Overview

- Background
- Requirements and design issues
- Herald and WING design
- WING (Web IMAP/NNTP Gateway)
- Herald implementation
- Herald reliability and problem history
- Some statistics
- Fitting the requirements
- Linux v Solaris

Background

- Before Herald, two central general purpose systems handled most University mailstore: Sable (18000 users) and Ermine (8000 users)
- Both DEC/Compaq AlphaServers running Compaq Tru64 Unix (The O/S Formerly Known As "Digital UNIX (The O/S Formerly Known as OSF/1)")
- Sable horribly overloaded (550 concurrent users on what was classed a Workgroup Server when it was EOLed last year)
- Decision to investigate setting up a separate mailstore server

Requirements

- Mail store for 30000 users
- Reliable
- Secure
  - against break-in *and* denial of service attacks
  - against attacks from the outside world *and* from our users
- IMAP and SMTP
- Browser interface available (but not compulsory) with no reliance on Java, Javascript, frames or plug-ins
- Quotas (one user's mistake/greed must *not* be able to affect mail delivery to any other user)

Design issues

- All-in-one software v. mix-and-match software
- Protocols: proprietary v. open
  - Open protocols (IMAP, SMTP, ...) are of critical importance for inter-operation with the Internet and support must be absolute
- Software: proprietary v. open source
  - Oxford's decentralised structure (a euphemism) tend to confuse/confound (ditto) the support channels of most proprietary software
- OS: Unix v. VMS v. OS/400 v. OS/390 v. ... v. NT
- Hardware: single server v. cluster

Sane design possibilities

- O/S: Unix
- Hardware: single server (e.g. Sun Enterprise E3K or E4K) or cluster of workstation-class machines.
- Narrowing down choices from other considerations leads to
  - O/S: Linux or Solaris
  - Hardware: SPARC or PC
- Possible configurations
  - Netscape Enterprise Server (or similar) on a Sun E3K/E4K
  - Mix-and-match
    - Cyrus/UW imapd plus Exim/Qmail/Postfix/... plus ...
    - Sun E3K/E4K or a workstation cluster

Herald and WING design

- Mix-and-match software on a workstation cluster
- UW imapd
- Exim
- Apache/mod_perl
- WING (Web IMAP/NNTP Gateway)
- PostgreSQL
- BIND
- Home-brew account and cluster management tools

WING (Web IMAP/NNTP Gateway)

- Open source, written in Perl for Apache/mod_perl
- Handles MIME attachments
  - Main message display lists attachments
  - Clicking on one sends it to the browser with the right MIME type
- Can MIME attach browser-local files when composing messages
- Allows arbitrary headers to be added to composed messages *except* for the "`From:`" header which cannot be changed/forged and is set to the user's canonical email address.
- Users can create and search any number of address books which can be shared with other users via ACLs

- Account management: changing password, querying disk usage and quota, setting mail forwarding and "vacation" messages
- Optional portal view (in testing) provides site-configured and personal links in a folding permanent list down the left hand side (requires frame support from browser)
- Scales up to thousands and tens of thousands of users
- See `http://users.ox.ac.uk/~mbeattie/wing.html`

Herald topology

- Two frontends: the only connection to the outside world
  - Act as firewalls
  - Redundancy and failover for reliability
  - Hold centralised information for cluster
  - Act as nameservers for `herald.ox.ac.uk` domain
- 24-port 100 Mbps switch connects all cluster nodes
- IMAP servers: hold all per-user filestore
- WING servers: hold only temporary per-session state
- Each user has a DNS entry `username.herald.ox.ac.uk`, an alias (CNAME) for their home IMAP node

Herald services

- IMAP servers
  - serve IMAP, POP, SMTP to public; export NFS to other cluster nodes
- WING servers
  - serve HTTP to public; talk client-side IMAP and NFS (to IMAP servers) and SQL (to frontend)
- Frontends
  - Serve DNS and act as SMTP and HTTP login gateways.
  - DBMS holds all user configuration data
  - Accept connections from User Registration server for automatic account creation and modification
  - Generate mailer tables and push them to other nodes

Herald security

- Firewalling
  - Frontends act as firewalls: only specific public service traffic is allowed through to cluster nodes
- Intra-cluster trust
  - IMAP and WING servers trust frontends completely (e.g. for rsh and NFS)
  - IMAP servers export `~foo/wing` directory owned by httpd for each user `foo`.
  - A break-in to the httpd account on WING servers leads to the ability to mess up all users' forward files, vacation messages and personal links but *not* their mail

If a Herald slave node dies

- Failure of a WING node means:
  - Current sessions on that node die (e.g. lose draft message)
  - 1 in n login attempts fail. Remove dead node from `/etc/wing.live`. Can/will be automated.
- Failure of an IMAP server node means:
  - Mail store is on RAID5, single disk failures do not matter
  - If node dies, all users on that node lose mail access
- Failure of the switch means:
  - Dead in the water until we plug in a new one.
  - Could use Beowulf channel bonding with two separate switches

If a Herald frontend dies

- DNS continues without a break
- IP traffic through to internal nodes (IMAP, HTTP) continues after a break since both frontends proxy arp for internal nodes. Failover can/will be auto-detected to minimise break.
- SMTP gateway service continues after same break (both frontends serve on same IP alias)
- SQL failover currently manual (bring up daemon on other frontend on same IP alias). Can/will be automated. Lose config changes since last sync.

Herald hardware: phase I

- Two frontends
  - PC/Linux, 512 MB RAM, 300MHz CPU, 2 x 6 GB disks
- Two IMAP servers
  - SPARC/Solaris, Ultra5, 320 MB RAM, 270MHz CPU, A1000 disk array with 6 x 4GB (10K RPM) disks (hardware RAID5)
- Two WING servers
  - PC/Linux, 512 MB RAM, 2 x 6GB disks
- One tape drive
  - DLT7000 uses DLT4 tapes (35 GB raw, 70 GB compressed)

Herald hardware: phase II and beyond

- Four more IMAP servers
  - PC/Linux, 512 MB RAM, 350MHz CPU, 6 GB disk, Mylex BT958D Differential Wide UltraSCSI-3 adapter
- Two D1000 disk arrays
  - 12 x 9GB (10K RPM) disks with 6 disks (Linux software RAID5) per bus, each bus connected to an IMAP server.
- Four more WING servers
  - PC/Linux, 512MB RAM, 350MHz CPU, 6 GB disk
- Summer '99
  - will replace Sun/Solaris boxes with PC/Linux

Herald software

- Almost all additional tools needed are in Perl
- Frontends
  - Apache/mod_perl, Exim, PostgreSQL, BIND
  - useradmind daemon acts on account management requests
  - Mailer and DNS tables auto-generated daily from cron
  - crsh, a simple "cluster rsh" helps push stuff around
- IMAP servers
  - UW imapd suite, Exim
- WING servers
  - Apache/mod_perl, WING, Exim

Herald software II

- Linux IMAP servers use Berkeley DB `passwd.db` and `shadow.db` in NSS for account information
  - How to use Berkeley DB in NSS is poorly publicised
  - READMEs in glibc source distribution reveal some
  - Reading the source reveals more
  - Wrote utilities {`user`,`shadow`}{`add`,`del`,`mod`}`db` to manipulate the files
- Autofs games let WING servers access cluster-wide NFS
- Patch to UW imapd ensures “black box” mode in correct directory

Frontend server problems

- In service since Sept ’98
- No problems
- No outages

IMAP server problems

- Sun servers in service since Sept ’98
- Linux servers in (very light) service since Jan ’99
- Two unexplained panic-and-reboot events on Sun IMAP servers
- No other outages

WING server problems

- In service since Sept ’98
- Early version of WING (last November) had an SQL deadlock race (my fault)
- During deadlock (few times a week), new WING logins not possible
- Restarting PostgreSQL daemon fixed it as workaround
- One weekend, 350 deadlocked httpd processes created on each WING server caused out-of-memory crashes on both
- SQL bug found and fixed within a few weeks
- No other outages before or since

SMTP drop-off server problems

- In service since Sept '98
- Frontend hosts SMTP on `smtp.ox.ac.uk` so that dumb clients can drop off their outgoing messages (Oxford mail hubs concentrate on throughput, not latency)
- Occasionally (three or four times over the last few months), the exim daemon dies silently
- Incoming mail to cluster is delivered directly to IMAP nodes and is not affected
- Cause unknown. Upgrade and auto-restart script planned

WING client bugs

- A few subversions of Internet Explorer mangle URLs by converting multiple consecutive slashes to a single slash
  - Upgrading IE fixes it
- Clients which use web proxies which source HTTP requests with different IP addresses within one WING session make WING server display a security alert
  - Workaround: disable proxying or choose a better ISP
  - Introduction of HTTPS and a port 81 will also help
- Various minor software bugs, niggles and feature requests
  - Surprisingly few (well, it surprised me anyway)
  - WING seems to be rather solid now

IMAP client problems

- Error in imapd “blackbox” patch (my fault) caused two new IMAP clients to induce imapd NULL-dereference
- Bug found and fixed

Some statistics

- During one recent "average" week, two IMAP servers and two WING servers handled:

```
82000 incoming messages to IMAP servers
113000 deliveries/outgoing by IMAP servers
95000 local deliveries by IMAP servers
18000 onward deliveries by IMAP servers
26000 outgoing messages from WING servers
66000 IMAP sessions (incl 38000 WING sessions)
120000 POP sessions
```

- Load averages hover near 0.00, occasionally hitting 0.1

Fitting the requirements

- Two IMAP reboots (disappointing)
- No WING outages since SQL deadlock bug fixed last November
- WING servers have already had transparent rolling hardware/software upgrades
- Fewer advisory calls: “How do I run pine?”, “How do I mail/attach a file from my PC?”, “How do I view/download this MIME attachment?”

Linux v Solaris

- Either would be adequate but Linux came out as preferred
- Too much necessary stuff is "third party" for Solaris
  - GNU grep, gzip, less, tar, make, gdb
  - Perl, vim, emacs, tcsh, top, lsof, Tcl, expect
  - imapd, Apache, mod_perl, PostgreSQL, tcpd, cracklib
- On a single large system, *maybe* it would be bearable to find/install them all but maintaining/upgrading them is another matter
- Solaris has journalling for its fs but ext2 fsck is adequate for now and expected number of crashes until ext2 journalling arrives is zero.
- Linux FSSTND/FHS filesystem standard is very useful

- RPM v SysV pkg tools
  - Some advantages of RPM are slight (e.g. speed, MD5 checksums)
  - One is major: can look at a package source dist, patch and rebuild RPM (e.g. for security, upgrade, local patch)
  - Debian dpkgs has same/similar advantages over SysV pkg
- Solaris has no built-in firewalling (or maybe it does now?)
- Cannot safely enhance systems level programs under Solaris
  - Enable/disable users? Retire/expire/force pw change?
  - Digital UNIX: SIA already permits it
  - Linux: Source shows locking strategy, dependencies, …
  - Solaris: no SIA, no source, PAM not enough: no solution
- Experience with Sun and DEC/Compaq support channels has convinced me that using source to fix any problems I might come across is definitely better