

Scaling Likewise-CIFS Beyond 50k Concurrent Connections (on a single node)

Gerald Carter

<gcarter@likewise.com>

Director of Engineering

Likewise Software

- ❑ Introduction to the Likewise Open project and to the Likewise-CIFS File Server
- ❑ Explanation of connection load stress test tool
- ❑ Likewise-CIFS Architecture
 - ❑ Thread Pools and Tasks
 - ❑ Drivers – SRV, PVFS
- ❑ Evaluation of Results
- ❑ Plans for future improvement

**“No one cares about performance
until it’s not there.”**

**“Workloads, much like people,
are rarely objective.”**

“Perfectionist”

**Favorite T-Shirt Slogan
@ThinkGeek.com**

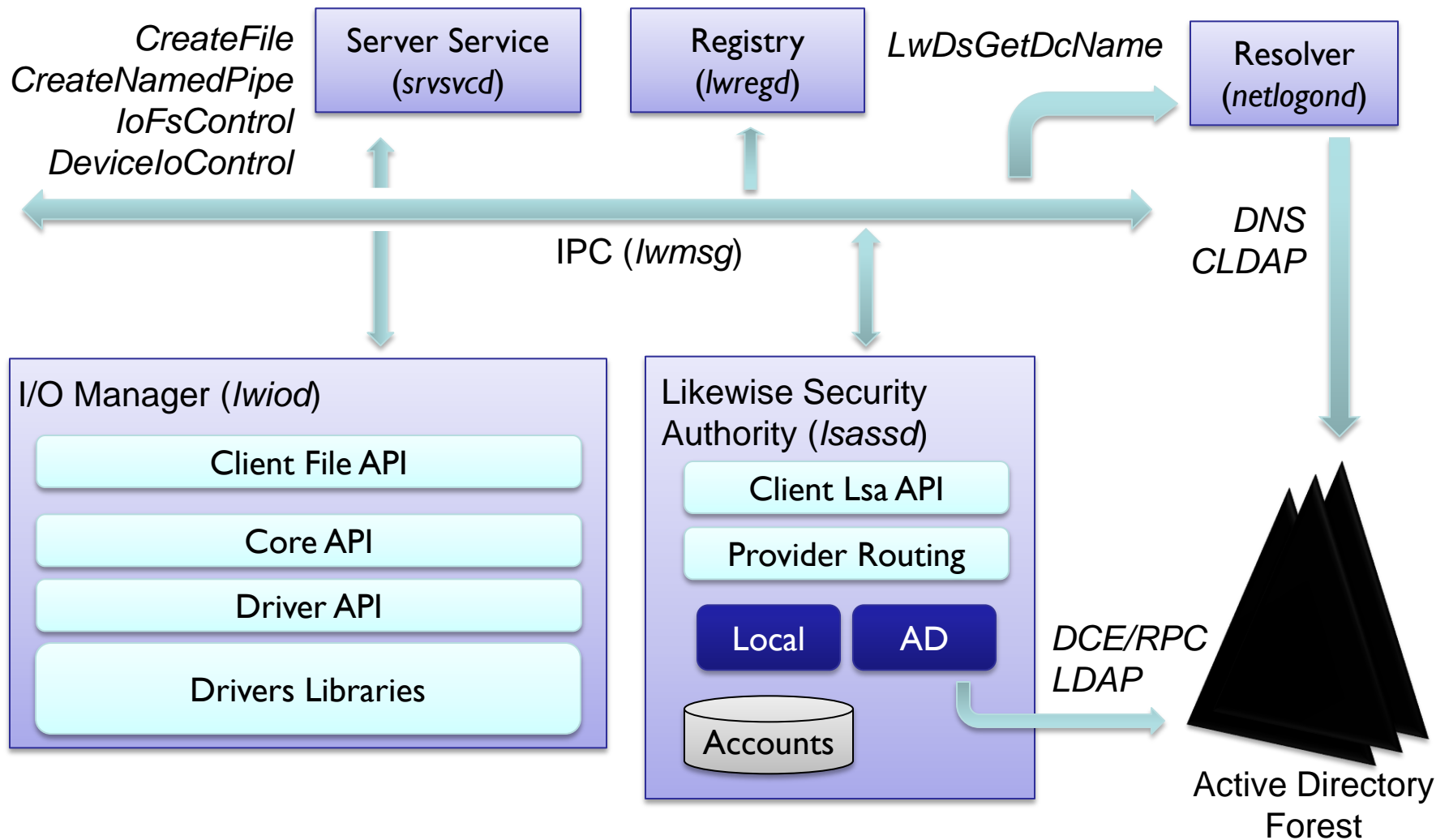
**Just behind “5 out of 4 people have trouble
with fractions.”**

Likewise Open – Background

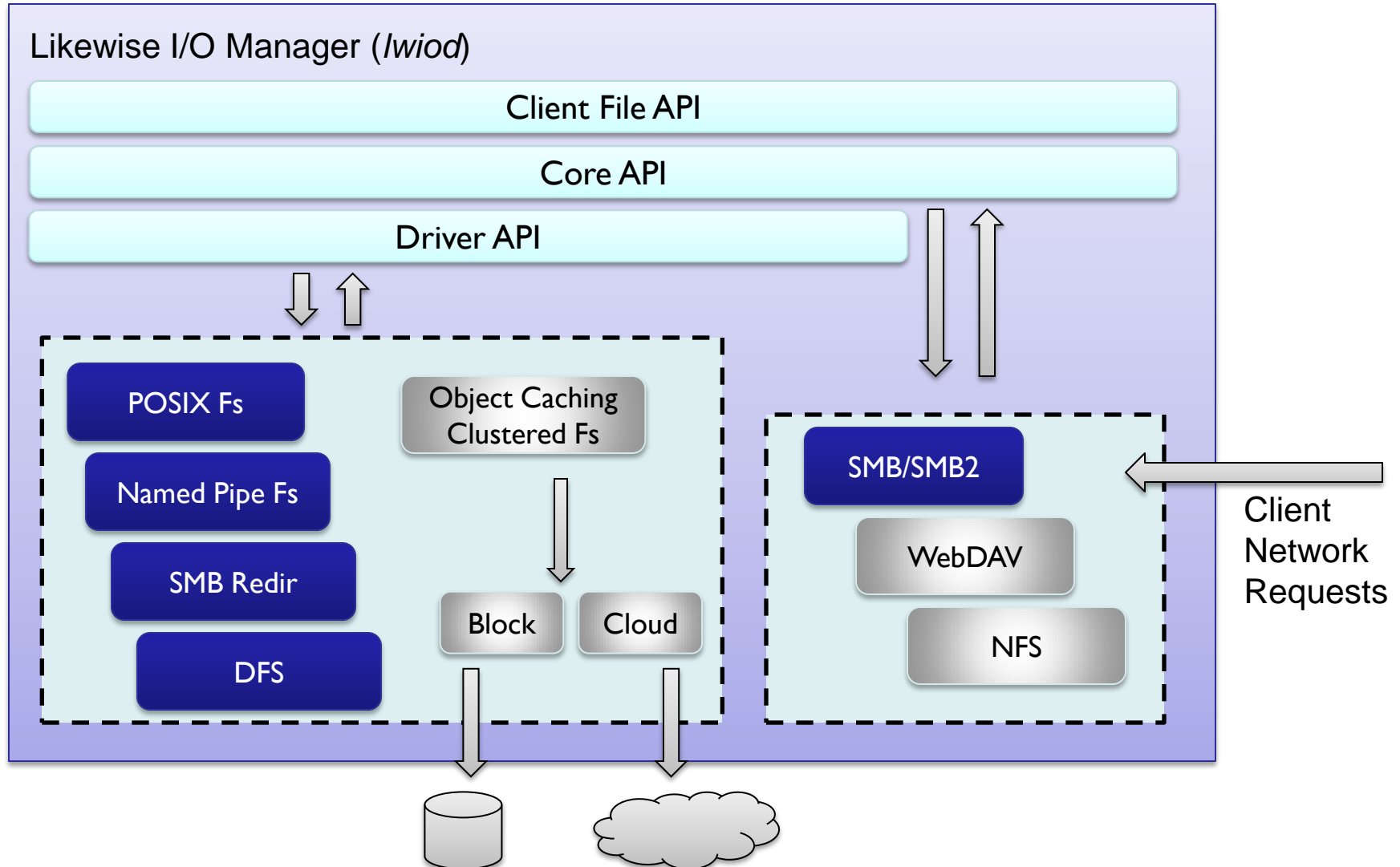
<http://www.likewiseopen.org/>

- ❑ Likewise Open Project is the umbrella project sponsored by Likewise Software designed to provide an interoperability platform for non-Microsoft clients and servers in Microsoft OS dominated networks.
- ❑ Likewise Open (product) refers to the open source authentication & Active Directory integration suite
- ❑ Likewise-CIFS is the file server software stack
 - ❑ Currently includes support for SMB and SMB2
- ❑ License
 - ❑ Choice: Commercial or GPLv2+/LGPLv2.1+
 - ❑ Single code base

Likewise-CIFS Core Components

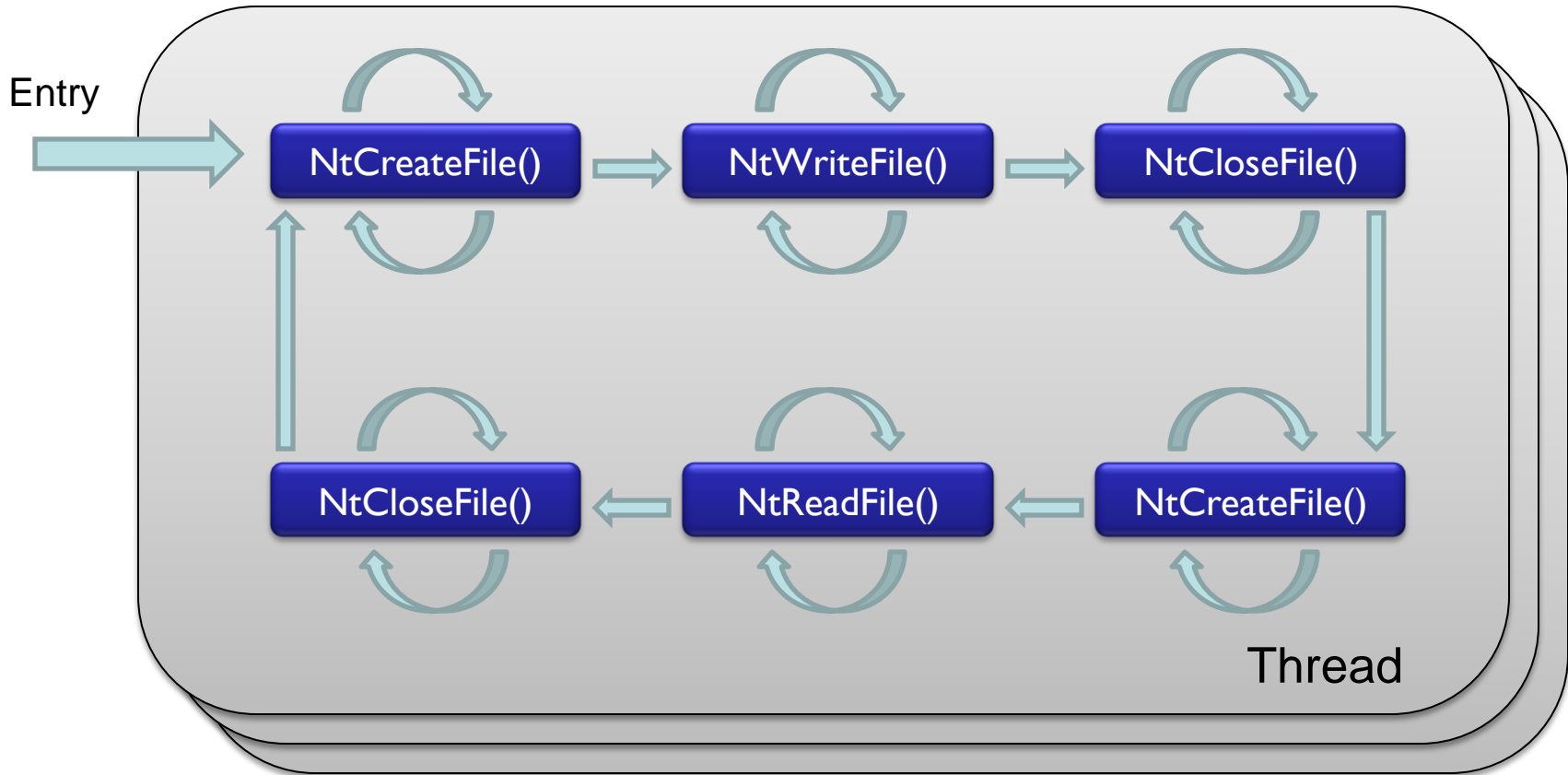


Likewise I/O Manager



- ❑ *test_load*
 - ❑ Included in “likewise-open/lwio/tests/test_load/”
 - ❑ Simple testing utility for spawning N number of connections per thread for X threads
 - ❑ Uses the Likewise I/O Mgr SMB redirector (RDR) for outgoing connections
- ❑ Focus on number of concurrent connections, session, trees, & files
 - ❑ Not a throughput test

Connection I/O Pattern



Usage: ./test_load

- ❑ `./test_load [options] <server> <share>`
- ❑ Options
 - ❑ `--iterations <int>`
 - ❑ `--threads <int>`
 - ❑ `--connections <int>`
 - ❑ `--{domain,user,password} <string>`
 - ❑ Looks for a local Krb5 ticket cache if no credentials supplied on the command line
 - ❑ `--continue-on-error`

Client Run

```
$ test_load \  
  --iterations 50 --threads 100 --connections 50 \  
  --user testload --domain AD --password testload \  
  ash.ad.plainjoe.org torture  
...  
[92] Starting iteration 1 of 10...  
[92] Opening 50 files for writing...  
[72] Starting iteration 1 of 10...  
...  
[51] writing to files...  
[89] Reopening files for reading...  
[7] writing to files...  
[59] writing to files...  
[21] writing to files...  
...  
[9] Closing files...  
[92] Reopening files for reading...  
[83] Starting iteration 6 of 10...  
...
```

Server Statistics

```
## Stats from SRV
$ lwio-cli --get-stats
Server statistics [level 0]:

Connections   [Current: 831] [Maximum: 831]
Sessions      [Current: 826] [Maximum: 826]
Tree connects [Current: 823] [Maximum: 823]
Files:        [Current: 823] [Maximum: 823]
```

```
## Stats from PVFS
$ test_pvfs --ls-open-files 100
Handles Delete Filename
-----
0         No      /
1         No      /data/torture/test-load-cattibrie-0.txt
1         No      /data/torture/test-load-cattibrie-900.txt
9         No      /data/torture
...

```

Likewise-CIFS Results

- ❑ Server - 4xAMD Opteron, 8Gb RAM
- ❑ 3 Clients
 - ❑ 20k, 20k, 10k
- ❑ Threads (4/8/4)
 - ❑ 4 socket tasks
 - ❑ 8 SRV workers
 - ❑ 4 PVFS workers
- ❑ *lwiod* initial RSS 10Mb
 - ❑ ~75Kb/connection

```
File Edit View Terminal Help
top - 17:15:10 up 33 days, 36 min, 5 users, load average: 7.84, 8.06, 8.11
Tasks: 7 total, 0 running, 7 sleeping, 0 stopped, 0 zombie
Cpu(s): 18.2%us, 12.5%sy, 0.0%ni, 63.1%id, 5.4%wa, 0.1%hi, 0.7%si, 0.0%st
Mem: 7943284k total, 7909016k used, 34268k free, 319868k buffers
Swap: 7815580k total, 37240k used, 7778340k free, 2920624k cached

  PID USER      PR  NI  VIRT  RES  SHR  S  %CPU  %MEM    TIME+  COMMAND
 8779 root        20   0 4310m 3.7g 5676 S   59  49.3 283:21.03 lwiod
 8773 root        20   0 296m  5144 3660 S   33   0.1 127:43.73 lwregd
 8804 root        20   0 592m   9m 7124 S   26   0.1 90:56.38 lsassd
 8798 root        20   0 297m 4868 3404 S    0   0.1 14:32.74 netlogond
 8670 root        20   0 179m 2264 1496 S    0   0.0  0:00.13 lwsmd
 8761 root        20   0 142m 3892 2556 S    0   0.0  0:00.06 dcerpcd
 8851 root        20   0 299m 5196 3440 S    0   0.1  0:00.03 srvsvcd

 1 Processes

Mon Aug 9 17:14:52 CDT 2010
Server statistics [level 0]:

Connections [Current: 50997] [Maximum: 50997]
Sessions [Current: 50027] [Maximum: 50027]
Tree connects [Current: 50022] [Maximum: 50022]
Files: [Current: 49802] [Maximum: 49803]

Mon Aug 9 17:15:07 CDT 2010
Server statistics [level 0]:

Connections [Current: 50997] [Maximum: 50997]
Sessions [Current: 50039] [Maximum: 50043]
Tree connects [Current: 50035] [Maximum: 50041]
Files: [Current: 49838] [Maximum: 49842]

 0 Statistics
```

Connection Architecture

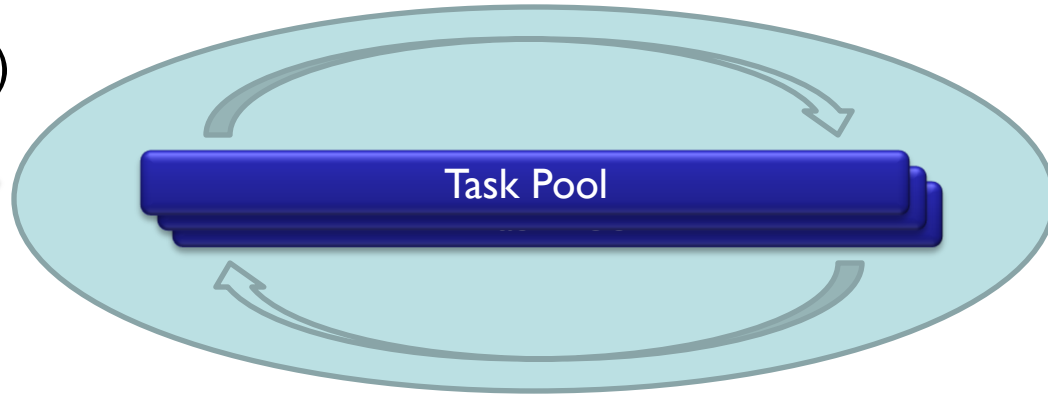
- ❑ SRV Driver
 - ❑ Thread Pools
 - ❑ Socket management
 - ❑ Connections, Session, Trees, and Files
- ❑ PVFS Driver
 - ❑ Open Handles vs. File System Objects

- ❑ Created by LwRtlCreateThreadPool()
 - ❑ Task Threads
 - ❑ Must be associated with an FD
 - ❑ Triggered by FD events: select(), epoll(), & kqueue()
 - ❑ Must not block
 - ❑ Worker Threads
 - ❑ Utilize a work queue FIFO model
 - ❑ Allowed to block

LW_THREAD_POOL

Tasks

LwRtlCreateTask()



Assigns Task Callback
to least loaded Task
thread

Task FDs



FD Events



select(), epoll(), kqueue()

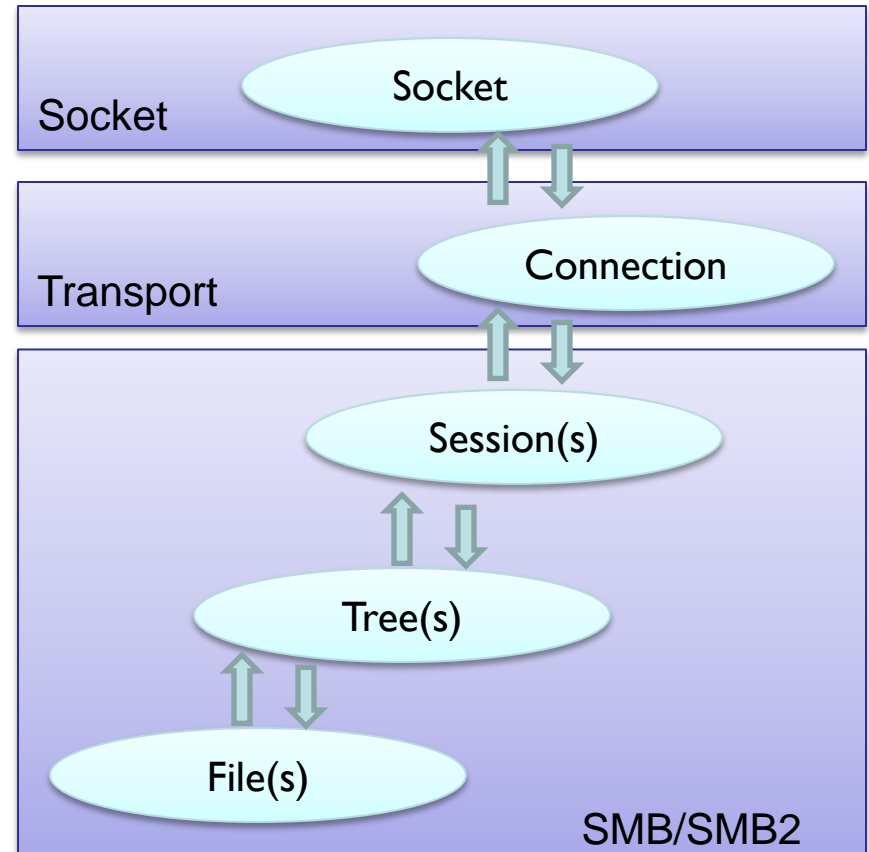
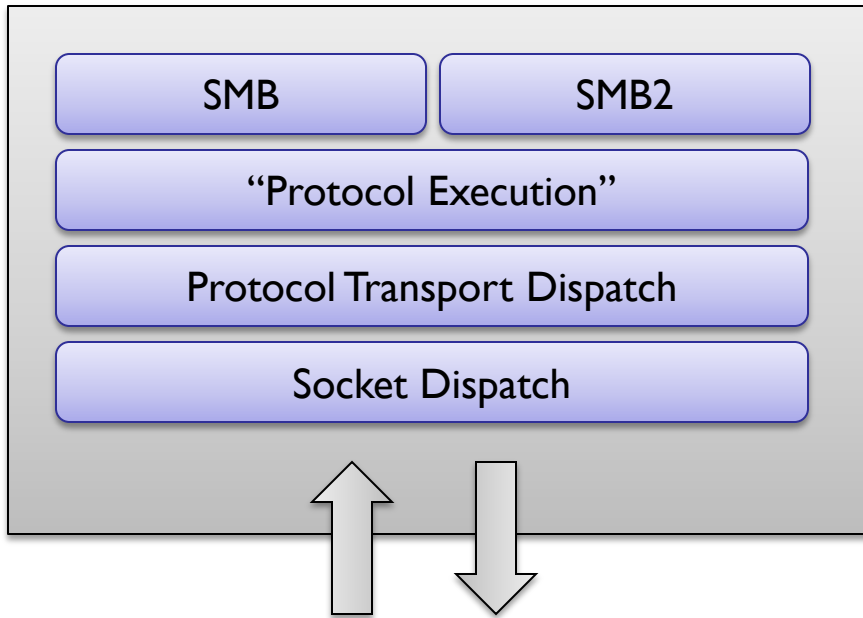
❑ Listener Task

- ❑ Listens on tcp/445 (Task for IPv4 & IPv6)
- ❑ Creates “PSRV_SOCKET pSocket” on accept()
- ❑ Assigns pSocket to a new Task

❑ Socket Tasks

- ❑ Creates new LWIO_SRV_CONNECTION
- ❑ Reads and Writes (including Zero Copy)
- ❑ Reads invoke Transport Protocol Dispatch layer

Connection & Session Management



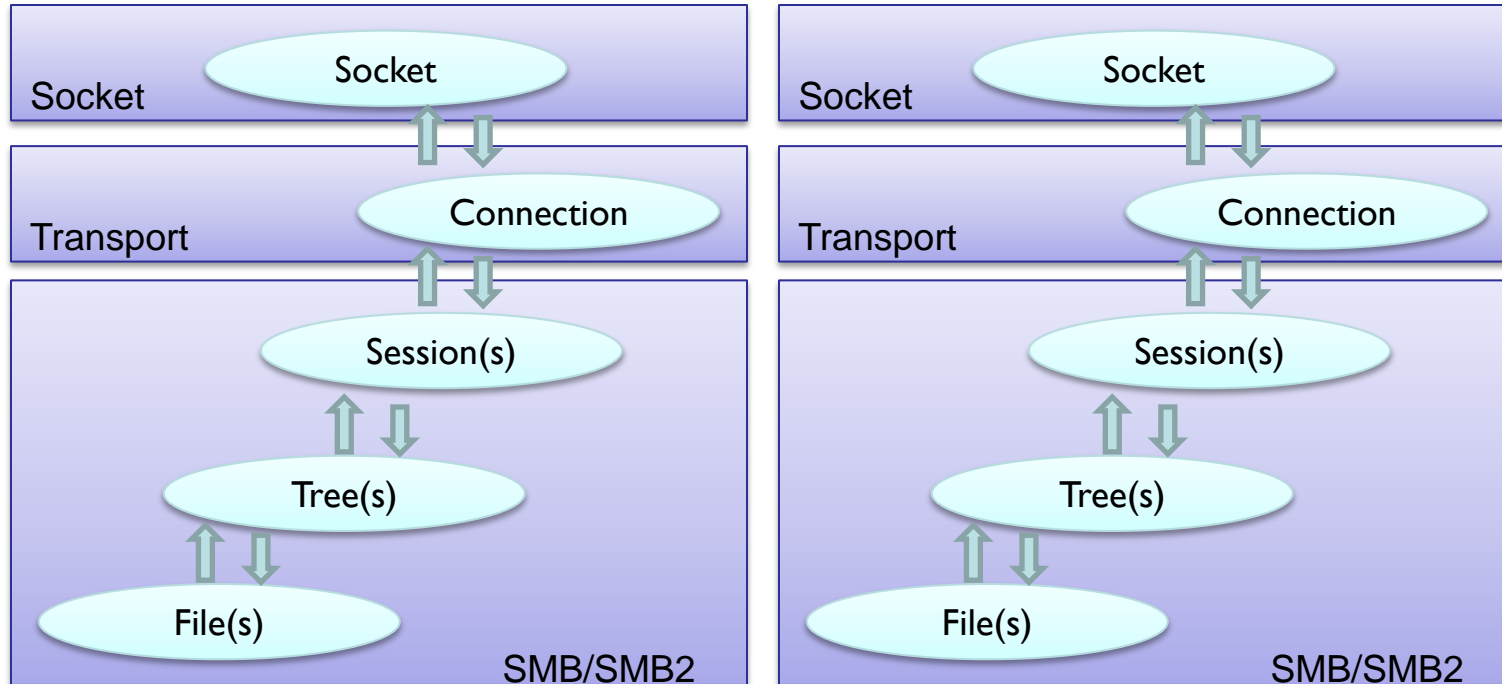
Socket and Transport Dispatch

```
typedef struct _SRV_CONNECTION_SOCKET_DISPATCH
{
    PFN_SRV_SOCKET_FREE                pfnFree;
    PFN_SRV_SOCKET_DISCONNECT           pfnDisconnect;
    PFN_SRV_SOCKET_GET_ADDRESS_BYTES   pfnGetAddressBytes;
} SRV_CONNECTION_SOCKET_DISPATCH,
*PSRV_CONNECTION_SOCKET_DISPATCH;
```

```
typedef struct _SRV_TRANSPORT_PROTOCOL_DISPATCH
{
    PFN_SRV_TRANSPORT_CONNECTION_NEW   pfnConnectionNew;
    PFN_SRV_TRANSPORT_CONNECTION_DATA pfnConnectionData;
    PFN_SRV_TRANSPORT_CONNECTION_DONE  pfnConnectionDone;
    PFN_SRV_TRANSPORT_SEND_PREPARE     pfnSendPrepare;
    PFN_SRV_TRANSPORT_SEND_DONE        pfnSendDone;
} SRV_TRANSPORT_PROTOCOL_DISPATCH,
*PSRV_TRANSPORT_PROTOCOL_DISPATCH;
```

- ❑ Test characteristics
 - ❑ Single session per connection
 - ❑ Single tree connect per session
 - ❑ Single file open per tree connect
- ❑ Conclusion:
 - ❑ SRV – Stresses socket and transport layers but not SMB/SMB2 session/tree/file data structures
 - ❑ PVFS – All unique file/directory opens create an entry in the File Control Block (FCB) table
 - ❑ Table lock contention, search time, etc...

Test Characteristics



Building Likewise-CIFS

- ❑ Simple build system for Linux & FreeBSD
- ❑ Step 1: Download the source code
 - ❑ `$ git clone git://git.likewiseopen.org/likewise-open`
- ❑ Step 2: Build the likewise-open components
 - ❑ `$ build/mkcomp [--noincremental] [--debug] all`
 - ❑ Installs all pieces to “staging/install-root/”
- ❑ Step 3: Generate RPMs/DEBs (Linux only)
 - ❑ `$ build/mkpkg [--debug] all`
 - ❑ Creates packages in “staging/packages/”

- ❑ Wish List for *test_load*
 - ❑ Better interface for *test_load* (e.g. ncurses frontend)
 - ❑ Less serialization of I/O
 - ❑ Random I/O Patterns
 - ❑ Track ops/second for report generation
- ❑ Test is freely available under the GPL
 - ❑ Upstream patch submission requires a signed contributor agreement

Questions?

Gerald Carter

Director of Engineering

Likewise Software

gcarter@likewise.com

<http://www.likewise.com/>

<git://git.likewiseopen.org/likewise-open>