Network Services

POP3 / IMAPv4
SPAM Filter
FTP

Johann Oberleitner
SS 2006

Agenda

- POP3
- IMAPv4
- Spam
- FTP

Message Storage Access

- Protocols to access mailbox
  - POP3 (RFC 1939)
  - IMAP4 (RFC 3501)
- Separate protocols
  - Primary idea
    - SMTP server not feasable to install on all machines
      - Resource consumption
      - Not always online
    - Use POP3/IMAP to access centralized mailbox (maildrop)

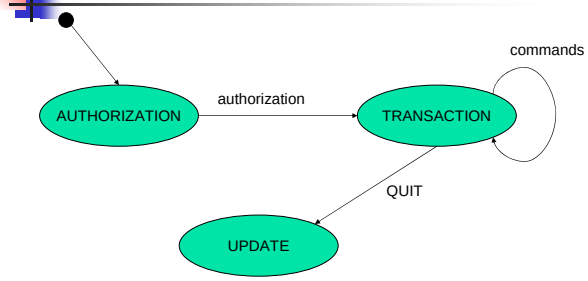
POP3 / 1

- Post Office Protocol
- Primary mechanism
  1. Download Mail from server (into email client)
  2. Delete from Server
- POP3 server listens on TCP Port 110

POP3 / 2

- Each POP3 session is state-based
- AUTHORIZATION state
  - Wait on authorization info
- TRANSACTION state
- UPDATE state
  - Removes mails from server maildrop

POP3 / State Machine



POP3

- Commands similar to SMTP
  - Keyword & text-based
  - Multiline responses end with "."
  - All commands terminated with <CRLF>
  - Each message has an id number

POP3 / 3

- USER & PASS (AUTHORIZATION)
  - Mailbox & Password – plaintext(!)
- APOP name digest (AUTHORIZATION)
  - Alternative to USER & PASS
  - Calculates shared secret based on server greeting (that must contain unique timestamp)
- STAT (TRANSACTION)
  - Status – information about number of messages in maildrop
- LIST [msgNr] (TRANSACTION)
  - Scan listing for (all) messages
    - Message number & message size in octets (=bytes)

POP3 / 4

- RETR msgNr (TRANSACTION)
  - Retrieves the contents of a message
- DELE msgNr (TRANSACTION)
  - Marks messages for deletion
- RSET (TRANSACTION)
  - Removes any deletions marks from a message
- TOP msgNr n (TRANSACTION)
  - Retrieves header + first n lines of body of a message
  - Important for retrieving header
- QUIT (TRANSACTION)
  - POP3 server removes all messages marked as delete

POP3 / Telnet Trace

```
C:  <open connection>
S:          +OK POP3 mail.xyz.at server ready
C:          USER joe
S:          +OK User name accepted, password please
C:          PASS blabla
S:          +OK Mailbox open, 20 messages
C:          LIST 20
S:          +OK 20 2696
C:          TOP 20 1
S:          +OK Top of message follows
....
C:          RETR 20
S:          +OK 2696 octets
...
C:          DELE 20
S:          +OK message 20 deleted
C:          QUIT
S:          +OK Sayonara

C:  Connection to host lost
```

POP3

- POP3 has no builtin support to distinguish between different types of emails
  - Example: no builtin support to distinguish between seen and not yet seen messages
  - Up to the (mail client)) application to determine which messages are new

IMAP4 / 1

- Internet Message Access Protocol
- IMAP4rev1
  - last release
- More features than POP3
  - Operations for Mailbox administration
  - Checking for new messages
  - Searching for messages
  - Message Flags
- IMAP4 server listens on TCP 143

IMAP4 / 2

- Keyword & text-based
  - All commands terminated with <CRLF>
  - Commands begin with unique identifier (tag)
    - Eg. A0001 SELECT mymailbox
  - Two different type of Responses
    - tagged response
      - Same tag as command was sent from client
      - Indiciates response for a command (eg. A0001 OK+)
    - Untagged response
      - Server messages that do not occur from commands
  - Client may have to send continuation data

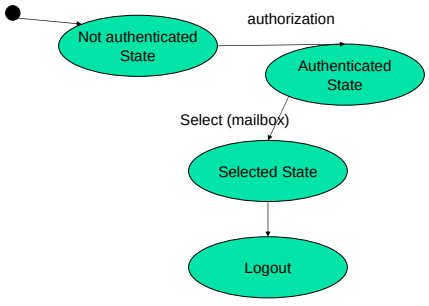
IMAP4 / 2

- Each message
  - unique identifier
    - MUST not change during a session
    - SHOULD not change between sessions
  - message sequence number
    - Relative position from 1 to number of messages in a mailbox
    - May be reassigned during a session

IMAP4 / 2

- Flags Message Attribute
  - 0-n named tokens associated with a message
  - Permant & Session-only flags
  - System flags = predefined
    - \Seen
    - \Answered
    - \Flagged (urgent/special attention)
    - \Deleted (marked as deleted)
    - \Draft (marked as draft)
    - \Recent (this IMAP session is first session notified about message)
  - Keywords
    - Not begin with "\"
    - Client may define new keywords in the mailbox

IMAP4/3



IMAP4 / 4

- Server
  - may send data at any time
    - Even if client did not request this data
  - Server MUST send mailbox size updates automatically
  - Untagged response while no command in progress
  - After some Inactivity (autologout time)
    - Automatic logout

IMAP4 / 5 – Client commands

- Any state
  - CAPABILITY
    - Requests listing of capabilities server supports
  - NOOP
    - No Operation
    - Prefered method to lookup new messages or status updates
  - LOGOUT
    - Server sends untagged BYE
    - Afterwards server sends tagged LOGOUT response

IMAP4 / 6 – Client commands

- Not authenticated
  - LOGIN
    - Plaintext password authentication (user name & password)
  - STARTTLS
    - Starts TLS/SSL negotation
    - On success all further commands under TLS layer
  - AUTHENTICATE
    - Indicates a SASL authentication mechanism to server
    - Server performs authentication protocol exchange to authenticate end identify client
    - May negotiate optional security layer for subsequent protocol interactions

IMAP4 / 7 – Client commands

- Authenticated State
  - SELECT mailbox
    - Selects a particular mailbox for subsequent requests
    - Only one mailbox can be selected in one connection
  - EXAMINE mailbox
    - Like SELECT, but read-only
  - APPEND mailbox messageData
    - Appends message to a mailbox
  - LIST refName mailboxName
    - Lists mailboxes relative to refName (eg. filePath)
  - Mailbox administration commands
    - CREATE,DELETE,RENAME

IMAP4 / 8 – Client commands

- Selected State
  - Based on currently selected mailbox
  - CLOSE & EXPUNGE
    - Removes all messages with \Deleted flag
    - Expunge sends untagged EXPUNGE response for each deleted message
  - SEARCH
    - Searches the mailbox for messages that match certain criteria (see RFC 3501 6.4.4)
  - FETCH
    - Retrieves data associated with a message (eg. Header, Body)
  - STORE
    - Alters data associated with a message

IMAP4 / 9 - Sample

```
C:  <opened connection>
S:  * OK [CAPABILITY IMAP4REV1 ...] mail.xyz.at
C:  A001 LOGIN joe mypasswd
S:  A001 OK [CAPABILITY IMAP4REV1 ...] User joe authenticated
C:  A002 SELECT mail/IEEE
S:  * 11 EXISTS
    * 0 RECENT
    * FLAGS (\Answered \Flagged \Deleted \Draft \Seen)
    * OK [UNSEEN 10] first unseen message in /home/joe/mail/IEEE
C:  A003 SEARCH ALL
S:  * SEARCH 1 2 3 4 5 6 7 8 9 10 11
    A003 OK SEARCH COMPLETED
C:  A004 FETCH 2:4 (BODY[HEADER])
S:  * 3 FETCH (BODY[HEADER]) {1085}
    ... mail messages ...
    A004 OK FETCH completed
C:  A005 LOGOUT
    * BYE mail.xyz.at IMAP4rev1 server terminating connection
    A006 OK LOGOUT completed
```

Message Disposition Notification

- RFC 3798
- Inform humans of the disposition of the message after successful delivery
- Additional message header field
  - "Disposition-Notification-To:"
- Sent as MIME message
- Problems:
  - Forgery (as regular emails)
  - Privacy
  - Non-Repudiation
  - Another way for Mail-bombing

Message Disposition

- Better solution
  1. Put message on Web server
     - Special URL that stores the message
  2. Send secret URL via email
  3. URL only accessible once

Phishing

- Sending an email to a user claiming to be another sender
- Attempt to acquire private information from the user
  - Passwords
  - Pins
  - Credit Card Numbers
  - Bank Account Numbers
- Frequent attempt
  - HTML Links in HTML emails
    - www.amazon.com?
  - Link appears as www.amazon.com but links to 66.22.33.22
- Simple Solution
  - Don't use HTML emails

Spam

- Different meanings
  - Unsolicited Bulk Email
    - Massive number of recipients
    - Unsolicited!
    - Primarily Mass mails with commercial content (other Name: Unsolicited Commercial Email)
  - Fraud emails (Nigeria Connection)
    - Chain letter via email
  - Nonsense Postings in Internet forums (Trolling)

Spam - Principles

- Internet has a friendly nature
  - Email sent back to sender when receiver does not react/exist
  - Otherwise error message to postmaster
- Spam
  - Sends emails to huge number of potential recipients
  - Postmaster gets error message for non existent addresses
  - Removes these addresses from recipient list

Spam – Countermeasurements / 1

- Mask published email addresses
  - on Web pages
  - "email: joe at infosys dot infosys dot ac dot at"
    - Frequent pattern & rather weak (easily analyzable)
  - Better something like this:
    - "email: name@domain where name = joe and domain = infosys.tuwien.ac.at"
- Complain about spammer at the spammer's provider
  - Often same person
  - Provider in foreign country
  - Spammer is a client of the provider

Spam – Countermeasurements / 2

- Legal measurements
  - Accusing spammers
  - Possible for large companies
  - Only if spammer works in developed countries
  - Slow
  - First success stories
- Filtering based on Content and Format
  - In control of end-user
  - In control of end-user's provider
  - Today most successful
  - Does not fight Spam at the originator

Spam Filtering

- Scan on MTA
  - Good place for centralized checks
  - User specific settings cannot be used
- Scan on MDAs / Message store
  - Supports user specific configurations
  - Move Spam to particular mailbox
  - Spam verification done only after message received the system
  - Has to be installed & maintained on every system
- Problem – Different kind of users
  - Some don't want spam
  - Some want all emails
  - Legal problem of NOT delivering emails
    - Eg. German university

EMail Classification for SPAM

- HAM = Real-Negatives
  - Message is no SPAM
- SPAM = Real-Positives
  - Message is SPAM
- False-Positives
  - Message classified as SPAM but isn't
- False-Negatives
  - SPAM, not marked as SPAM
  - Goal of Spam Filtering is to minimize False-Negatives

Heuristic Filtering

- Set of common rules to specify characteristics of spam
  - Rules are preconfigured
  - May be written by administrators
- Problem
  - Everyone uses a similar set of rules
    - Spammers can react on this
- Example
  - SpamAssassin (without Bayesian Filtering)

Sender Policy Framework

- SPF
  - At potential sender domain
  - To allow reverse MX records
  - Mail receiver can query DNS if sending host was authorized
  - http://www.openspf.org/mechanisms.html
- Additional records for DNS
  - Uses TXT resource record, starts with v=spf1
- Prevents not Spam, but forgery
- Example:

maydomain.com IN TXT "v=spf1 +ptr –all"

  - Means: "sender was authorized if its IP address can be reverse looked-up within the sending domain (+ptr) (via PTR DNS queries), fail in all other cases (-all)"

Spam Lists

- Lists contain sender
  - domain names
  - Email addresses
- Whitelists
  - Don't want email filtered
- Blacklists
  - Emails are Spam
    - Eg. DNSBL: emails sent or relayed from certain hosts are very likely Spam

Statistical Filtering

- Based on 3 components
  - Historical dataset
    - Stores the corpus = total of user's email set
  - Tokenizer
    - Splits email into tokens
  - Analysis engine
    - Provides result if email is spam or ham

Statistical Filter - Process

1. Tokenization of the email
   - Usually on word boundaries
     - Some filters support word chaining (two word tokens)
     - Some filters support phrases
   - Assigning token values (from 0.0 – 1.0)
2. Construction of a decision matrix
   - Consists of 15 – 27 of the most interesting tokens (peak values, with largest distance from 0.5)
3. Evaluating decision matrix

Tokenization

- Example:
  - Spam mail: "Buy an academic degree!"
- Tokens:
  - Buy,an,academic,degree!
    - Sometimes: "degree" and "degree!" are considered as different tokens

Grahams Approach for assigning token values

- Assign token values based on values in historical dataset:
  - SH = total number of appearances of a token in all spam mails
  - IH = total number of appearances of a token in all innocent mails (HAM)
  - TS = total number of spam mails in users corpus
  - TI = total number of HAM mails in users corpus
  - P = (probability that token is identifier for spam)

$$= \frac{SH / TS}{SH/TS + IH/TI}$$

Graham's approach

- Biasing
  - Reduce number of false positives by doubling number of occurences for a token

$$\frac{SH / TS}{SH/TS + 2*IH/TI}$$

Total Spam (TS)	250
Total Ham (TI)	118

Token	Spam (SH)	Ham (IH)	P	Biased P
degree!	46	3	0,8786	0,7835
an	17	53	0,1315	0,0704

Decision Matrix

Token	Spam	Ham	Probability
degree!	46	3	0,7835
an	17	53	0,0704
…	…	…	…

Tokens are sorted based on its distance from 0.5 (= absolute value of (0.5 – P)), means that significant tokens (Spam identifying and Ham identifying are considered)

Bayesian Combination

- Combine N first values of sorted decisision matrix with bayesian statistics

$$\frac{A * B * \ldots * N}{A*B* \ldots * N + (1 - A) * (1 - B) * \ldots * (1 - N)}$$

- Relatively extreme values
- Graham uses 15 first values
- Brian Burton uses 27 first values
  - A single token may populate two slots if it appears at least two times in a message
    - Leads to better results for small messages

Bayesian Filtering

- Requires training phase
  - Collection of messages that are definitively SPAM
  - Collection of messages that are definitively NO-SPAM
  - Finds token in messages based on these mesages
    - Words or word groups
- Known Statistical Filters:
  - SpamProbe
  - DSpam

File Transfer Protocol (FTP)

- RFC 959
- Already from 1971(!), RFC 114
- Goal: File transfer from one host to another
- Based on 2 connections
  - Control connection (server listens on TCP port 21)
    - Transfers commands
  - Data connection created each time a file is transfered
    - For Data transfer
- Uses TELNET NVT protocol on control connection
- Limited number of file types supported
  - ASCII, Binary

Active FTP

1. Client initiates connection to server control port
2. Client opens random data port for listening
3. Server connects to this open client data port with its own port 20

- Firewall problem
  - Server has to go through client firewall

Passive FTP

1. Client initiates connection to server control port
2. Server listens on data port
   - NOT port 20 (!)
3. Client connects to open data port
- Not all FTP clients/servers support passive FTP

FTP commands

- Access control
  - USER & PASS
  - CWD (change working directory)
- Transfer Parameter commands
  - PORT – specifies data port
  - PASV – passive mode
  - TRANSFER MODE (stream,block,compressed)
- Service Commands
  - RETR - retrieve a file
  - STOR - store a file
  - LIST – list files
  - ...

Summary

- Email Access Protocols
  - POP3
  - IMAPv4
- Spam
- FTP