YouServ: A Web Hosting and Content Sharing Tool for the Masses*

Roberto J. Bayardo Jr., Rakesh Agrawal, Daniel Gruhl, Amit Somani

Presented By:
BINAMRA DUTTA

OBJECTIVE

- Design, implementation, and a successful intranet deployment of the YouServ system, and compare it with several alternatives.
- A system that allows the users to pool existing desktop computing resources for *high availability* web hosting and file sharing
- Exploits standard web & internet protocols
- No need to install special purpose software
- Is cost effective

IMPORTANCE

- Alternative P2P systems like napster, kazaa need special purpose software to be installed in the users machine
- Youserv exploits peer-to-peer techniques to provide easy to use, low-cost *web* publishing of content with users using their own hardware and internet connections to put content up
- Users are immediately assigned a convenient domain name which always directs to your site content, even if your ISP assigns IP addresses dynamically
- By pooling resources of a group of friends, your content can remain accessible even after your computer is turned off or disconnected from the network.

- You can publish content on the web from your own machine even if you are behind a firewall (or cannot accept inbound connections for any other reason).
- You can secure content on your site without assigning, distributing, and maintaining any accounts or passwords.
- No restrictive space or disk quotas, no advertisement bombarding, simple to use and cost effective
- Machines of the end users do all the work, only server side components are dynamic DNS server & the Youserv Coordinator

RELATED WORK

- Badblue implementing the Gnutella protocol, MacOS personal web server, Microsoft personal web server, lotus notes, instant messenger etc. can be use for file sharing and putting up web content
- XDegrees Corp. provides similar software which provides a uniform namespace for the organization's distributed information assets
- Farsite project provide a serverless fault tolerant distributed file system
- Mojo Nation replicates content to support swarm distribution, allowing multiple bandwidth limited peers to cooperatively serve files faster and with improved availability
- Freenet is a system that supports anonymous, uncensorable content publishing. Freenet replicates content as it is pulled across the network in response to a query, ensuring the most highly accessed content remains the most highly replicated
- Project JXTA is an open source project led by Sun Microsystems that is creating a common platform for building a range of distributed services composed of addressable and communicating peers running on arbitrary devices
- WebDAV is an extension of the HTTP protocol to better support web authoring

YOURSERV INTERFACE

- Being run in the IBM corporate environment for past 9 months
- Download & install the YouServ Software, login through the IBM corporate ID and password based on the existing e-mail account
- Employees site is assigned a domain name based on this e-mail (bayardo@us.ibm.com → bayardo.userv.ibm.com)
- It creates a brand new empty folder populating it with a default homepage (containing employee information) and two access-controlled subdirectories
- User can change the contents of the shared folder at any time
- Content outside the access controlled subdirectories can be accessed by anyone capable of determining the URL

## YouServ lets a user quickly publish a file on the web

---
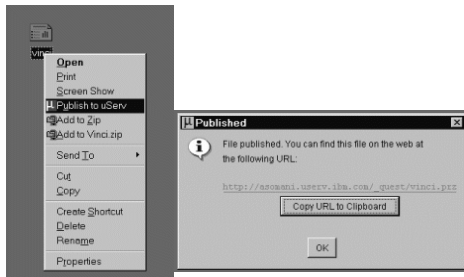
- For sharing a file or an entire folder of files a user can either copy it to his or her shared folder, or simply right-click over it and select a "Publish to YouServ" desktop menu option that is installed in Windows environments along with the peer software
- Also integrates with Lotus notes creating a toolbar icon which copies the selected files to the users shared folder and pastes the URL pointing to the file into the message
- Keeps attachments from getting replicated in every mailbox and allows the sender to monitor when and from where the content was viewed

---
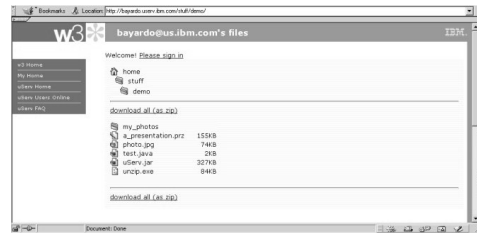
- A unique feature of YouServ is that it allows site visitors to download the entire contents of a shared folder hierarchy with one click in ZIP format

YouServ GUI Console
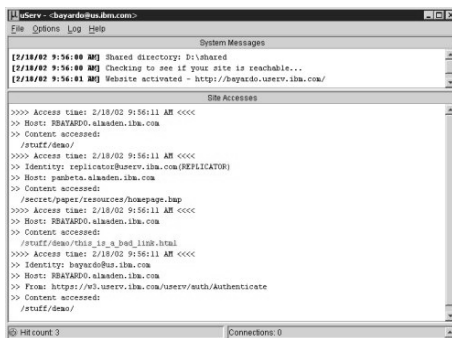
- Displays a log that lists any files that have been accessed, when they were accessed, from which host, and the referring page (if available)
- This log also flags error requests (such as file not found), which facilitates site debugging.
- The identity of anyone who accesses a file is also displayed in the log if it can be determined.
- YouServ requires that anyone who attempts to access secured content reveal their identity by signing in with the site. Signing in initially involves providing a corporate e-mail id and password to a trusted authentication server. Subsequent sign-ins, even with other YouServ sites, requires only a single click for the duration of the browser session.
- Users who are not interested in monitoring the activity of their site can close the GUI window. A YouServ control-tray icon allows the GUI to be restored as desired.

REPLICATION

- YouServ supports site replication and shared hosting
- Any user can list other users ("replicators") who are willing to host their content when they are offline.
- These users should also specify the "masters" whose content they would be hosting completing the two way agreement.
- Once the masters and replicators are specified, replicas synchronize with the master site automatically, and replicas are activated automatically by the YouServ coordinator when the user disconnects, even when the user does not cleanly shut down

---

PEER TO PEER PROXYING

- Members of the YouServ community who are able to accept inbound port 80 connections can be called upon to accept them on behalf of users who are not
- On starting up the YouServ software on the users machine detects whether a proxy is needed
- If needed the Coordinator forwards the contact information of another user willing to serve as a proxy
- The users machine then connects to the proxy machine which then accepts connections on its behalf
- It non-intrusively indicates which machine is serving as a proxy and checks if the user machine can be reconfigured so that proxying can be avoided
- As proxying only consumes bandwidth when someone is downloading files from the proxied users site hence CPU time consumed is typically negligible.

YOUSERV COMPONENTS

➢ Browsers - any machine running a standard web browser and accessing YouServ content.
➢ Peer nodes - the machines of users who have set up a YouServ site by running the YouServ peer software.
➢ Dynamic DNS - a centralized component that speaks the DNS protocol for resolving YouServ domain names to machine IP addresses.
➢ YouServ Coordinator- a centralized component that provides user authentication, proxy and replica matchmaking, IP sniffing and firewall detection, site availability monitoring, and other administrative tasks.
➢ For supporting standard web browsers DNS & HTTP are used
➢ YouServ specific protocols are implemented using Vinci libraries
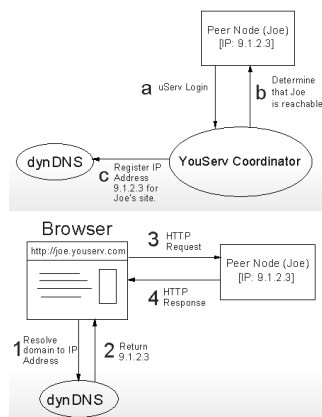
ACCESSING CONTENT

➢ Three modes

Basic: A peer node is online and capable of accepting inbound connections, and serves its own site.

Peer-hosted: A peer node is offline, and a replica of its site is served by another peer node.

Proxied: A peer node is online but unable to accept inbound connections, and serves its site through a proxy which accepts connections on its behalf.

➢ YouServ uses dynamic DNS service for mapping the domain name to the machine of the user to whom the domain name belongs or to the machine of another peer node capable of serving content from a site replica or map to a machine serving as proxy for the site.

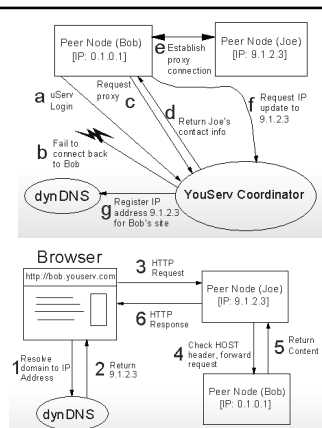➢A peer node that can accept inbound connections comes online.

➢Accessing content from a peer node that can accept inbound connections.

➢A peer node goes offline and another node that is replicating his site takes over.
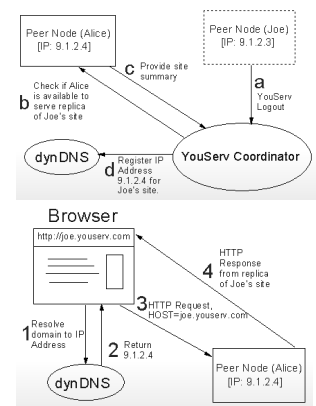
➢Accessing content from a site whose owner is offline.

➢A peer node that cannot accept inbound connections (Bob) comes online.

➢Accessing content from a peer node unable to accept inbound connections.

REPLICA MAINTANENCE

➢ Peer nodes are themselves entirely responsible for the bulk of replica maintenance.
➢ Replicator sites (sites which host replicated content) initiate contact with their Master sites, and also initiate content synchronization when necessary.
➢ A replicator determines when its replicated content is out of date by periodically (3 mins by default) comparing a short summary of its replicated content with the master's summary.
➢ For each file that needs to be updated, the replicator site will download the entire file from the master site using a standard HTTP GET request.
➢ Replicators monitor the availability of their masters and notify the coordinator if they go offline.
➢ Assist the coordinator in monitoring site availability.

OTHER ISSUES

- DNS Caching : YouServ DNS entries should be cached for at most the value of the TTL setting (2 minutes), allowing a replica to become accessible by users very shortly after it is activated by the coordinator.
- Scalability : The system bottlenecks are limited to the centralized dynamic DNS and YouServ coordinator components.
  - → To reduce DNS traffic in YouServ, the coordinator could in addition be reprogrammed to recognize sites that use static IP addresses and rarely if ever fail over to replicas, and increase their TTL values accordingly.
  - → Even more capacity could be provided by running additional coordinator instances on other servers.
- Security : When someone whose identity is unknown attempts to access secured content, YouServ forces a redirect to the main YouServ authentication site, where the person attempting to access the site must provide his or her ID and password.
- YouServ is written in Java which makes it robust (if not immune) to buffer overflow attacks such as those used by Code Red and other hacking tools to install unauthorized programs. Execution of scripts is also not allowed.

USABILITY

- Makes publishing content on the web as easy and universal as accessing it.
- As it exploits existing web protocols, YouServ content can be accessed with any standard web browser without installing special software.
- By relying primarily on existing desktop infrastructure, the YouServ service can be provided at an extremely low cost.
- Currently handling over 2900 users, the system is projected to scale to at least tens of thousands more lending credence to it being a superior alternative to paid hosting services.

Make it Fresh, Make it Quick -- Searching a Network of Personal Webservers

Roberto J. Bayardo Jr., Mayank Bawa, Sridhar Rajagopalan, Eugene J. Shekita
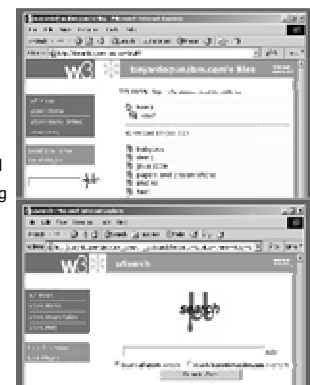
OBJECTIVE

- Propose and describe *YouSearch*, an easy to use, low cost application that seamlessly integrates with a personal webserver
- Was architected to be extremely simple and released within the IBM intranet in mid September 2002
- YouSearch provides fresh, fast and complete searches over personal webserver-hosted content
- YouSearch—enabled webserver allows
  - → Easy publishing of personal content
  - → Easily enabled web searches over published content
- Is important as the searches offered by Web's existing search tools are inadequate being based on crawling
- Location by navigation cannot be used as content on personal webservers is highly transient, poorly arranged and not cross-linked as other HTML documents.
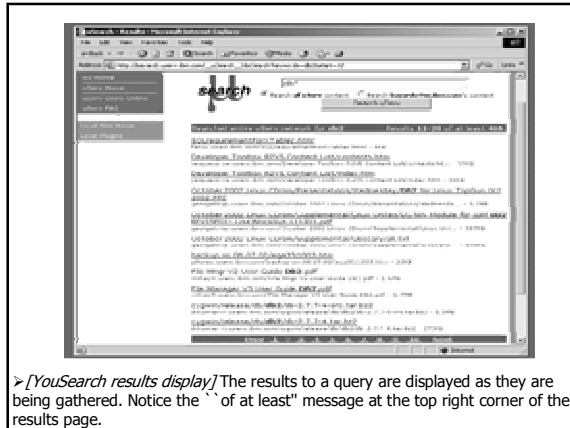
RELATED WORK

- Gnutella performs searches by flooding each query to all the hosts in the network reachable within a fixed horizon imposed by a time-to-live parameter on query packets
- KaZaa improves on this basic scheme by delegating most of the work to *super peers* - machines in the network with high bandwidth and large compute resources at their disposal
- Napster provided search over peer-hosted music files by adopting a hybrid scheme that is similar to YouSearch
- PlanetP is a research project that proposes another searchable P2P network.
- BadBlue can be used to search over a (private) network of hosts running BadBlue or other-Gnutella compliant software.
- Coopnet, Backlash & Proofs are proposals for a P2P caching scheme to address flash crowds at a host.
- Squirrel and MangoSoft's CacheLink product propose organizing a P2P cache for web objects
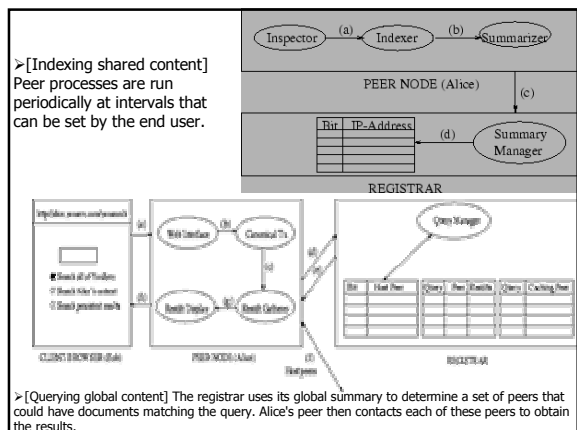
User Interface

- *[YouSearch interface]* A user browsing a YouSearch enabled website can issue queries using the above form-based search interfaces. **Sharing Searchable Content**

➤[YouSearch results display] The results to a query are displayed as they are being gathered. Notice the ``of at least'' message at the top right corner of the results page.

➤YouSearch offers no single centralized query form, each participating host offers its own web-accessible search interface

➤YouSearch peers by default preserve index only publicly shared, non-hidden content

➤Files are searchable not only via the keywords that appear in their filenames/URLs, but also through the content of the files

➤While the user views the available search results, the system continues to gather additional results in the background

➤The hits from each individual host are appropriately ranked using standard text search metrics

➤YouSearch can restrict searches to local hosts and also supports user defined "groups"

➤Has low barrier to conducting searches as YouServ is mostly used for constructive purposes and each user can be identified through the URL of the file they share

ARCHITECTURE

- ➤ peer nodes who run YouSearch-enabled clients
- ➤ browsers who search YouSearch–enabled content through their web browsers
- ➤ Registrar which is a centralized light-weight service that acts like a "blackboard'' on which peer nodes store and lookup (summarized) network state
- ➤ Each peer node closely monitors its own content to maintain a fresh local index
- ➤ A bloom filter content summary is created by each peer and pushed to the registrar
- ➤ The querying peer first queries the summaries at the registrar to obtain a set of peers hosting relevant documents
- ➤ The results for the query issued by the above peer are cached for a limited time
- ➤ This cache entry is notified to the registrar so that other peers can use the results, for a similar query
- ➤ Peers in YouSearch use the centralized registrar as a blackboard by posting the state of their node content for other peers to query


➤[Indexing shared content] Peer processes are run periodically at intervals that can be set by the end user.

➤[Querying global content] The registrar uses its global summary to determine a set of peers that could have documents matching the query. Alice's peer then contacts each of these peers to obtain the results.

➤Bob's query is received by Alice's peer node via a web interface and is forwarded to a *Canonical Transformer* that converts the query into a canonical form consisting of sets of terms labeled with the associated modifier

➤The canonical query is forwarded to the *Result Gatherer*.

➤The Result Gatherer sends the canonical query to the registrar where the *Query Manager* computes the hash of keywords to determine the corresponding bits for each of the k bloom filters

➤The registrar looks up its bit position to IP address mapping and determines the intersection R of peer IP address sets. The set R is then returned to the querying peer (Alice).

➤The Result Gatherer at Alice's peer obtains R

➤It then contacts each of the peers in R and obtains a list of URLs U for matching documents

➤The results are then passed to *Result Display* which then appropriately formats and displays U

➤If the Result Gatherer recognizes the query to be a local query it looks up its local index to find documents that match the query.

CACHING QUERY RESULTS

➤Every time a global query is answered that returns non-zero results, the querying peer (Alice's peer node) caches the result set of URLs U

➤The registrar is then informed of the fact

➤The registrar then adds a mapping from the query to the IP-address of the caching peer (Alice) in its cache table

➤The caching peer itself monitors and expires entries in its cache based on their lifetime, and informs the registrar of any such changes

➤Each global query result is displayed with a check box that allows the peer to indicate if he found the result relevant and would like to recommend it

➤On doing so the query and the selected result are sent to the registrar who maintains such mappings from query to the recommended URL

➤The information pool is editable, allowing peers to modify and augment shared information

➤Each peer informs the registrar of any changes in its index or cache entries

➤Whenever a peer leaves the network, it asks the registrar to remove its entries from the index and cache tables
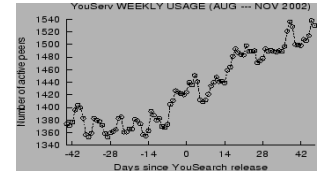
PERFORMANCE ANALYSIS
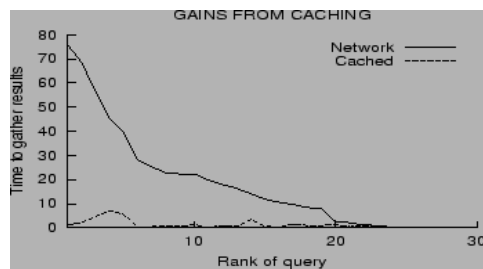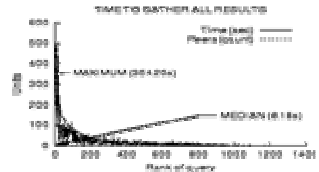
➢Distribution of session durations of YouServ peers.

PEER SESSION DURATIONS (AUG --- NOV 2002)
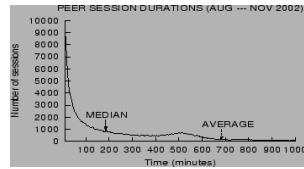
➢The top 15 of 43 countries with YouSearch users.

GEOGRAPHIC DISTRIBUTION OF PEERS
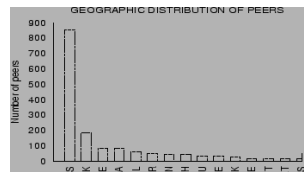
➢Characteristics of bloom filters from approximately 340 peers.

YouServ WEEKLY USAGE (AUG --- NOV 2002)

➢Time taken to gather all results

TIME TO GATHER ALL RESULTS

GAINS FROM CACHING

Results with and without caching

ADVANTAGES

➢It scales gracefully and costs little since its centralized resource requirements are small

➢As it is web compatible at it core hence it closely mimics the existing user experience of conducting a web search

➢Enhances the shared context among its users

➢Personal webservers can be aggregated into overlapping, user specified groups, and these groups search just as individual nodes

➢Any group member can persist result recommendations so that others can draw upon its knowledge

➢With the use of tunable parameters we can tune the network as it evolves

➢Unlike purely centralized search architectures, the plethora of compute, storage, and bandwidth available to the set of YouSearch peers as a whole puts few constraints on further enhancement

---

CRITICAL EVALUATION

➢ Still has relatively a very low profile compared to other sharing programs with the user base being mostly IBM employees and some students

➢ Remains to be seen if it can burst into the scene and penetrate the market

➢ Is not exactly cutting edge in terms of the technology it employs with more emphasis being given to user friendliness

➢ The full set of functionality described in the paper for YouSearch, hasn't been deployed yet

➢ Eventually legal issues will crop up as always

QUESTIONS

1) How does YouServ help users who are unaware of the exact URL of file?

2) How does YouServ deal with the problem of a hosting machine going offline?

3) If the firewall of the users machine disallows inbound port 80 connections how does YouServ deal with it?

4) How does YouSearch benefit users sharing the same context and with what operator?

5) How does YouSearch handle Repetitive queries?