Diplomacy in Storage: Communicating Through the Proper Channels in Envoy

Russ Ross
rgr22@cl.cam.ac.uk
30 August 2005

The Goals

- VMM clusters have different storage demands
  - One big file system: like a regular cluster
  - Many little file systems: unrelated VMs
  - Implicit sharing: bandwidth costs and shared base images
  - Snapshots and clones: easy set up, backup
  - Use cheap hardware: disks in every machine
  - See Parallax for more

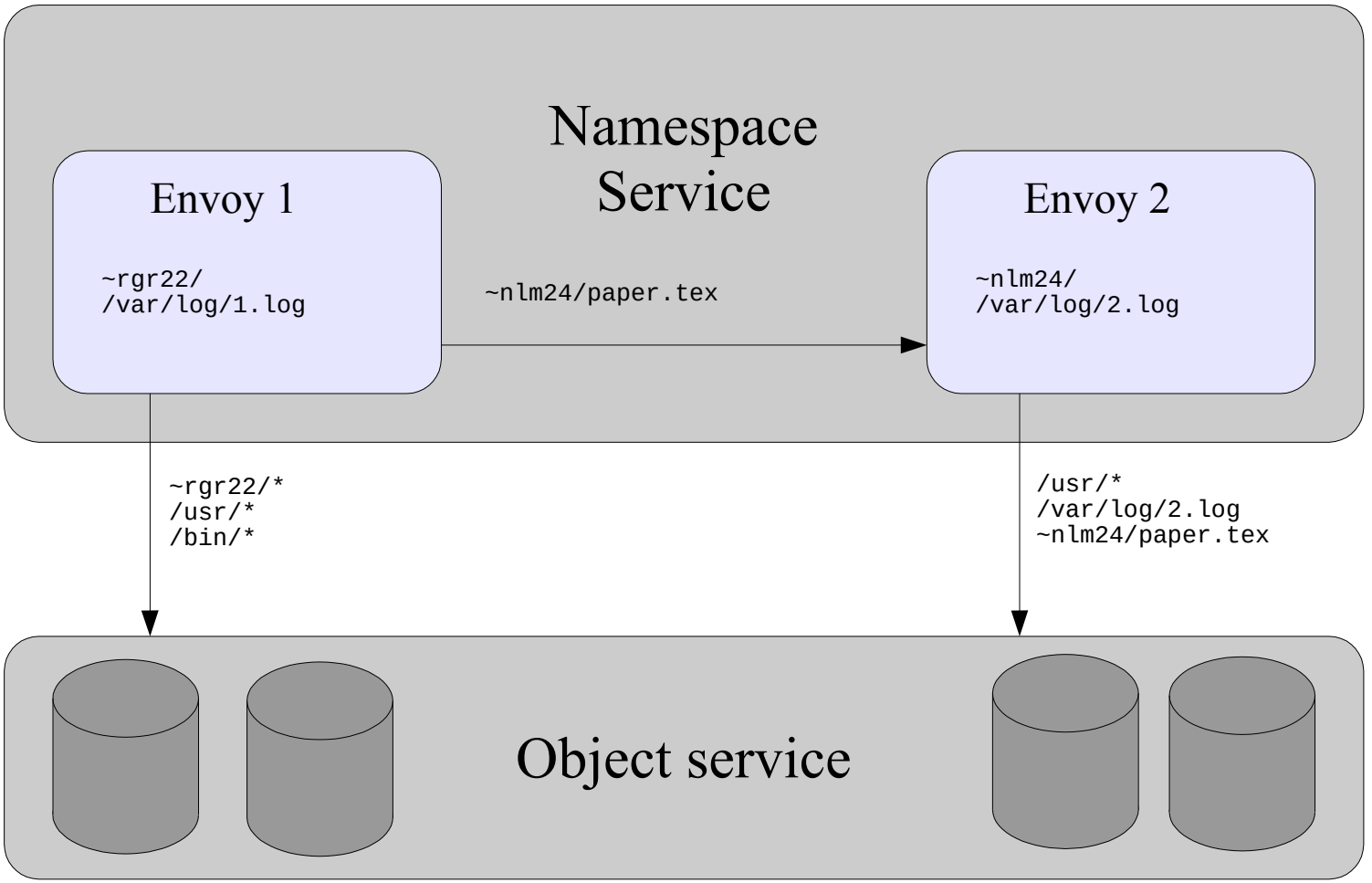
The Advantages

- VMM clusters give us some advantages
  - Can still scale at physical machine level
    - Zillions of users, but only thousands of hosts
    - Sharing cache is easy: put it in the admin VM
    - Staging area with persistent cache
  - Don't need to trust clients: trusted admin VM
    - We can punt on many security problems
    - ~~Microkernel~~VMM gives us (relatively) stable host
  - Target environment uses well-provisioned hosts
    - Stable network layout, too

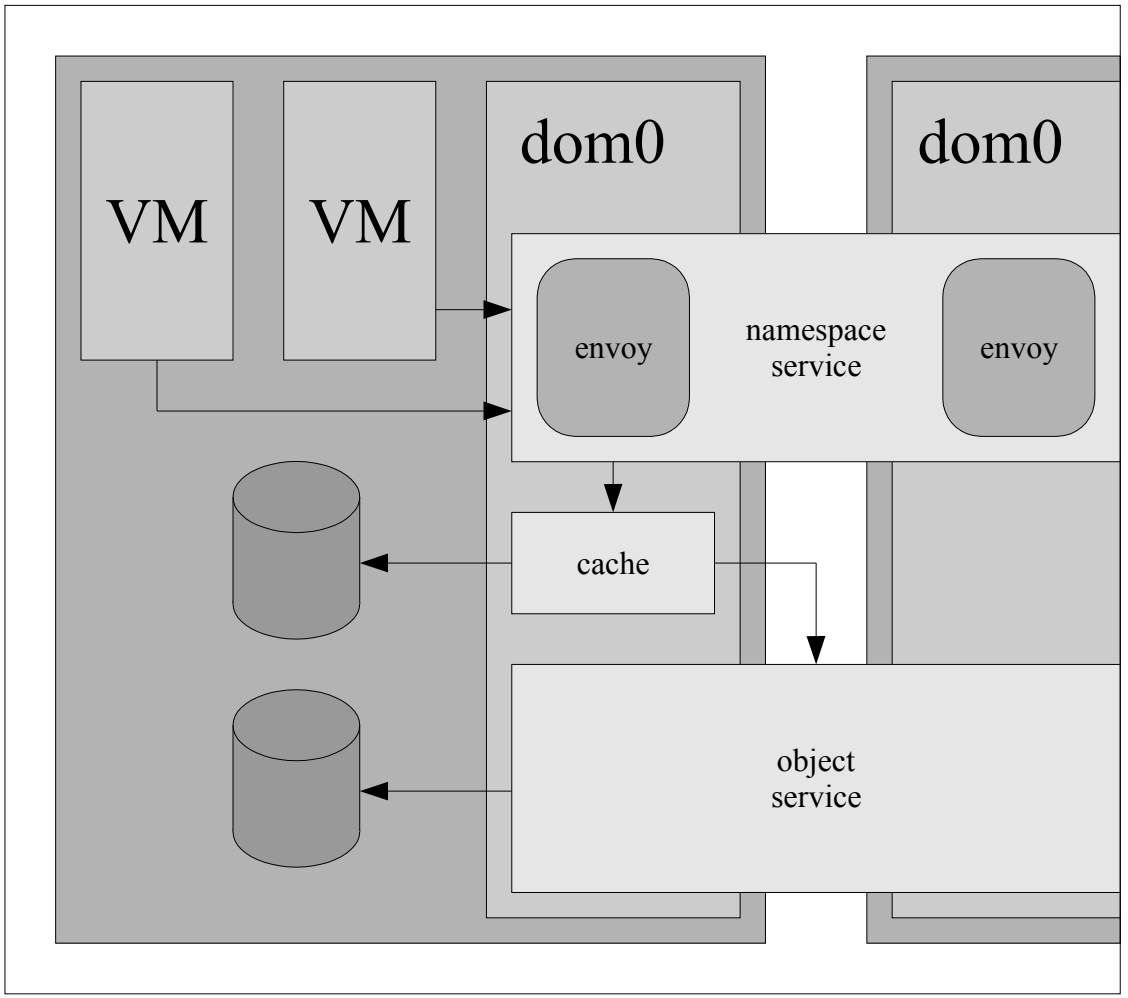
Distributed Storage Governments

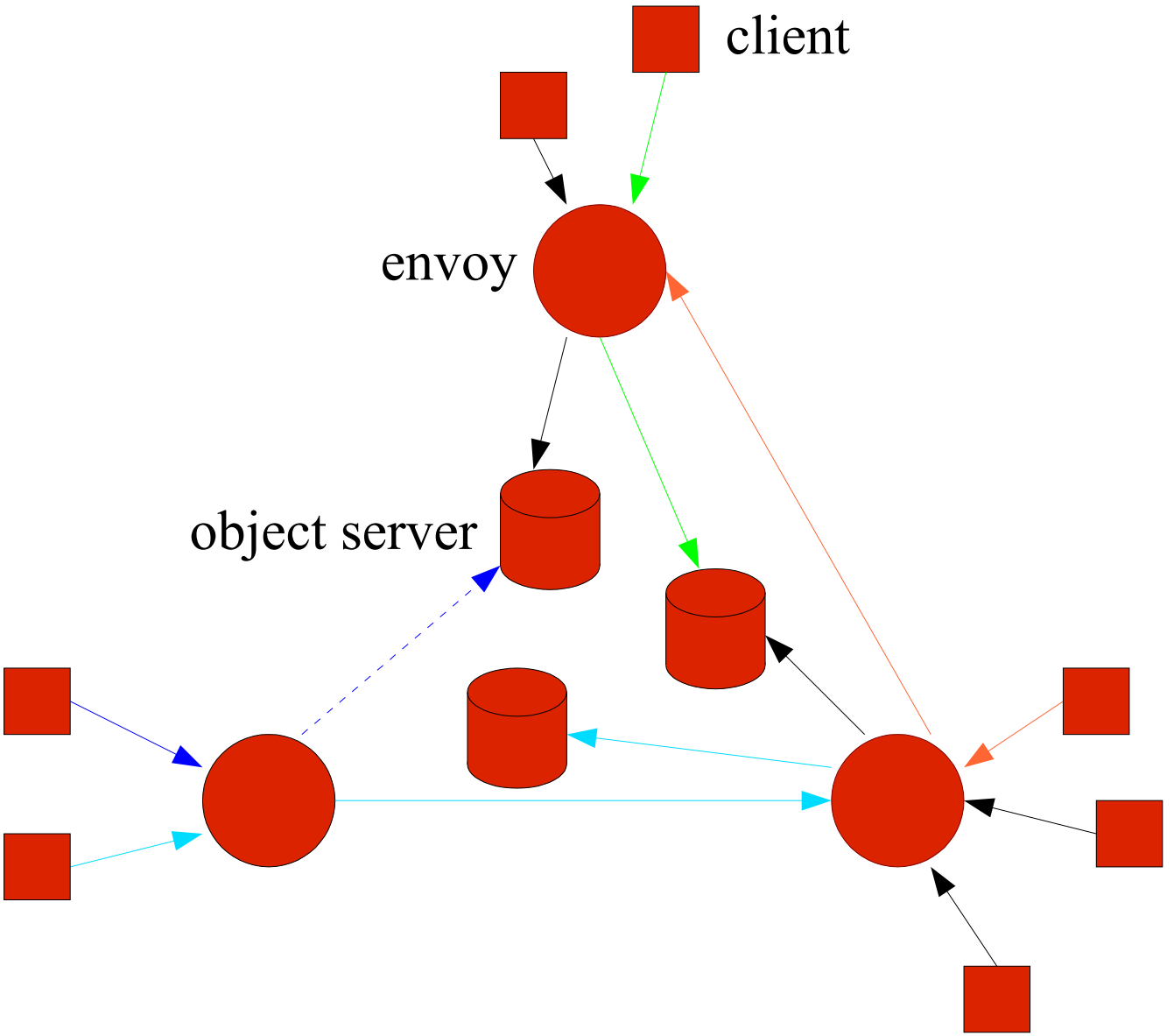
- Dictatorship
  - Single server owns the data with lot of caching
- Democracy
  - Speak directly to a majority of the replicas
- Representative democracy
  - Coordinate through metadata server
- Hippy commune
  - Share the data freely, watch out for the hairy guy
- Federation
  - Carve it up, work through the envoy from each territory

High-level architecture



Node architecture




client
envoy
object server

Steady State Data Paths

- Local territory
  - Persistent cache
    - 0 hops
  - Direct to object layer
    - 1 hop

Like Parallax for files

- Other territory
  - Envoy cache hit
    - 1 hop
  - Envoy cache miss
    - 2 hops
  - Weakly synced reads, direct to object layer
    - 1 hop

Like NFS over Parallax

Implementation

- Using 9P2000.u protocol (Plan 9 for Unix)
  - Simple, stateful protocol
  - No cache between client and envoy service
- Extension of 9P between envoys
  - Request forwarding, migration
  - Cache invalidation
- Linux 2.6, TCP (SCTP and Xen in future)
  - Started using OCaml, started over in C