Understand opportunistic locking (oplock) and related issues

Royce Lu
fruitfoxlu@gmail.com

Who need to understand driver

- NT file system driver developer
- filter driver developer
- Any kernel driver that will try to read or write file .

Relaed issues

- Handle oplock improperly will cause hang

- A mechanism to avoid sharing violations

What is oplock?

- Since NT 3.1

- It is not a lock for synchronization.

What is oplock

- It was design for remote file system performance.
- If we want to reduce network use, enhance performance on remote file access. What will we do? Cache data on local machine

Oplock is for reducing network usage.

- If only client 1 want to r/w remote file A, server will grante a <u>Level 1 oplock</u>.

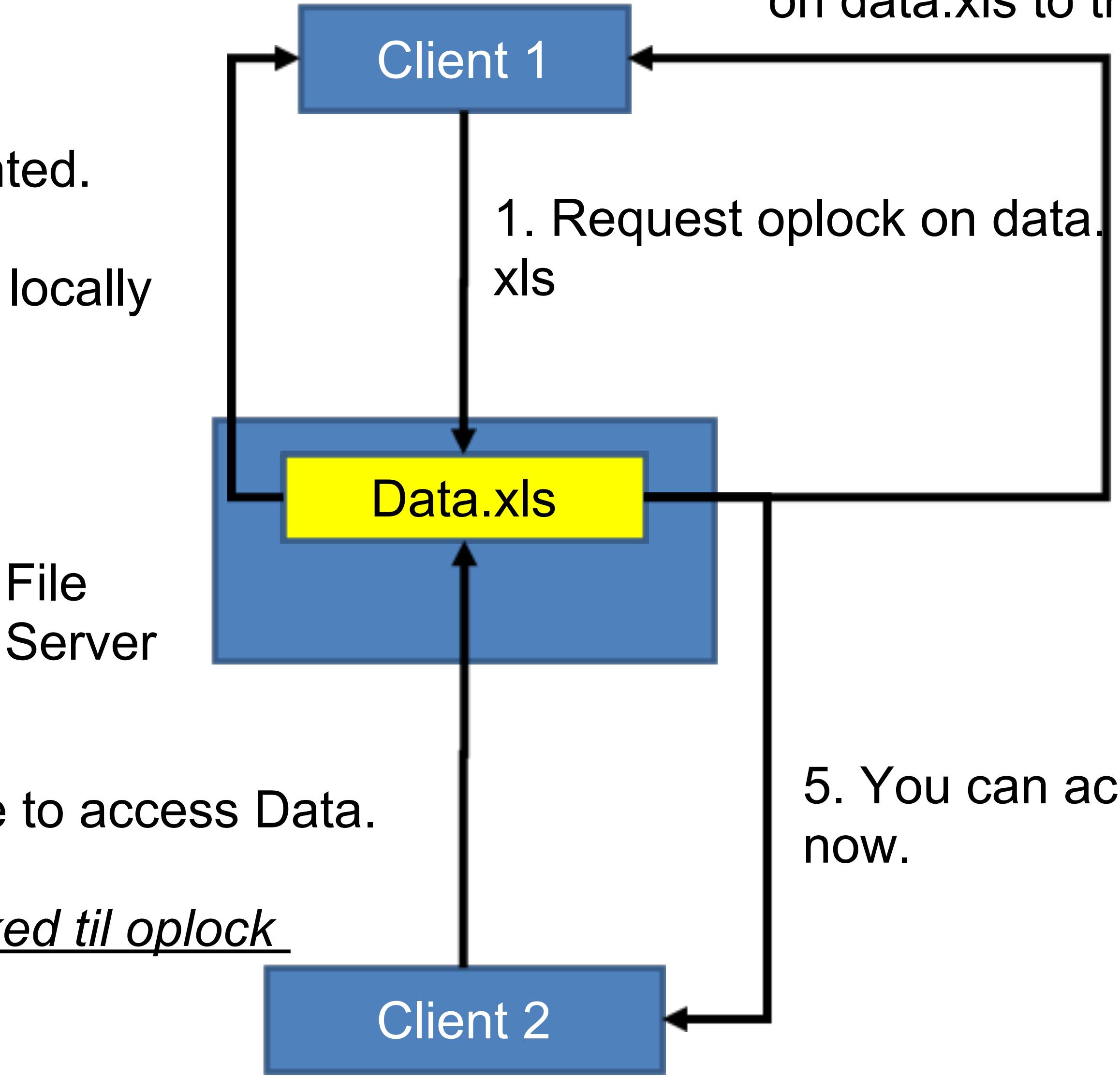
- Level 1 oplock means client 1 can cache r/w from the file A locally, flush the changes back when file closed or oplock is broke.

When will oplock be broke?

- Now client 2 want to r/w file A before client 1 close file handle.

- Lv 1 oplock will be break to no oplock. Because under this scenario performing cache will cause the file content inconsistent.

- Client 1 has to flush all the changes back to the server, then acknowledge the break.



1. Request oplock on data.
xls

2.Oplock granted.
Cache
your changes locally

3. I would like to access Data.
xls
*(Will be blocked til oplock
break)*

4.Oplock break, return your
changes
on data.xls to the server.

5. You can access
now.

Optimize it

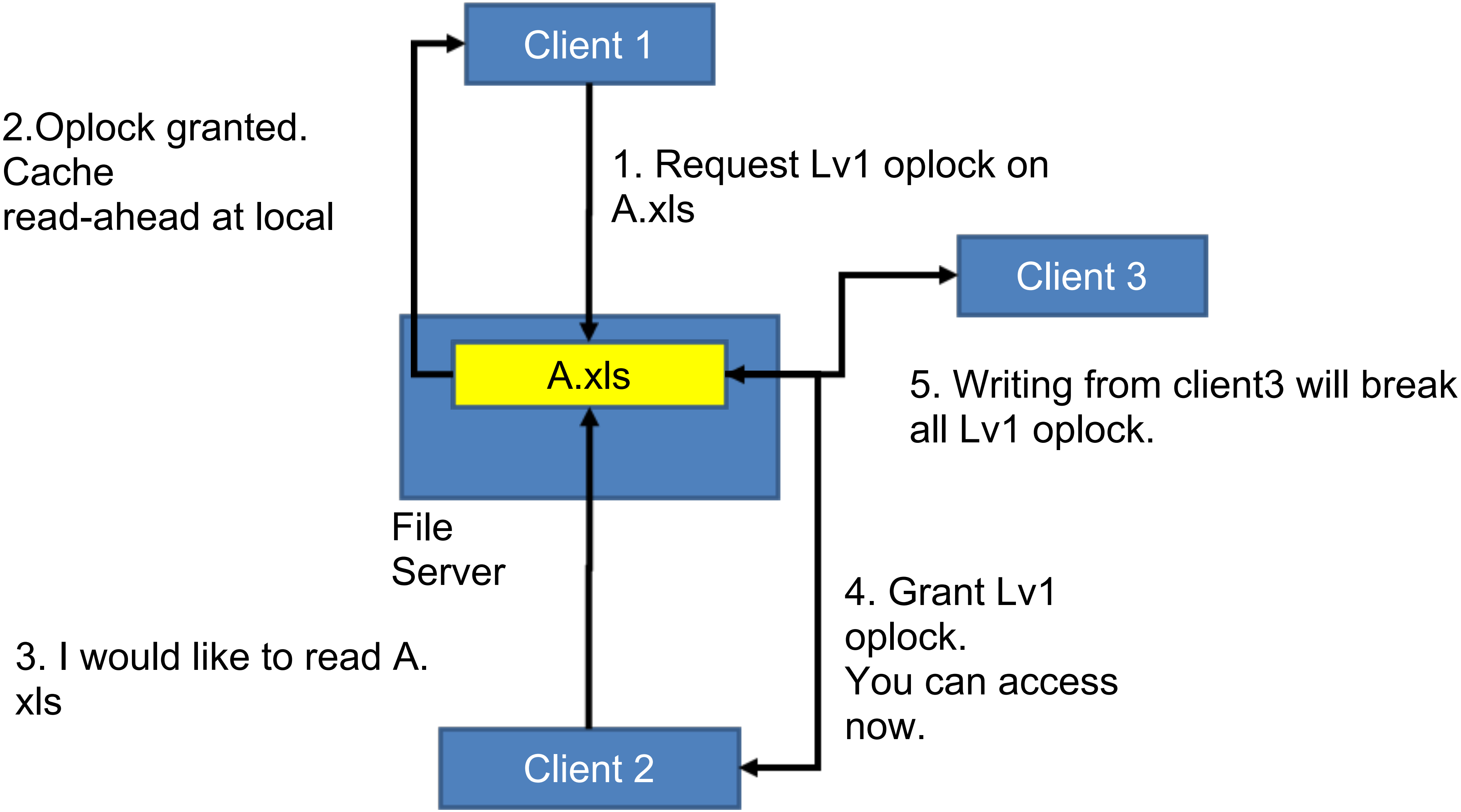
- If both client 1 and 2 just want to read file A, can they share an oplock?

- In this case, both of them will be granted a Level 2 oplock. Which means they can read-ahead and cache the result at local.

Level 2 oplock

- If there is any IRP_MJ_WRITE to file A before client 1&2 close the file, level 2 oplock will be broke to no oplock.

- Without oplock, Client 1&2 have to read file A remotely .



Client 1

2.Oplock granted.
Cache
read-ahead at local

1. Request Lv1 oplock on
A.xls

Client 3

A.xls

5. Writing from client3 will break
all Lv1 oplock.

File
Server

4. Grant Lv1
oplock.
You can access
now.

3. I would like to read A.
xls

Client 2

Level 1 break to Level 2

- If client 1 has a Lv1 oplock on file A, and client 2 just want to read file A.

- Server will break client 1’s oplock to Lv 2. After acknowledge the break, client 2 will be granted a Lv2.

- If there is any IRP_MJ_WRITE on file A, all Lv2 will be broke to no oplock.

Breaking Oplock

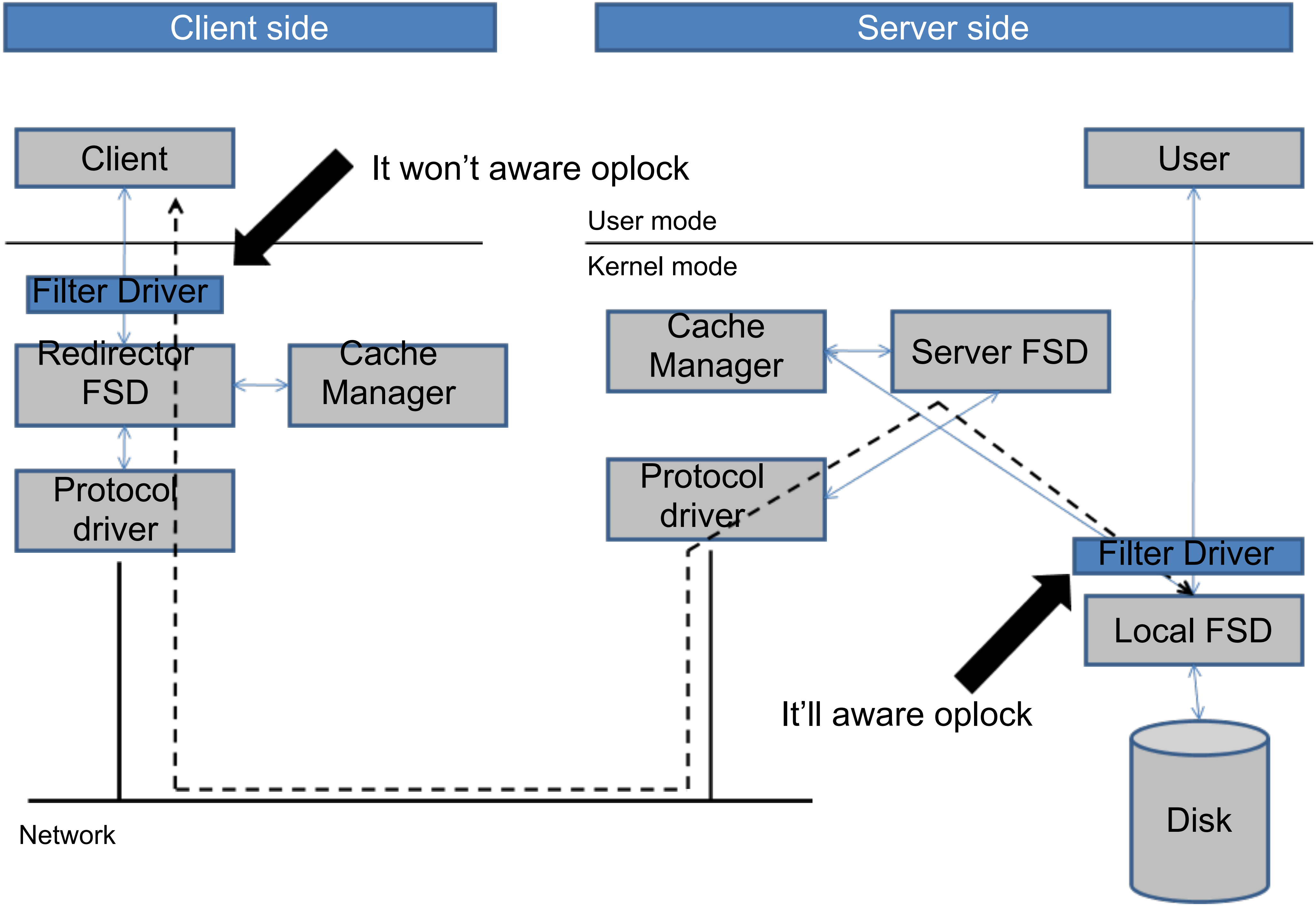
- Breaking oplock can happen on following IRP
  - IRP_MJ_CREATE
  - IRP_MJ_READ
  - IRP_MJ_WRITE
  - IRP_MJ_CLEANUP
  - IRP_MJ_LOCK_CONTROL
  - IRP_MJ_SET_INFORMATION
  - IRP_MJ_FILE_SYSTEM_CONTROL
- Detail is for file system developer
- All states are documented
  - http://msdn.microsoft.com/en-us/library/dd445269.aspx

Basic concept

- Oplocks is stream handle based

- For file systems that do not support ADS, ex: FAT, file handle = stream handle.

- The oplock breaks even if it is the same process or thread performing the incompatible operation.



Client side

Server side

Client

Filter Driver

Redirector FSD

Cache Manager

Protocol driver

Network

It won’t aware oplock

User

User mode

Kernel mode

Cache Manager

Server FSD

Protocol driver

Filter Driver

Local FSD

Disk

It’ll aware oplock

We observe

1. Client and filter driver at client side will not aware oplock
2. Server FSD & Redirector FSD will use oplock as their cache coherency protocol
3. Because user at server side will access the file that Remote FSD is using, Local FSD also need to support oplock protocol.

FILE_COMPLETE_IF_OPLOCKED

- If file is oplocked, any action that will result in oplock break will be blocked until acknowledge from the owner.
- If the client don’t want to be blocked, maybe will deadlock if blocked, they can specify FILE_COMPLETE_IF_OPLOCKED in the CreateOptions for ZwCreateFile / IRP_MJ_CREATE

FILE_COMPLETE_IF_OPLOCKED

- If oplocked will be broke, the thread will not be blocked. FSD will return STATUS_OPLOCK_BREAK_IN_PROGRESS

- It is a success status code.
- NT_SUCCESS(status) == TRUE

FILE_COMPLETE_IF_OPLOCKED

- If we see FILE_COMPLETE_IF_OPLOCKED in pre-create, any action on the target file that will be blocked until acknowledge of oplock break is unexpected for the caller

- Potential deadlock.

FILE_COMPLETE_IF_OPLOCKED

- If we see FILE_COMPLETE_IF_OPLOCKED in post-create and status code is STATUS_OPLOCK_BREAK_IN_PROGRESS, means target file is oplocked.
- Any action that will be blocked until acknowledge, is unexpected for the caller.

Oplock type

- Lv 1 : Exclusive own, only one handle can have it. Can perform r/w at local.
- Lv 2 : Shared. Can perform read data at local

Oplock type

- Batch : Exclusive. Keep stream open on the server. Support scenario that open/close file repeatedly. Client also can do read /write caching.

- Filter : Exclusive. Act like Lv1 lock but only break when sharing violation happened. For filter to “get out”.

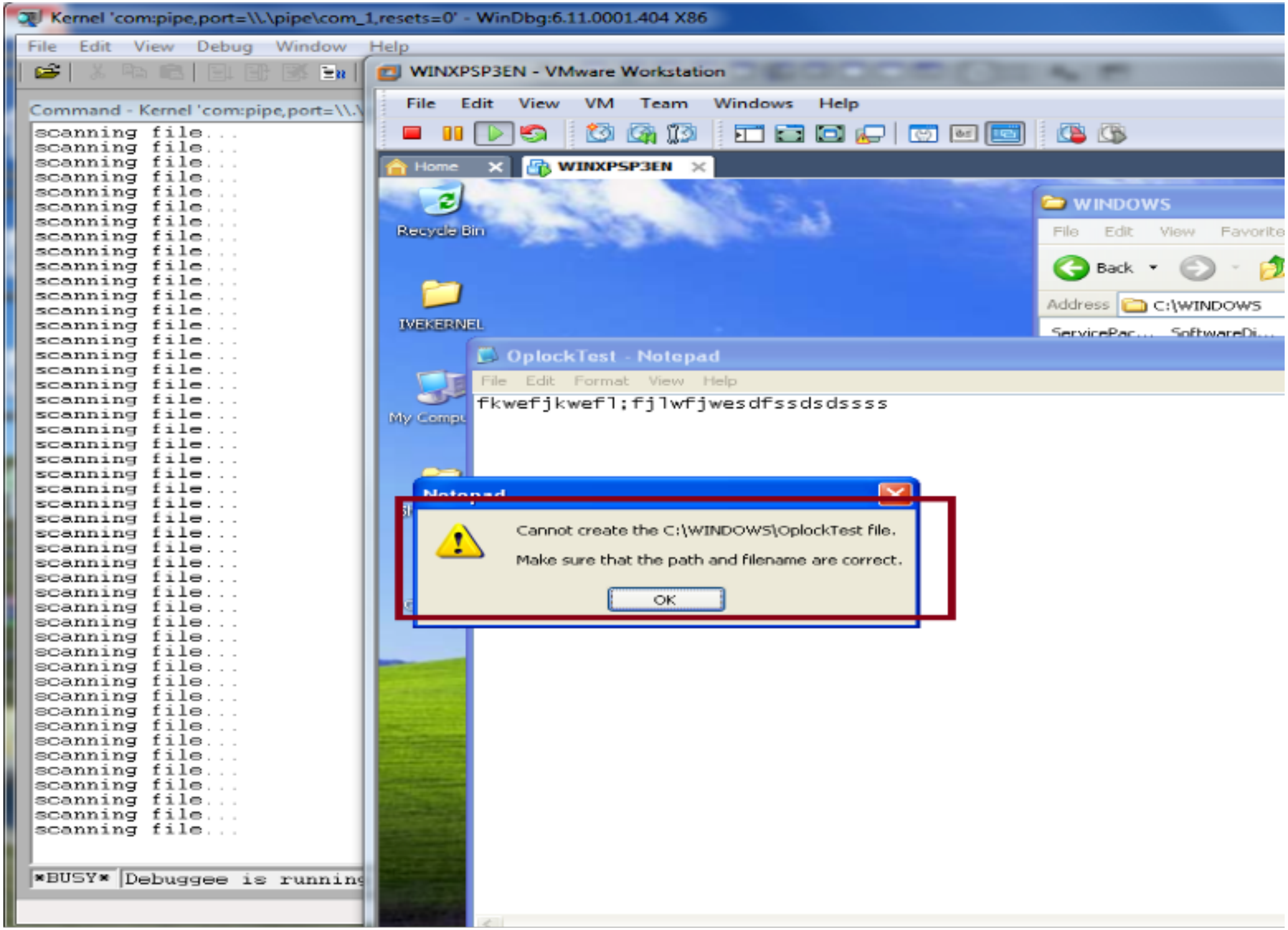
Filter Oplock

- Utilize oplock protocol to notify filter sharing violation.
- Filter acknowledge the break, and get out. User will be unaware to sharing violation.

How to require filter oplock

1. ZwCreateFile
2. ZwFsControlFile
    1. FSCTL_REQUEST_FILTER_OPLOCK
    2. Provide an event handle
3. Event will be signaled if the sharing violation happened.

No filter oplock : sharing violation



Kernel 'com:pipe,port=\\.\pipe\com_1,resets=0' - WinDbg:6.11.0001.404 X86
File Edit View Debug Window Help
Command - Kernel 'com:pipe,port=\\.\
WINXPSP3EN - VMware Workstation
File Edit View VM Team Windows Help
Home
WINXPSP3EN
Recycle Bin
IVEKERNEL
My Compu
WINDOWS
File Edit View Favorite
Back
Address
C:\WINDOWS
ServicePac... SoftwareDi...
OplockTest - Notepad
File Edit Format View Help
Notepad
OK

```
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
```

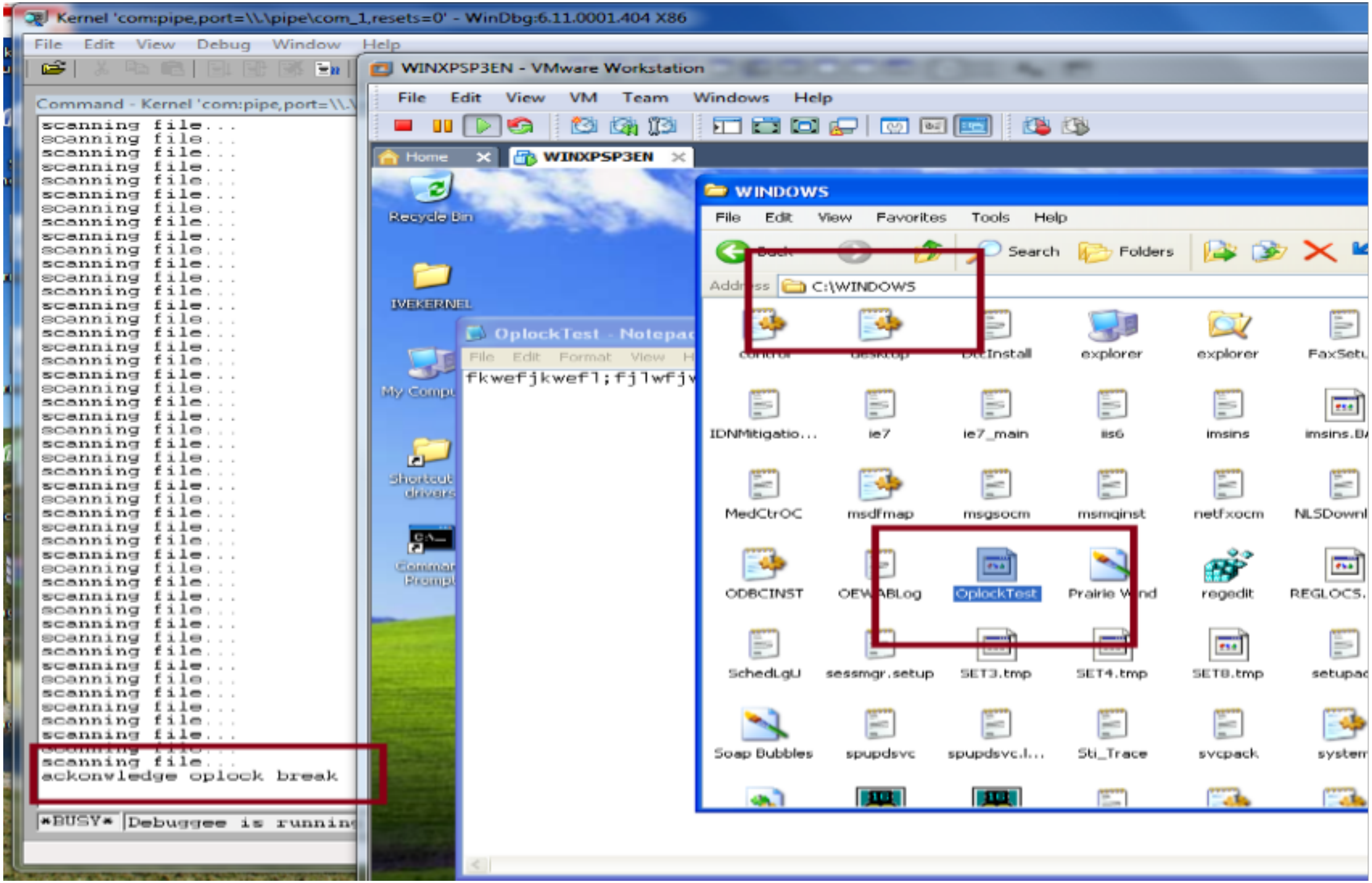
```
*BUSY* Debuggee is running
```

```
fkwefjkwefl;fjlwfjwesdfssdsdssss
```

Cannot create the C:\WINDOWS\OplockTest file.

Make sure that the path and filename are correct.

With filter oplock : acknowledge break and get out



```
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
scanning file...
ackonwledge oplock break
```

```
*BUSY* Debuggee is runnin
```

```
fkwefjkwefl;fjlwfjw
```

Windows 7 oplocks

- Support Oplock upgrade
- Support RWH (batch), RW (LV2), RH, R (LV1)
- Multiple handles can share an exclusive oplock through a “lease key” (GUID).

Q&A