# Spinnaker Networks

## Scaling

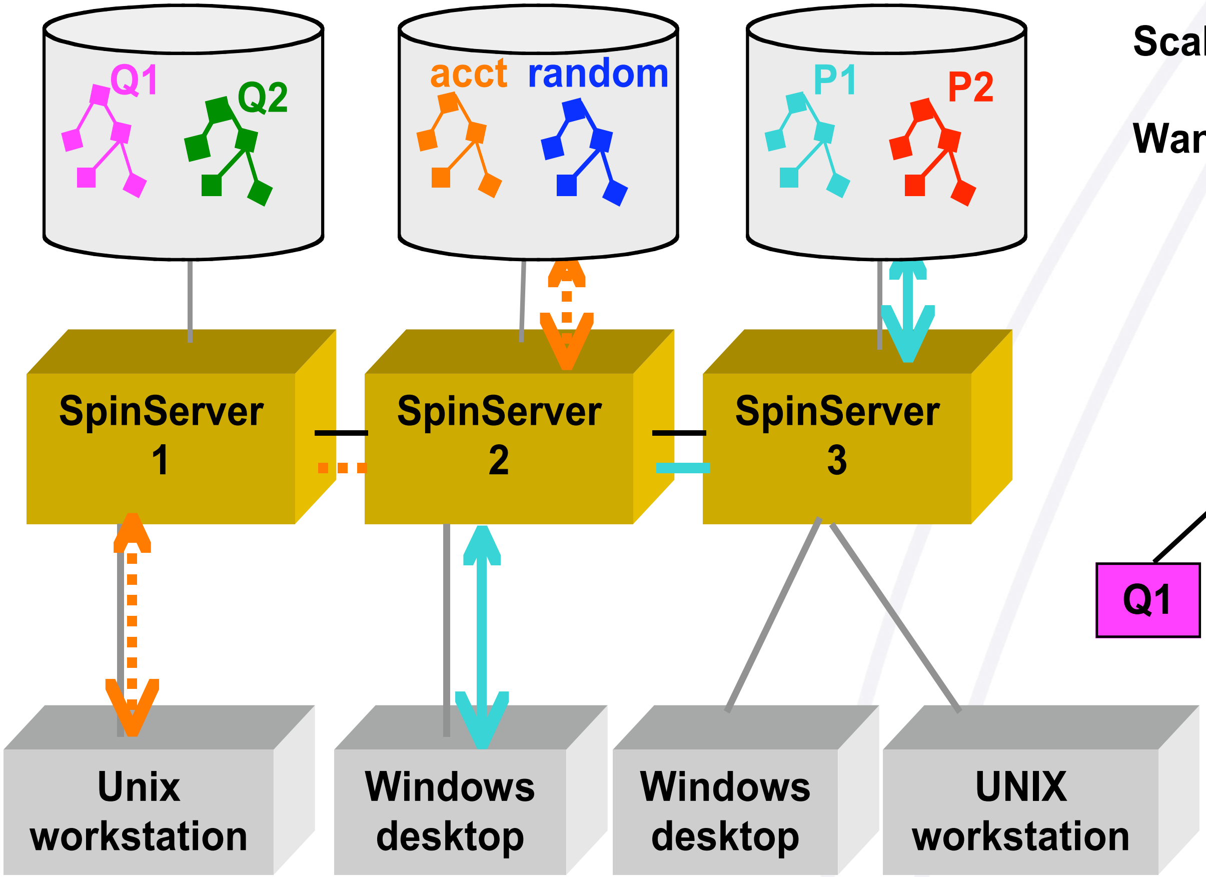
12/4/2004

SPINNAKER
NETWORKS

# Scaling Goals

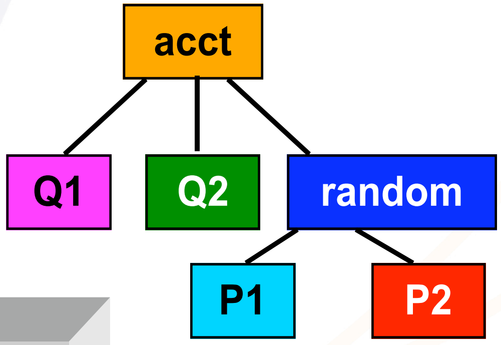
- Need to scale to 100s of servers
  - already have 1000s of clustered machines
  - may all be accessing one file
  - or small numbers of files
- Mix is mostly reads and writes
  - half the time mostly reads alone
  - far fewer need massive create/delete rate

# Clustering – Global Name Space



Scales with even load

Want to stripe VFSes



# Structure of File Systems

- Three types of file system data
  - Overall file system structure
    - dirs, symlinks, etc
    - not heavily modified in big clusters
  - File attributes
    - eg. file times, length, ACLs, versions
    - frequently updated (esp. times and length)
  - File data
    - actual data bytes
    - most heavily used parts of file system
- Need 3 types of locations as well

# Striping Architecture Observations

- **Striping works better than cluster locking**
  - **so stripe when possible, lock only when necessary**
  - **use delegations when possible for locked state**
- **File data stripes naturally**
- **File length & times locking can be optimized**
  - **length info needed for reads and writes**
  - **could use delegations revoked when file shrinks**
  - **often, heavily shared files don’t change size anyway**
- **Other attributes**
  - **used for getattr calls**

# Information Granularity

- Info stored per FS
  - dirs, symlinks, etc. (FS consistency)
- Info stored per file
  - file length
  - mtime, ctime, atime
    - atime already maintained loosely
- Info stored per stripe
  - version information (don’t make global)
  - data delegations
    - whole strip delegations
    - per-strip range delegations (if supported)
  - read and write data

# NFS Problems

- Times as version numbers
  - uses ctime
    - semantically ambiguous (utimes problem)?
    - per-file instead of per-strip (too global)
- No concept of file system structure
  - separation of meta data, file attrs, file data
  - no file striping geometry descriptions
  - no use made of length’s semantics
- No fine grained delegations

# Suggested Fixes for NFS

- File systems
  - meta-data server
- Files
  - file attribute server
  - data geometry for striping (RAID too?)
- Real version numbers
  - per strip server
  - not time stamps
- Delegations
  - per strip
  - byte range?

SPINNAKER
NETWORKS