SCALING NFS SERVICES

Tom Haynes
tdh@sun.com
Sun Microsystems, Inc.

Multiple Flavors per Export

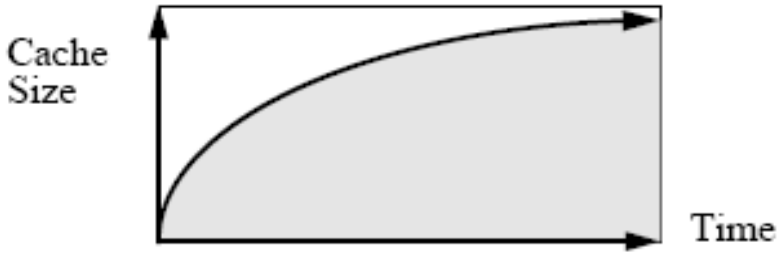
- Presented in Connectathon 1995
- Check every NFS request
- Netgroups
  - Not capped
  - Membership can change
- Use an Authentication Cache

The Authorization Cache

SunSoft
A Sun Microsystems Company

Kernel Cache

- Calls to authd typically at mount time, or any NFS request after a server reboot.
- For each export, caches client address, flavor & permission.
- Cached entries flushed only if filesystem unshared or if share information changes or upon VM request.



NFS Client Authentication

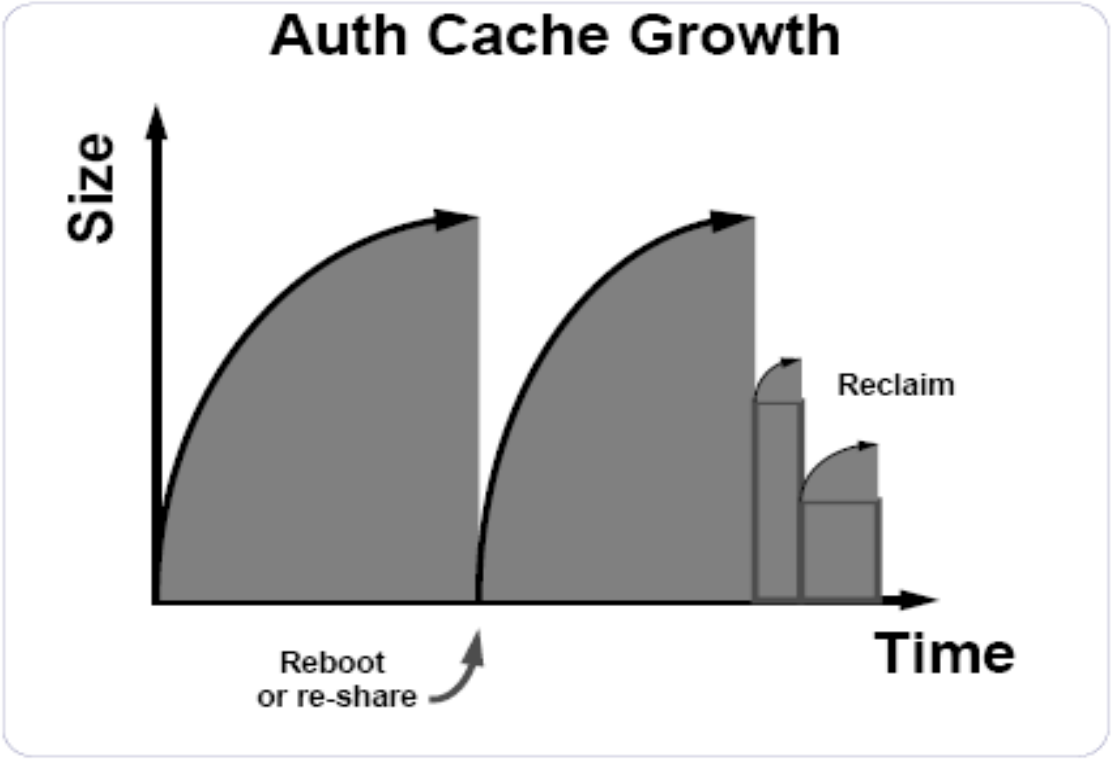
- Presented in Connectathon 1996
- Provide per client security
  - No spoofing
  - Allow multiple security flavors
- Netgroups
  - Not capped
  - Membership can change
- Use an Authentication Cache

The Authorization Cache

SunSoft
A Sun Microsystems Company

NFS Client Authentication

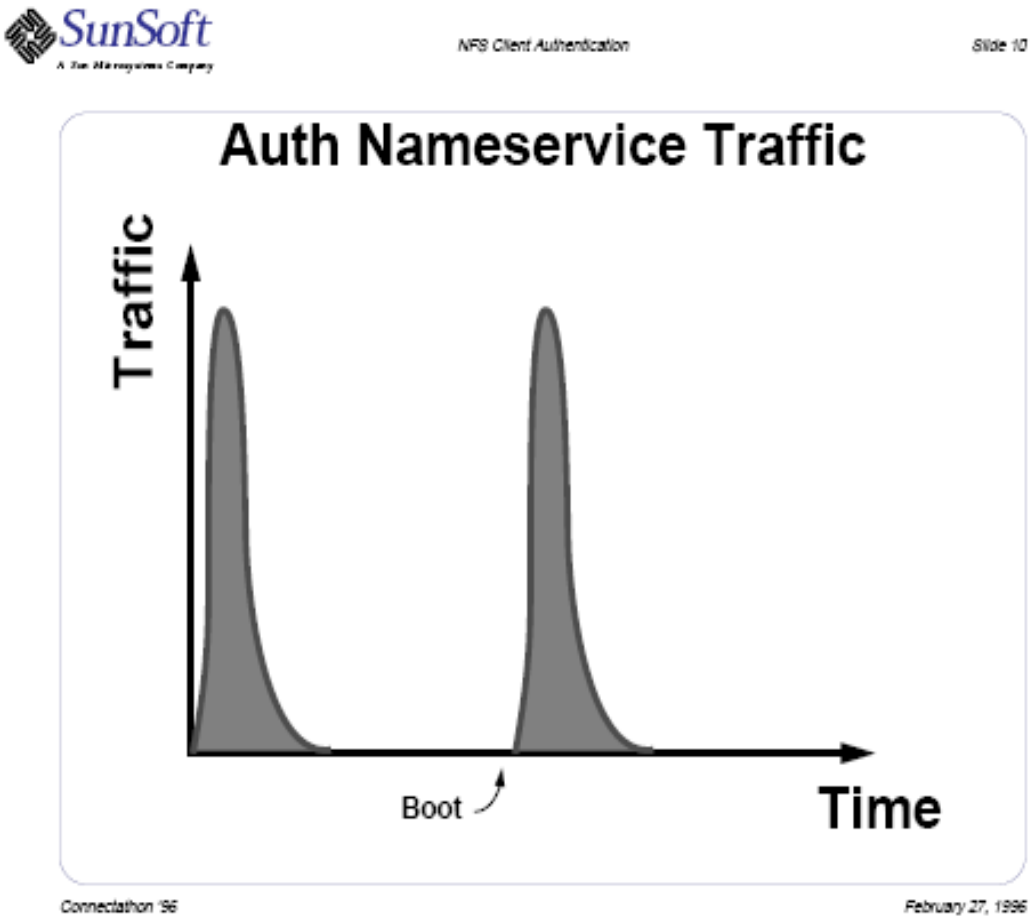
Slide 9



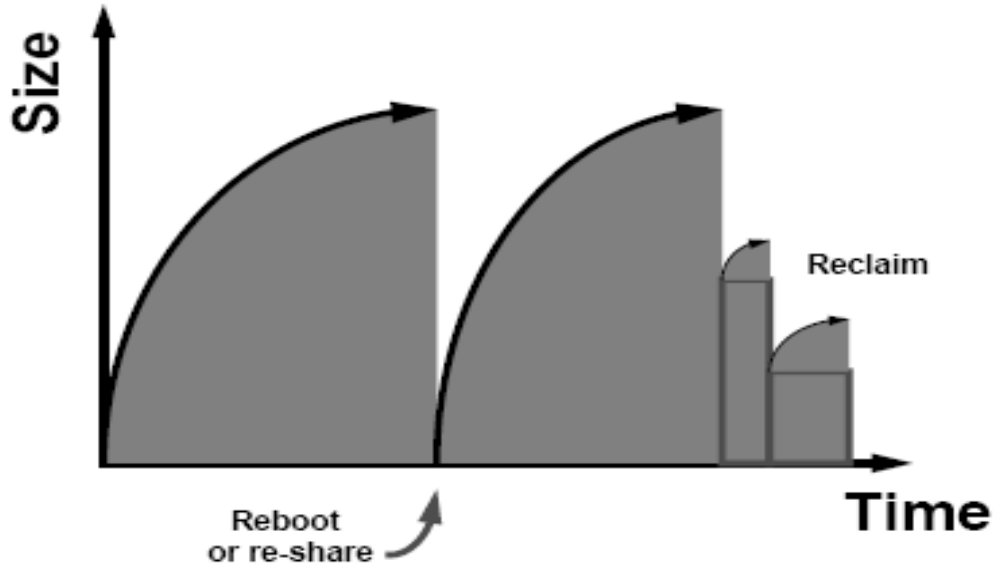
Connectathon '96

February 27, 1996

Traffic over time



Where's the Beef?



- How large is the Auth Cache?
  - I.e., is anything every paged out by VM?
- Are the curves that smooth?

How big is the chasm?

- What happens during a cache miss?
  - > Slow name servers?
  - > Dead name servers?
  - > Overloaded name servers?

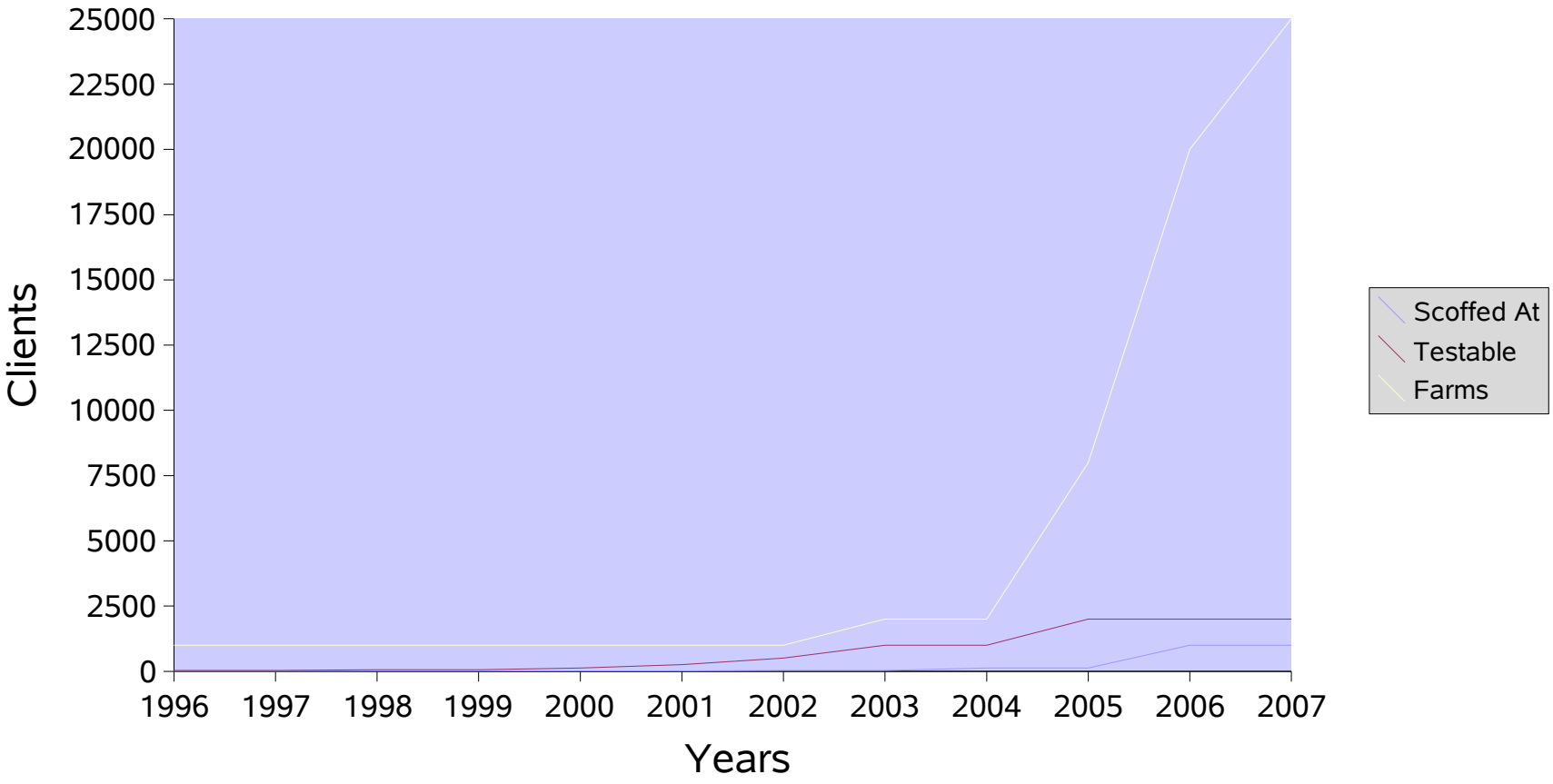
Factors Impacting Cache Misses

- Number of exports
  - Typical Unix server does not allow sub-mounts
- Number of clients
  - How many boxes could you afford?
- Automounters
  - More prevalent today

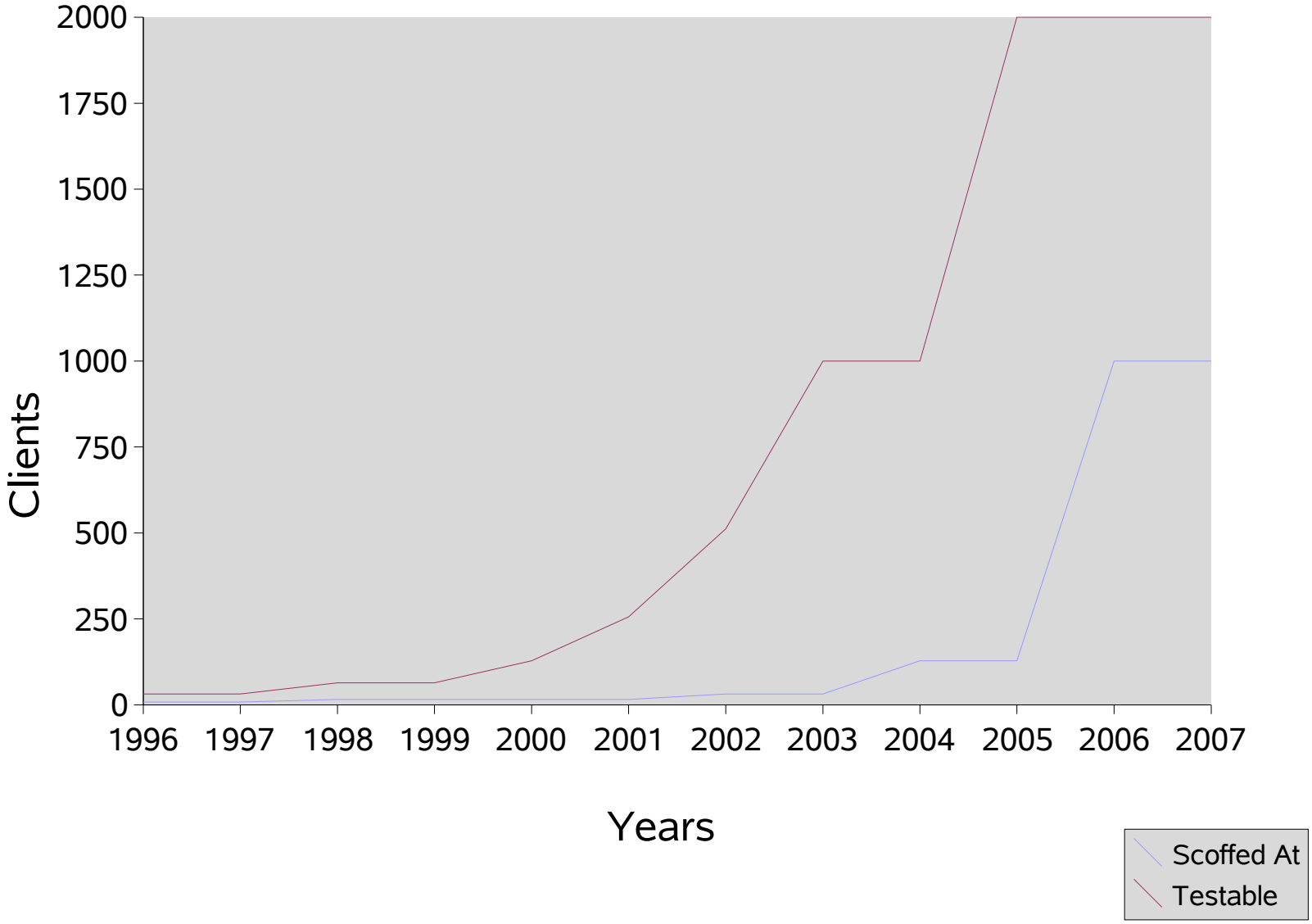
Number of Exports

- Data ONTAP allows 10,240 exports
  - *Does allow for sub-exports*
- Some rough Solaris data
  - Jurassic (4 weeks ago) 12 exports
  - Jurassic (1 week ago) 300 exports
  - Jurassic (today) 1300 exports
- ZFS testing is driving larger export sizes

Rise of the Machines



Focus on acceptable testing



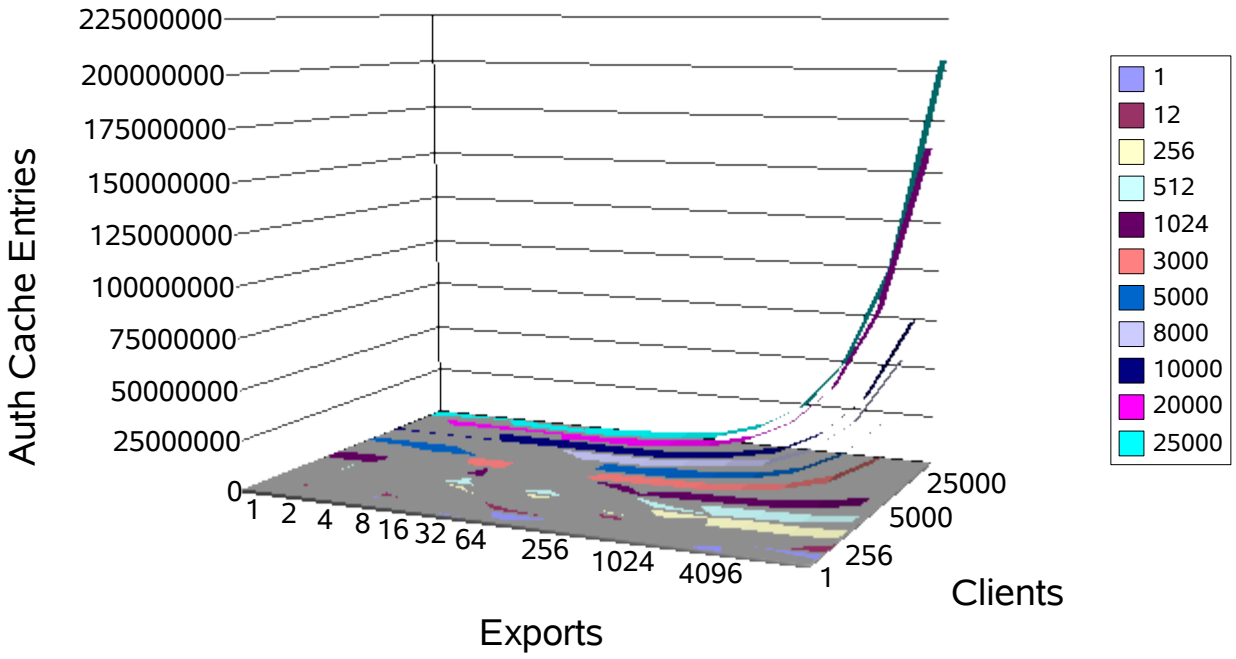
Automounter Hell

- Some automounters check all exports
- Some client farms all try to access at the same time
  - All get tools
  - All compute for a bit, then
    - Write results
    - Get new input
- Site admins manually try to stagger cron jobs

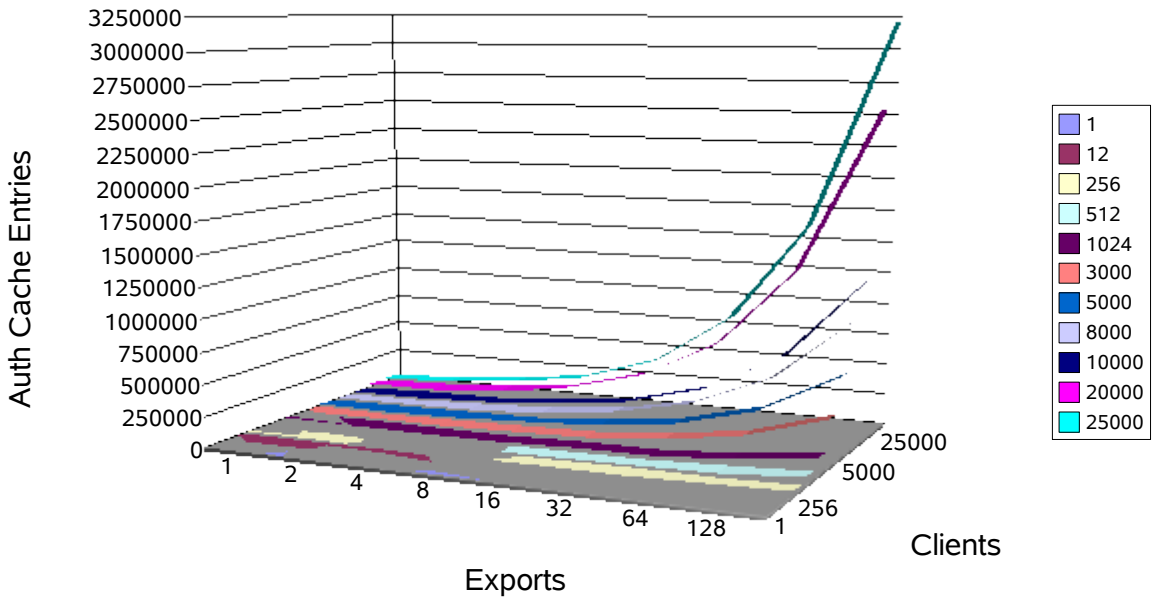
Deep in the Valley

- Many clients + many exports = name server storm
- Large Auth Cache size results in VM pressure reclaim
- Automounters synchronize storms
- Drop-offs occur more frequently

Explosion of Auth Cache Entries



Focus in on Old Realistic Sizes



Prior Scalable Solutions

- Limit size of netgroups
- Precompute all netgroups
- Not flexible with netgroups

Data ONTAP 6.4

- Translate all names into IP when loading exports
- I.e., precache
  - > What if name servers are not available?
- Avoids name servers, fast
- Only mechanism to refresh cache is to reload exports

Cray UNICOS

- Translated all names into IP when loading exports
- Use a Radix Tree to store information
- Reuses IP routing concepts
- Can compact neighbors

Best of Both Worlds

- Reuse name server information between exports
  - Netgroup entries
  - Reverse name lookups
- Make Auth Cache persistent across reboots
- Try to reuse data across changing export rules

Questions

Tom Haynes
tdh@sun.com