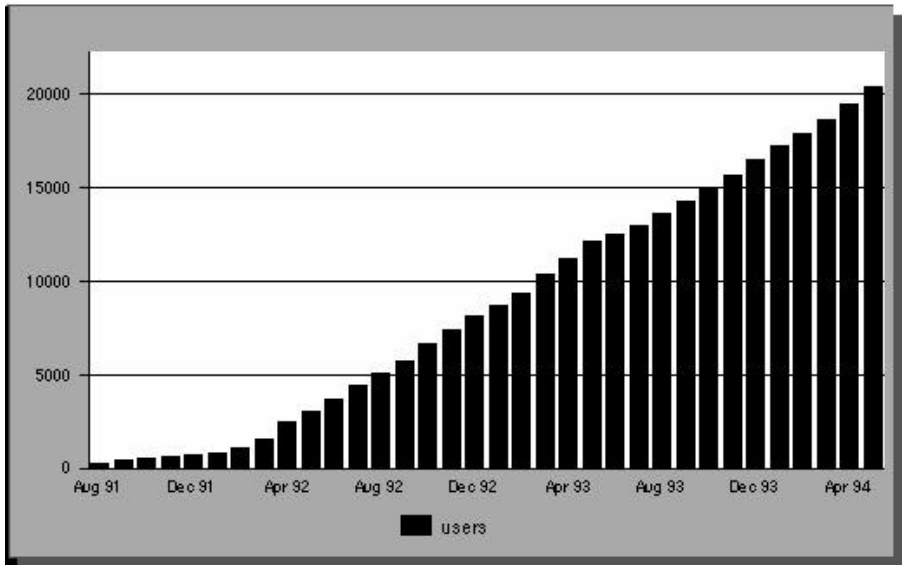

Figure 1. Number of Users of E-Print Archives

The bar chart shows the combined number of e-print-archive users over the period beginning August 1991 and ending April 1994. The data include users of the following e-print archives: High-energy particle theory (formal), started August 1991; Algebraic geometry, started February 1992; High-energy particle theory (phenomenological), started March 1992; Astrophysics, started April 1992; Condensed-matter theory, started April 1992; Computational and lattice physics, started April 1992; Functional analysis, started April 1992; General relativity/Quantum cosmology, started July 1992; Nuclear theory, started October 1992; Nonlinear Sciences, started March 1993; Economics, started July 1993; High-energy experimental physics, started April 1994; Chemical physics, started April 1994; Computation and language, started April 1994.

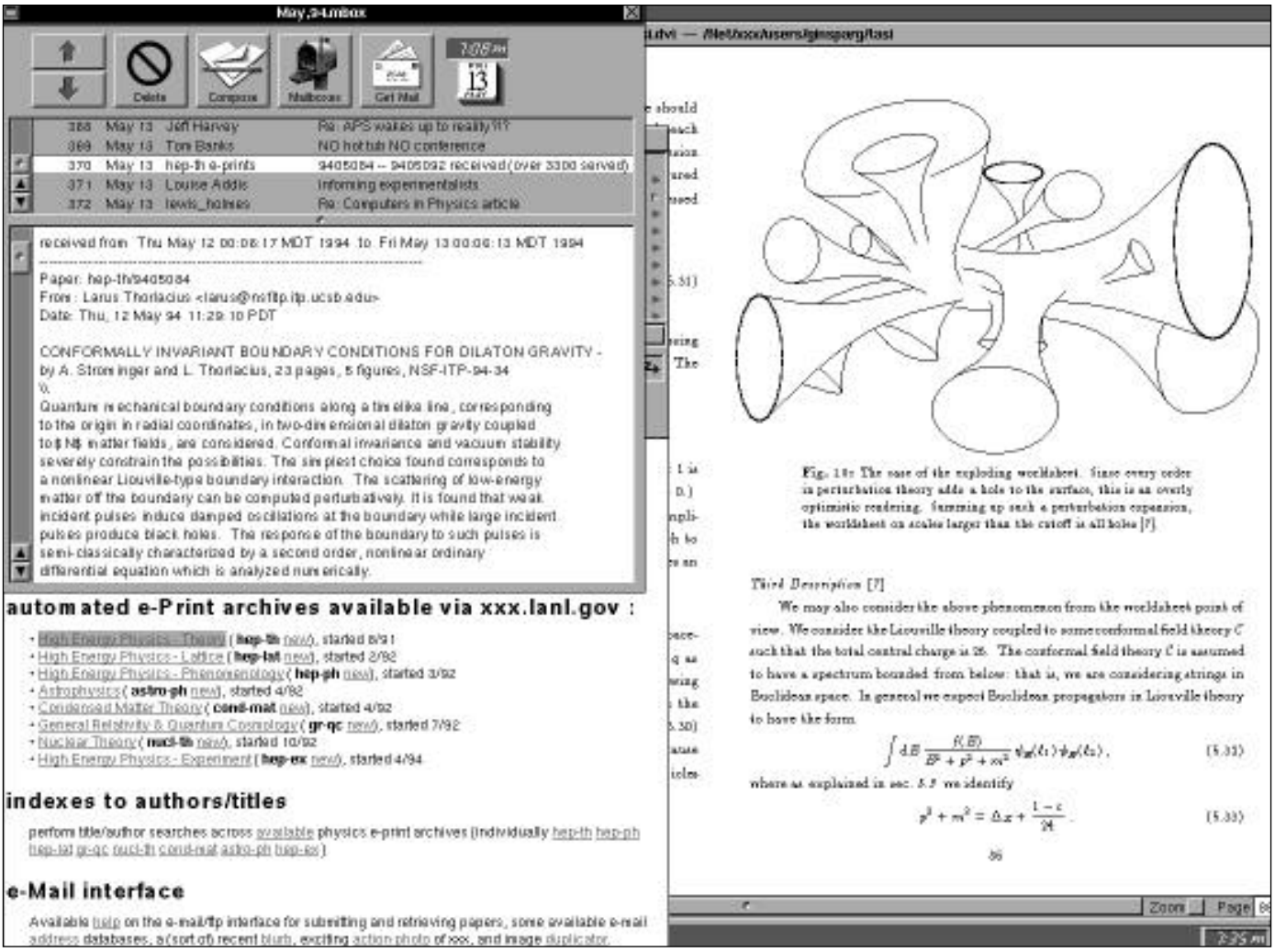

Figure 2. User Interface for E-Print Archives

The left side of the sample screen above shows two user interfaces for accessing the e-print archives. The window in the upper left corner shows abstracts received through e-mail, and the window in the lower left corner shows the graphical user interface provided by a WorldWideWeb client (in this case OmniWeb.app running under NeXTstep) accessing the frontpage http://xxx.lanl.gov/. (The underlined text signifies network hyperlinks that bring up new hypertext when clicked upon.) A paper extracted from the e-print archive appears in the window on the right side, where it can be read or sent to a printer.

rapid-access gigabyte disk drives costing under $1,000 can hold 25,000 papers at an average cost of 4 cents per paper. Slower-access media for archival storage cost even less: A digital audio-tape cartridge, available from discount electronics dealers for under $15, can hold over 4 gigabytes, that is, over 100,000 such papers. The data equivalent of multiple years of most journals is often far less than the amount many experimentalists handle every day. Moreover, the costs of data storage will only continue to decrease.

Since storage is so inexpensive, an archive can be duplicated at several distribution points, minimizing the risk of loss due to accident or catastrophe and facilitating worldwide network access. The Internet runs 24 hours a day—with virtually no interruptions—and transfers data at rates of up to 45 megabits per second (that is, less than a hundredth of a second per paper). Projected upgrades of NSFnet to a few gigabits per second within a few years should be adequate to accommodate increased usage for the academic community. The commercial networks that will constitute the nation's electronic data highway will have even greater capacity.

These technological advances—combined with a remarkable lack of response to the electronic revolution from conventional journals—rendered the development of e-print archives "an accident waiting to happen." Perhaps more surprising has been the readiness of scientific communities to adopt this new tool of information exchange and to explore its implications for traditional review and publication processes. The exponential growth in archive usage suggests that scientific researchers are not only eager—but indeed impatient—for completion of the proposed "information superhighways" (though not necessarily the tollbooths of "information turnpikes").

Implementation

Having concluded that an electronic preprint archive was possible in principle, I spent a few afternoons during the summer of 1991 writing the original software. It was designed as a fully automated system in which users construct, maintain, and revise a comprehensive database and distribution network without outside supervision or intervention. The software is rudimentary and allows users with minimal computer literacy to communicate e-mail requests to the Internet address hep-th@xxx.lanl.gov. Remote users can submit and replace papers, obtain papers and listings, get help on available commands, search the listings for author names, and so on.

The formal communication provided by an "e-print archive" should be distinguished from the informal (and unarchived) communication provided by electronic bulletin boards and network news. In the case of an e-print archive, researchers are restricted to communication by means of abstracts and research papers suitable for publication in conventional research journals. Electronic bulletin boards are more akin to ordinary conversation or written correspondence; that is, they are neither indexed for retrieval nor stored indefinitely. The e-print archives allow a submitter to replace his or her submission, and the program automatically checks on database integrity to ensure, for example, that the person replacing a submission is indeed the original submitter. In addition, the system maintains permanent records of submissions and the dates they were submitted, and it records the number of user requests for each paper. Subscribers to the system receive a daily listing of new titles and abstracts (see Figure 2).

The initial user base for hep-th was 160 addresses assembled from pre-existing e-mail distribution lists in the subject of two-dimensional gravity and conformal field theory. Within six months the user base grew to encompass most of the workers in formal quantum field theory and string theory, and now includes the 3600 subscribers mentioned above. Its smooth operation has transformed it into an essential research tool—many users have reported their dependence on receiving multiple "fixes" each day. The original hep-th archive now receives roughly 200 new submissions per month, responds to more than 700 e-mail requests per day, and transmits more than 1000 copies of papers on peak days. Internet e-mail access time is typically a few seconds. The system originally ran as a background job on a small UNIX workstation (a 25-megahertz NeXTstation with a 68040 processor purchased for roughly $5,000 in 1991), which was primarily used for other purposes by another member of my research group, and placed no noticeable drain on CPU resources. The system has since been moved to an HP 9000/735 that sits exiled on the floor under a table in a corner.

For those directly on the Internet, the system allows anonymous FTP access to the papers and listings directories. Now access can be gained through WorldWideWeb for those with the required (public-domain) client software (see Figure 3). Local menu-driven interfaces can be set up to automatically pipe selected papers through text formatters directly to a screen previewer or printer. (Such software has been set up to cache and redistribute papers on many local networks.) The WorldWideWeb interface for the multiple archives at xxx.lanl.gov currently processes over 5000 requests daily (see Figure 4). While that is only a small fraction of the overall usage of the e-print archives, WorldWideWeb is ex-

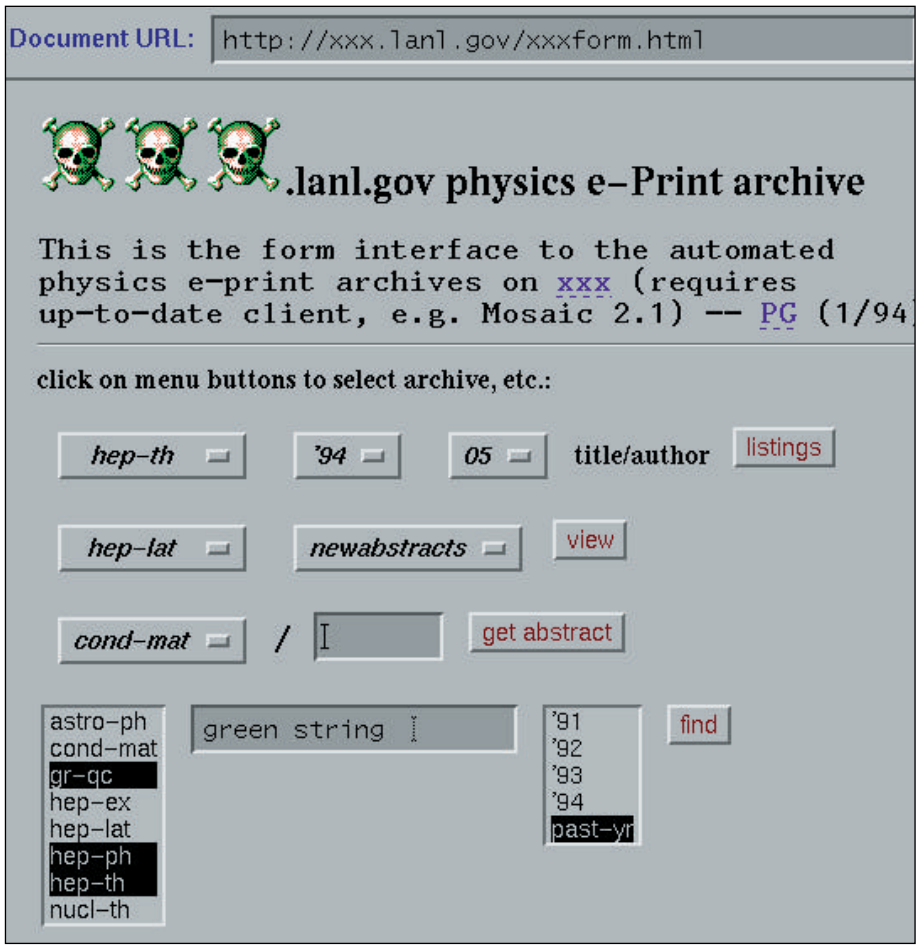

Figure 3. The "Form Interface" on xxx.lanl.gov

This screen grab shows another WorldWideWeb client, Mosaic, accessing the "form interface" on xxx.lanl.gov. The Motif "buttons" allow the client to choose an archive to view monthly listings or daily abstracts received, or to search the title/author listings of selected archives for given time periods. Listings are displayed in hypertext with included hyperlinks that retrieve paper abstracts or full text in either TeX or Postscript format.

pected to become the dominant mode of access.

An active archive such as hep-th requires about 70 megabytes per year (that is, $70 per year) for storing papers, including figures. Its network usage is less than 10^{-4} of the lanl.gov backbone capacity, so it places a negligible drain on local network resources. It requires little intervention and has run entirely unattended for extended periods while I have been away on travel. It is difficult to estimate the potential of future dedicated systems because the resources of the current experimental system (run free of charge) are so far from saturation.

Storage and retrieval of figures. Although software for technical illustrations has not yet been standardized, the vast majority of networked physics institutions have screen previewers and laser-printers that display and print Postscript files created by a wide variety of graphics programs. Figure files are typically submitted as compressed Postscript, and papers can be printed with the figures embedded directly in the text. High-resolution digital scanners will soon become as commonplace as fax machines, thus permitting the inclusion of figures of almost any origin. (Of course it is already possible to fax figures in any format to a machine equipped with a fax modem, convert them to bitmapped Postscript files, and then append them to the paper.) Using appropriate data compression and Postscript conversion, figures typically increase paper-storage requirements by an inconsequential factor of 2.

Some measure of the success of e-print archives is given, first, by numerous testaments from users that they find it an indispensable research tool—effectively eliminating their reliance on conventional print journals; second, by decisions of numerous institutions to discontinue their preprint mailings in

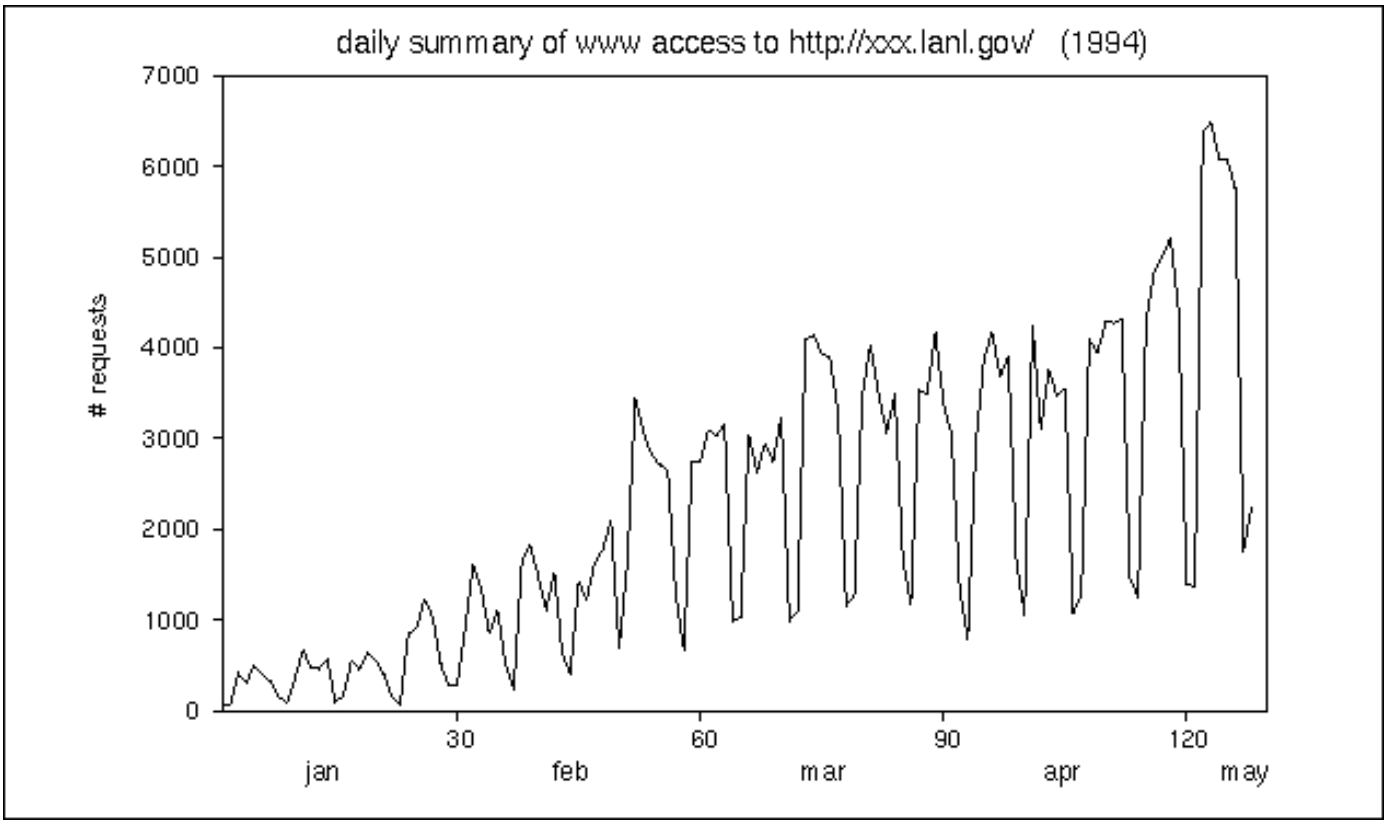

Figure 4. Requests to xxx.lanl.gov through WorldWideWeb

The graph shows the number of http (hypertext transfer protocol) requests to the WorldWideWeb interface on xxx.lanl.gov per day beginning January 1, 1994. The seven-day periodicity is evidence that many physicists still do not have readily available network access on weekends.

recognition of the superior service provided by e-print archives; and third, by the fact that in some of the fields served by e-print archives, it has become customary to provide a paper's electronic-archive index number as a reference rather than a local report number or a published reference.

Prospects and Concerns

The system, in its present form, was not intended to replace journals but only to organize what was once a haphazard and unequal distribution of electronic preprints. It is increasingly used as an electronic journal, however, because retrieving a copy of a paper electronically is more convenient than physically retrieving a paper from a file cabinet. Aside from minimizing geographic inequalities by eliminating the "boat-mail gap" between continents, the system institutes a form of democracy in research wherein access to new results is granted equally to everyone from beginning graduate students to seasoned operators. No longer is it crucial to have the correct connections or to be on exclusive mailing lists to be kept informed of progress in one's field. The pernicious problem of lost or stolen preprints experienced by some large institutions is also definitively exorcised. The many institutions that have eliminated their hard-copy distribution of preprints have already seen significant savings in time and money; others have specifically requested that hard copy no longer be sent to them, since electronic distribution has proven reliable and more efficient. Implementing a billing system for the use of an e-print archive would be fairly straightforward; however, such archives cost so little to set up and maintain that they can be offered virtually free. Overburdened terminal resources at libraries are not an issue, since access is typically via the terminal or workstation on one's desk or in the nearest computer room.

Electronic research archives will prove particularly useful for new and

emerging interdisciplinary areas of research for which there are no existing print journals and for which information is consequently difficult to obtain. In many such cases, it is advantageous to avoid a proliferation of premature or ill-considered new journals. Cross-linking of various databases provides an immediate virtual meeting ground for researchers who wouldn't ordinarily communicate with one another. Researchers can quickly establish their own dedicated electronic archive when it is appropriate and ultimately disband if things do not pan out—all with far greater ease and flexibility than is provided by traditional publication media.

Electronic access to scientific research will be a major boon to developing countries, since the expense of connecting to an existing network is infinitesimal compared with that of constructing, stocking, and maintaining libraries. (I frequently receive messages from physicists in developing countries confirming how much better off they find themselves even in the short term with the advent of electronic distribution systems—they are no longer "out of the loop." Others report feeling that their own research gets a more equitable reading—their research is no longer dismissed for the superficial reasons of low-quality printing or paper stock.) Now that much of the technology has ripened, Eastern European and third-world nations may rapidly develop their electronic infrastructures to the level that took developed nations over a decade to reach—a level at which data-transmission lines are as common as telephone service and terminals and laser-printers as common as typewriters and copy machines. (Similar comments apply equally to the less well-endowed institutions in the U.S., and the changes experienced by physics and biology departments are soon to be repeated by the full range of conventional academic institutions, including teaching hospitals, law schools, humanities departments, and ultimately public libraries and public grade schools.)

E-print archives will eventually bring great changes to the scientific-journal industry as well. Over the past decade publication companies have been somewhat irresponsible—increasing the number of journals and as well the subscription price per journal (some single journal subscriptions to libraries now run well over $10,000 per year) during a period when libraries are experiencing a decrease in both funds and space. Publishers have been slow to incorporate electronic communication into their operation and distribution, although such a move would ultimately result in dramatic savings in cost and time for all involved.

Some members of the community have voiced their concern that electronic distribution will somehow increase the number of preprints produced, or encourage dissemination of preliminary or incorrect material. This concern, however, confuses the method of production with the method of distribution—most researchers are *already* producing at saturation. Moreover, once posted to an archive, the electronic form is instantly publicized to thousands of people. Thus the embarrassment over incorrect results is, if anything, *increased*. Such submissions cannot be removed; they can only be replaced by a note that the work has been withdrawn as incorrect, leaving a more permanent blemish than a hard copy of limited distribution that is soon forgotten.

The widespread use of e-print archives does not necessarily make refereed forums obsolete. In some disciplines, the refereeing process plays a useful role in improving the quality of published work, filtering out large amounts of irrelevant or incorrect material, and validating research for the purpose of job and grant allocation. A refereeing mechanism could be easily implemented for the e-print archives in the form of either a filter prior to electronic distribution or a review after submission by volunteer readers and/or selected reviewers. In either case, the archives could be partitioned into one or more levels of refereed and unrefereed sectors. Thus, lifting the artificial financial constraints on dissemination of information and decoupling it from the traditional refereeing process will allow for more innovative methods of identifying and validating significant research.

Problems may arise, however, as computer networking spreads outside of the academic community. For example, hep-th would be somewhat less useful if it were to become inundated by submissions from "crackpots" promoting their perpetual-motion machines. It is clear that the architecture of the information highways of the future will somehow have to reimplement the protective physical and social isolation currently enjoyed by ivory towers and research laboratories.

Increased standardization of networking software and electronic storage formats during the 1990s encourages us to fantasize about other possible enhancements to scholarly research communication—in particular, discussion "threads" in which users respond to one another's comments on a specific topic. Usenet newsgroups, for reasons such as their lack of indexing and archiving and their open nature, are unlikely to prove adequate for serious purposes. On the other hand, it is now technically simple to implement a WorldWideWeb form-based submission system to build hyperlinked threads, accessible from given points in individual papers and also started from a subject-based linked discussion page. All posted text could be indexed by the WAIS (Wide Area In-

formation Server) scheme for easy retrieval, and related threads could interleave and cross-link in a natural manner, with standard methods for moving forward and backtracking. A histogram-like interface showing the activity on each thread would facilitate finding threads of current interest, and the index could allow location of all postings by a given person (including self) with the date of latest follow-up to facilitate tracking of responses. This would provide a much more flexible format than Usenet, specifically avoiding awkward protocols for group creation and removal as well as avoiding potentially unscalable aspects of nntp (the network news transfer protocol). For the relatively circumscribed physics research community, a central database (copied onto many nodes, as usual) would have no difficulty with storage or access bandwidth. To enable full-fledged research communication with in-line equations or other linkages, we require slightly higher quality browsers than are currently available. But with hypertext transfer protocols (http) now relatively standardized, network links and links to other application software can be built into underlying TeX documents (and configured into standard macro packages) to be either interpreted by dedicated TeX previewers or passed by a suitable driver into more archival formats (such as Adobe Acrobat PDF) for greater portability across platforms. Multi-component messages could also be assembled in a graphical user interface for composing MIME (multipurpose internet mail extension) messages to be piped to the server by means of the http POST protocol, thereby circumventing some of the inconvenient baggage of Internet sendmail or FTP protocols.

While the above is technically straightforward to implement, there remains the aforementioned issue of limiting access to emulate that effective insulation from unwanted incursions afforded by corridors and seminar rooms at universities and research laboratories. One method would be to employ a "seed" mechanism—that is, to start from a given set of "trusted users" and let them authorize others (and effectively be responsible for those beneath them in the tree), with guidelines such as that the new users must have doctorates or be doctoral candidates, and make permission to post/authorize revocable at any time, retroactive one level back in the tree. To allow global coverage, application to the top level for authorization could be allowed to start a new branch. The scheme entails some obvious compromises, and other schemes are easily envisioned, but the ultimate object remains to determine the optimal level of filtering for input access to maintain an auspicious signal-to-noise ratio for those research communities that prefer to be buffered from the outside world. This would constitute an incipient "virtual communication corridor," further facilitating useful research communication in what formerly constituted both pre- and post-publication phases, and rendering ever more irrelevant individual researchers' physical location.

Finally, we mention that the e-print archives in their current incarnation already serve as surprisingly effective inducements for computer literacy, and they have motivated some dramatic changes in computer usage. Researchers who previously disdained computers now confess an addiction to e-mail. Many researchers who for years had refused to switch to UNIX or to TeX are in the process of converting; others have suddenly discovered the power of browsing with World-WideWeb. The system's effectiveness in motivating these changes justifies the philosophy of providing dual functionality in the form of top-of-the-line search, retrieval, and input capabilities for cutting-edge power users, while maintaining "lowest-common-denominator" capabilities for the less "network-fortunate."

Conclusions and Open Questions

These systems are still primitive, and they represent only tentative first steps in the optimal direction. To summarize, thus far we have learned.

- The exponential increase in usage of electronic networking over the past few years opens new possibilities for both formal and informal communication of research information.

- In some fields of science, electronic preprint archives have been on-line since mid-1991 and have become the primary means of communicating research information to many thousands of researchers within the fields they serve. It has been established that people will voluntarily subscribe to receive information from these systems and will make aggressive use of them if they are set up properly. It is anticipated that such systems will grow and evolve rapidly in the next few years.

- From such experimental systems, we have learned that open (unrefereed) distribution of research information can work well for some disciplines and has advantages for researchers in both developed and developing countries. We have also learned that the technology and network connectivity are currently adequate to support such systems, the performance of which should benefit from the continuing improvements in technology.

I conclude with some unanswered questions to amplify some of my earlier comments:

- Who will ultimately be the prime beneficiaries of electronic research communication (that is, researchers, publishers, libraries, or other providers of network resources)?
- What factors influence research communities in their rate and degree of acceptance of electronic technology, and what mechanisms are effective in facilitating such changes?
- What role will be played by the conventional peer-refereeing process in the electronic media, and how will it differ from field to field?
- What role will be played by publishing companies, and how large will their profits be? If publication companies do adopt fully electronic distribution, will they pass along the reduced costs associated with the increased efficiency of production and distribution to their subscribers? Can publishing companies provide more value than an unmanned automated system whose primary virtue is instant retransmission?
- What role will be played by library systems? (Will information be channeled through libraries or, instead, directly to researchers?)
- How will copyright law be applied to material that exists only in electronic form? At the moment publishing companies are "looking the other way," living with the dissemination of the electronic preprint information as they did with the earlier preprinted form—claiming that it would be antithetical to their philosophy to impede dissemination of information. Will they continue to be so magnanimous when libraries begin to cancel journal subscriptions?
- What storage formats and network utilities are best suited for archiving and retrieving information? Currently we use a combination of e-mail, anonymous FTP, and window-oriented utilities, such as Gopher and WorldWideWeb, combined with WAIS indexing to retrieve TeX and Postscript documents. Will something even better—for example, Acrobat or some other format currently under development—soon merge with the above or emerge as a new standard?
- How will the medium itself evolve? Conservatively, we can imagine "interactive" journals in which equations can be manipulated, solved, or graphed; citations can instantly open references to the relevant page; comments and errata dated and keyed to the relevant text can be inserted as electronic "post-it notes" in the margins, and so on. Ultimately we will have a multiply interconnected network hypertext system with transparent pointers among distributed databases that transcends the limits of conventional journals in structure, content, and functionality, thereby transforming the very nature of *what* is communicated. These are the kinds of benefits for which we should certainly be willing to pay. Certainly we do not wish to clone current journal formats (determined as they are by the constraints of the print medium) in the electronic medium—we are already capable of distinguishing information content from superficial appearance. Who will decide the standards required to implement any such progress?

This began for me as a spare-time project to test the design and implementation of an electronic preprint distribution system for my own relatively small research community. Its feasibility had been the subject of contentious dispute, and its realization was thought—even by its proponents—to be several years in the future. Its success has led to an unexpectedly enormous growth in usage. It has expanded into other fields of research and has elicited interest from many others—I have received over one hundred inquiries into setting up archives for different disciplines. Each discipline will have slightly different hardware and software requirements, but the current system can be used as a provisionary platform that can be tailored to the specific needs of different communities. Despite the success of this project, for three years it remained a spare-time project with little financial or logistical support. Only very recently have the Laboratory, certain government funding agencies, and certain professional societies moved to increase their levels of involvement.

Further development will require coordination among interested researchers from various disciplines, computer and networking staff, and interested library personnel. In particular, it will require dedicated staffing. At the moment, hardware and software maintenance of existing automated archives remains a loosely coordinated volunteer operation, and little further progress can be made on the issues raised by the current systems without some thoughtful direction. Perhaps the centralized databases and further software development will ultimately be administered and systematized by established publishing institutions—if they are prescient enough to reconfigure themselves for the inevitable. Since it has been researchers who have taken the lead thus far, however, we should retain this unique op-

portunity to continue to lead the development of such systems in optimal directions and on terms maximally favorable to ourselves. ■

Acknowledgements

Many people have contributed (consciously or otherwise) to the development of these systems. The original distribution list from which hep-th sprung in 1991 was assembled by Joanne Cohn, whose incipient efforts demonstrated that members of this community were eager for electronic distribution (and Stephen Shenker recommended that the original archive name not include the string "string"). Continual improvements have been based on feedback from users too numerous to credit (although among the most vocal have been Tanmoy Bhattacharya (T-8), Jacques Distler, Marek Karliner, and Paul Mende). People who have administered some of the remote-based archives include Dave Morrison, Bob Edwards, Roberto Innocente, Erica Jen (T-7), and Bob Parks. Joseph A. Carlson (T-5) and David Thomas set up the original Gopher interfaces in late 1992. The Network Operations Center at Los Alamos National Laboratory has reliably and uncomplainingly supplied the requisite network bandwidth @lanl.gov, and Joseph H. Kleczka (C-8) has been available for crisis control. Louise Addis and the staff at the SLAC library moved quickly to incorporate e-print information into the SPIRES database, furthering their decades of tireless electronic service to the high-energy physics community. Dave Forslund (Advanced Computing Laboratory) and Richard Luce (CIC-14) helped lobby for support from within the Laboratory, and the Advanced Computing Laboratory has in addition provided some logistical and moral support. Finally, Geoffrey B. West (T-8) repeatedly and against all obvious reason insisted that the Los Alamos National Laboratory is an appropriate sponsor for this activity, while simultaneously bearing the bad news both from within the Laboratory and from certain government funding agencies

Paul H. Ginsparg received his A.B. in physics from Harvard University in 1977 and his Ph.D. in physics from Cornell University in 1981 under the direction of Kenneth G. Wilson. He then joined the physics department at Harvard University as a Junior Fellow, eventually becoming an Associate Professor. In 1990 he came to the Elementary Particles and Field Theory Group in the Laboratory's Theoretical Division, where he carries out research in relativistic quantum field theory.