SDC 19

September 23-26, 2019
Santa Clara, CA

INTRODUCE IN-KERNEL SMB3 SERVER CALLED CIFSD

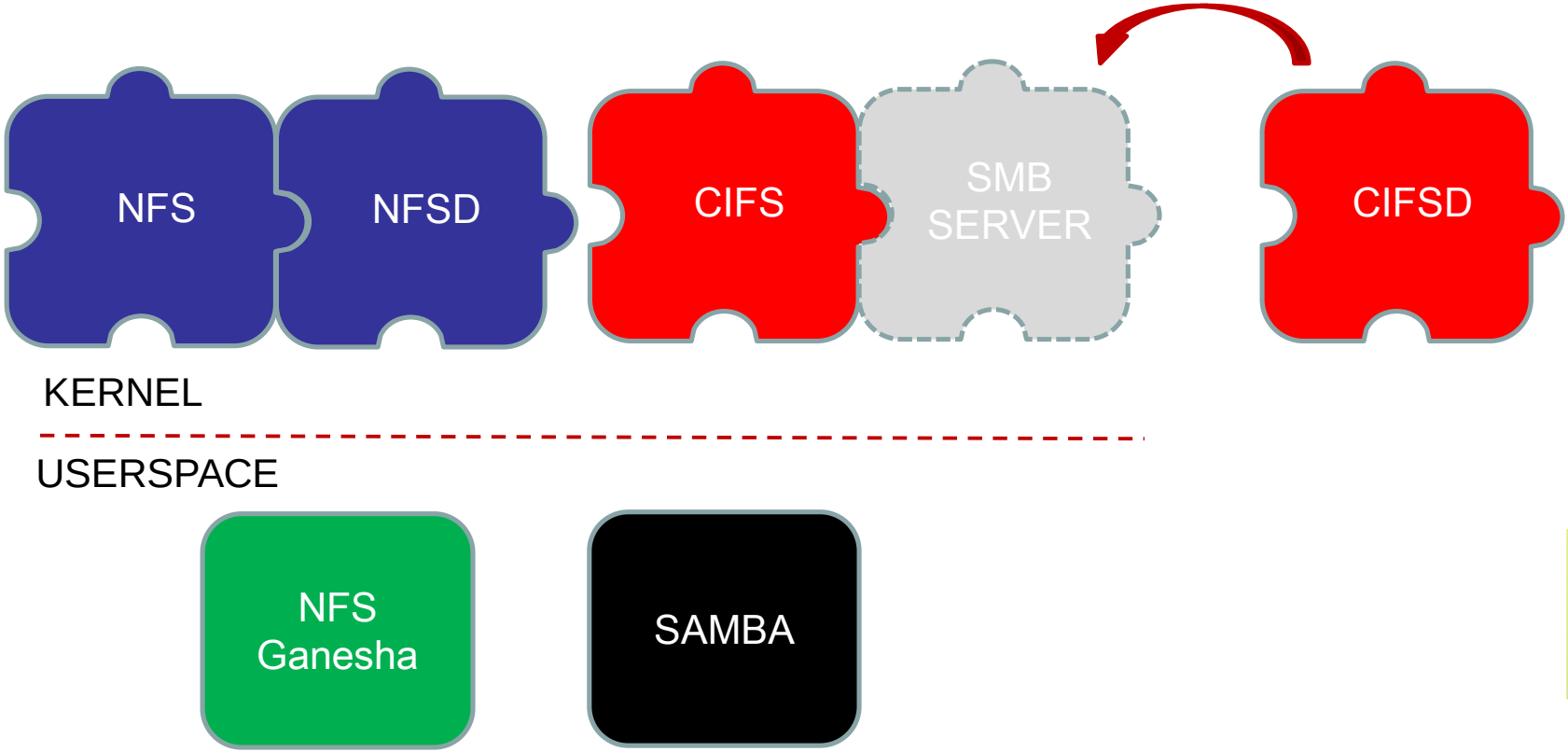
Namjae Jeon
Samsung Electronics
Hyunchul Lee
LG Electronics

TOPIC

- Introduction
- Architecture
- Performance/Stability/Compatibility
- SMB over RDMA
- Plans

INTRODUCTION

Fill the blank of kernel SMB Server



Full SMB stack support

- SMB1 ~ 3.1.1 version support
- SMB1, 2.0 optional and disabled by default
- Pending SMB2 notify
  - fsnotify functions are not exported
  - raise a request to export them

Supported Features

- Authentication support
  - NTLM/NTLMv2
- Performance features
  - Oplock/lease, compounding, copy offload and SMBDirect
- Security features
  - Signing (HMAC-SHA256, AES-CMAC)
  - Encryption (AES-CCM/GCM)

linux-cifsd project

- Github Repos
  - https://github.com/cifsd-team/cifsd
  - https://github.com/cifsd-team/cifsd-tools
- Mailing-list
  - linux-cifsd-devel@lists.sourceforge.net
- Supported Linux kernel versions
  - Linux 4.1 kernel or later

Test automation

- Use Travis CI
  - Github integration to test per pull request
- Build test
  - Linux 4.1 kernel and The latest Linus-tree
- Stability test
  - Xfstests(72)
  - Smbtorture(180)

DCERPC

- Need minimum implementation
  - Commands needed to work as file server
- DCERPC engine is located in userspace
- DCE/RPC response is prepared from the user space
- passed over to kcifsd kernel thread

Configuration

- User configure parameters in smb.conf
- Format is compatible with samba's one
- Minimum implementation
  - Select what user generally use a lot
  - Add more parameters users requested
- List-up supported parameter list in smb.conf.example in cifsd-tools

SMB1 disabled by default

- Keeping for clients which doesn't switch to SMB2 yet
- Planning to remove SMB1 in cifsd when contributing to mainline

OpenWRT Built-in cifsd

- OpenWRT is wireless router open source project
- Built-in cifsd as file sharing function
- OpenWRT had built-in Samba, Why ?
  - Easy cross compile without special handling
  - Small binary size(page 16)
  - Low CPU usage and better performance

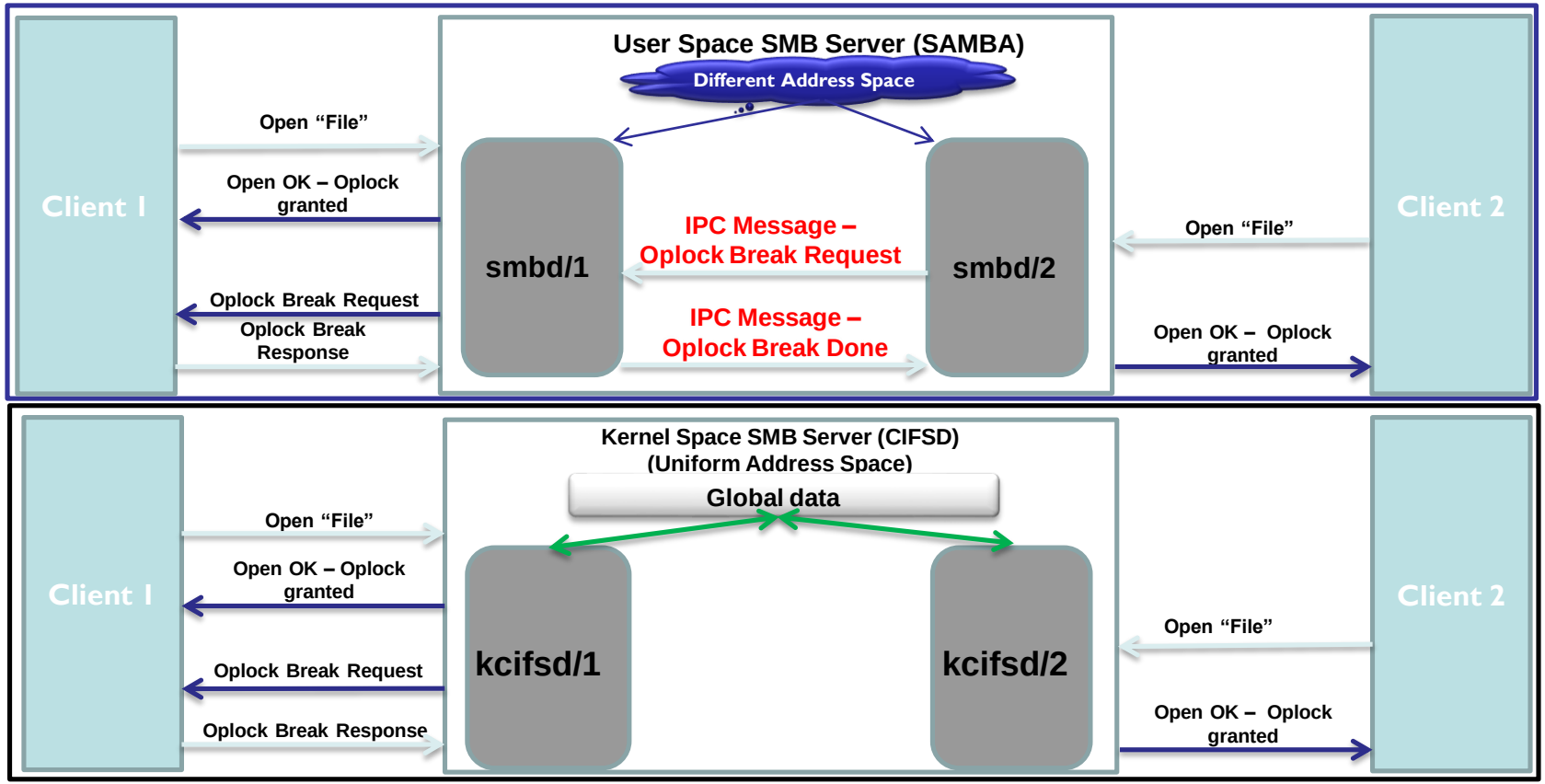
KEY CONCEPTS

In-kernel server

- No system call cost
- Shorter path to use VFS and network stack
- no duplicate memory allocation (Filesystem metadata)

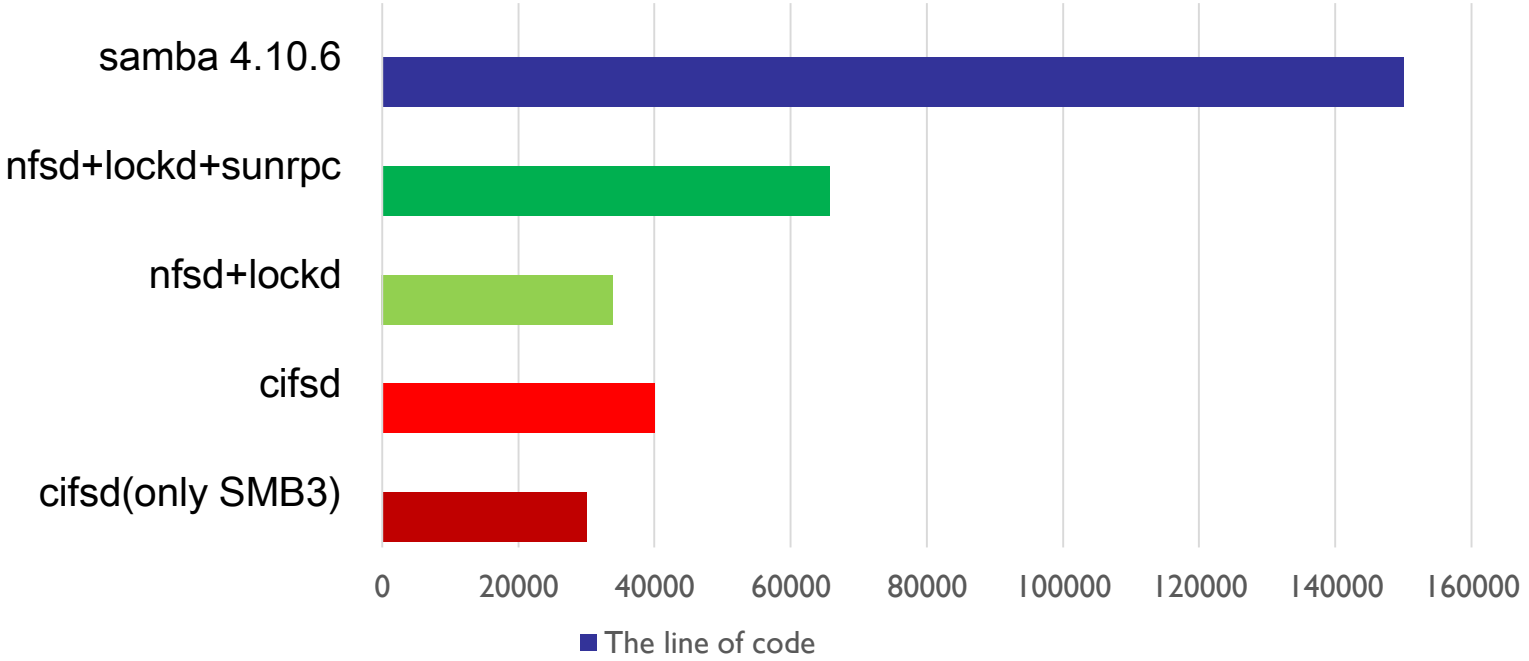
Optimized SMB over RDMA

Oplock lease/better handling



Minimal SMB3 server

Code line count comparison



Lightweight file share

- Reported by Andy Walsh(OpenWRT)

Binary Size	Main	Extra	Total
cifsd	128KB(cifsd kmod, tools)	61KB(crypto kmods) + 872KB(glib2)	1061KB
samba4	6MB(samba libs, server package)	64KB(libtirpc, etc)	6064KB

ARCHITECTURE

User and Kernel thread

ucifsd (User-level thread)

Work related to non-performance, more memory/ more resource feature. i.e. DCERPC, user/share management, configuration.

USERSPACE

kcifsd(Kernel-level thread)

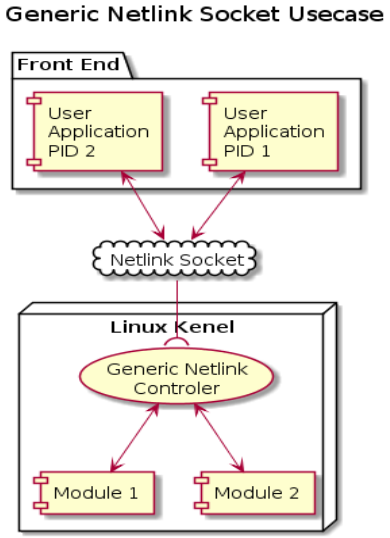
Work related to performance, It contain SMB engine, transport, server handler, oplock/lease.

KERNELSPACE

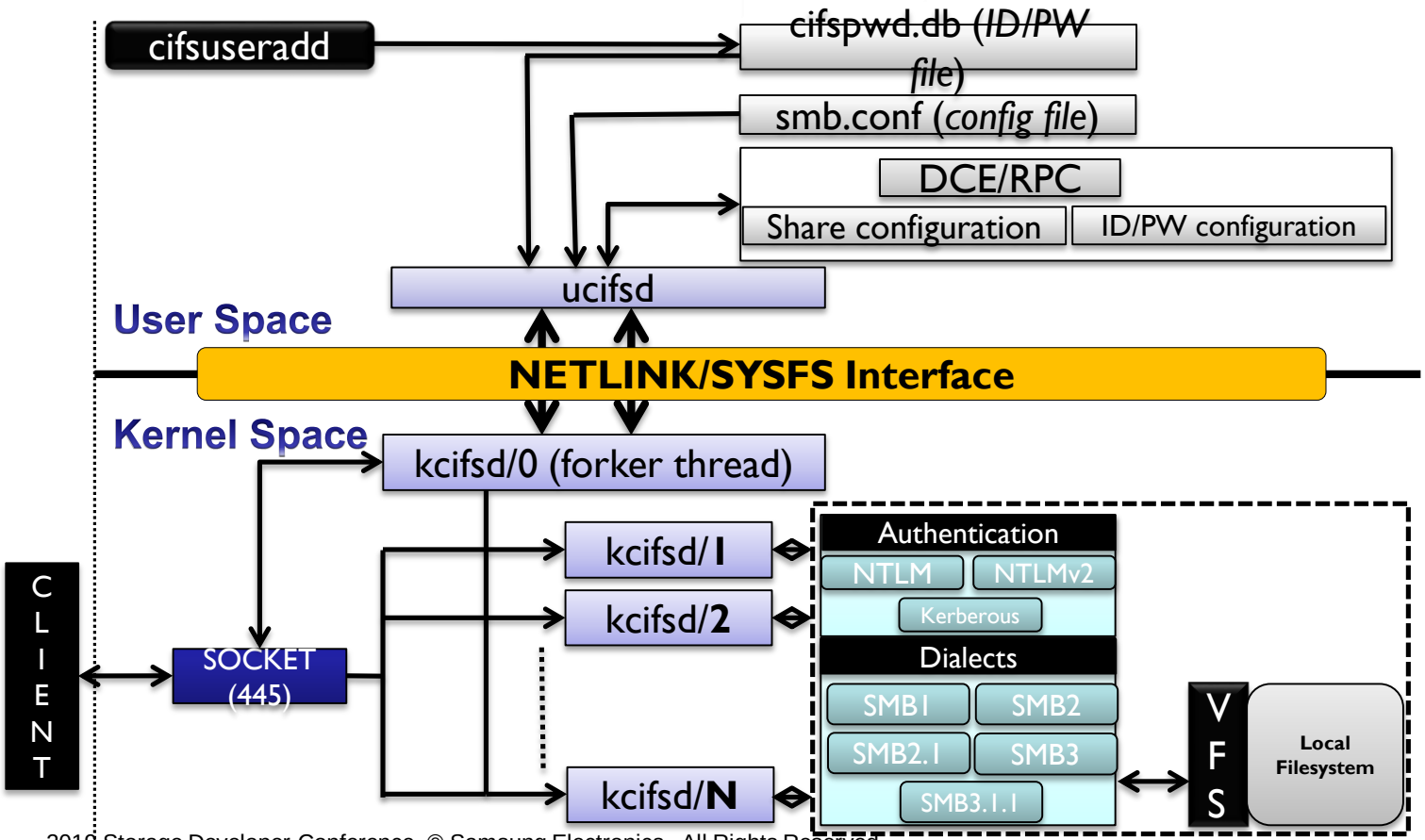
IPC between kernel/user

- Use Netlink interface
- Specify a few command-set

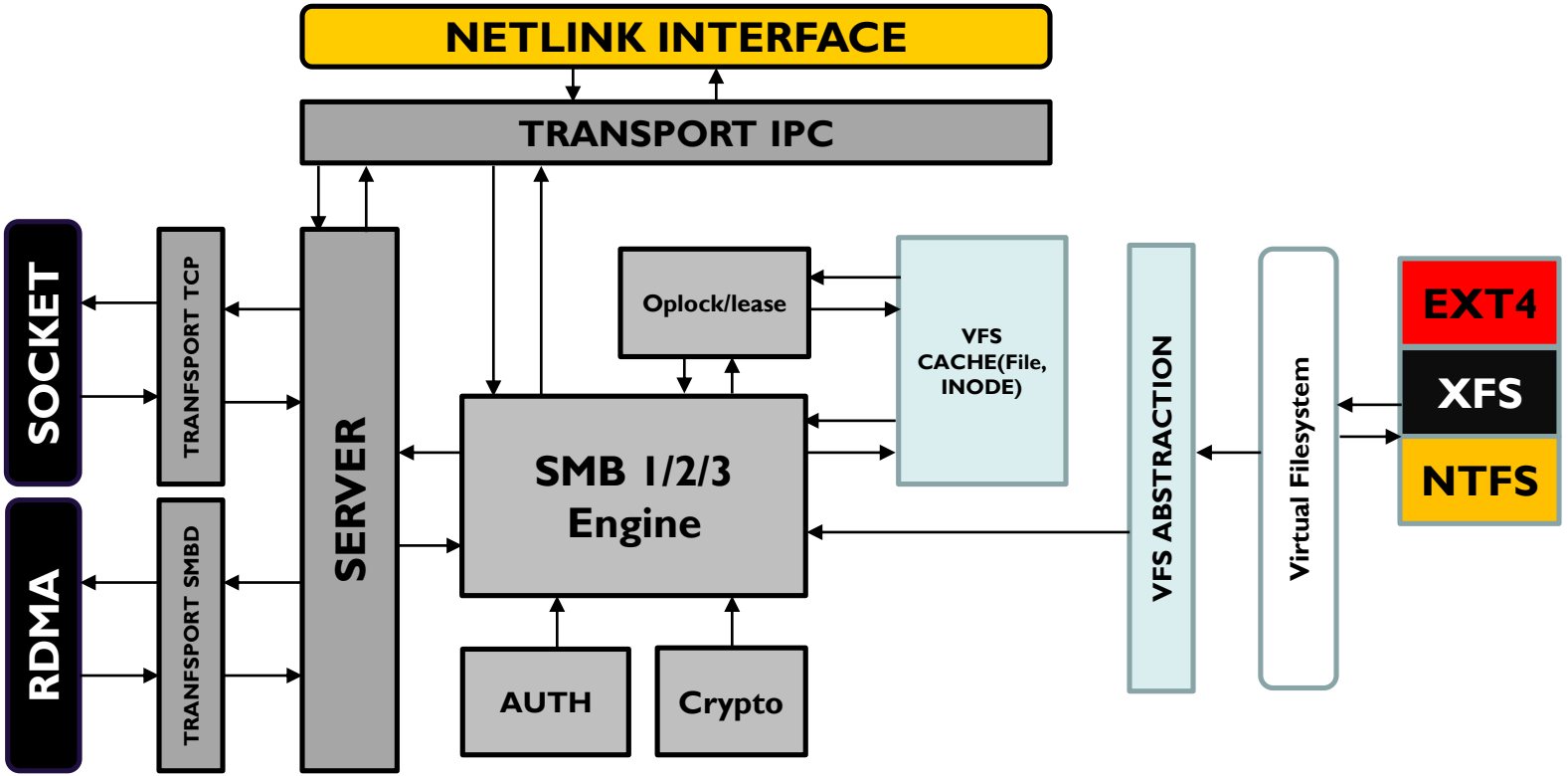
Name	Purpose
CIFSD_EVENT_HEARTBEAT_REQUEST	Monitor cifsd is alive
CIFSD_EVENT_STARTING_UP CIFSD_EVENT_SHUTTING_DOWN	Transfer the initial information necessary for the start and shutdown
CIFSD_EVENT_LOGIN_REQUEST CIFSD_EVENT_LOGIN_RESPONSE	Transfer the user account / password information necessary for login
CIFSD_EVENT_SHARE_CONFIG_REQUEST CIFSD_EVENT_SHARE_CONFIG_RESPONSE	Transfer the share configuration
CIFSD_EVENT_TREE_CONNECT_RESPONSE CIFSD_EVENT_TREE_DISCONNECT_REQUEST	Transfer the tree connect info
CIFSD_EVENT_RPC_REQUEST CIFSD_EVENT_RPC_RESPONSE	Transfer DCERPC requests



Architecture



Components

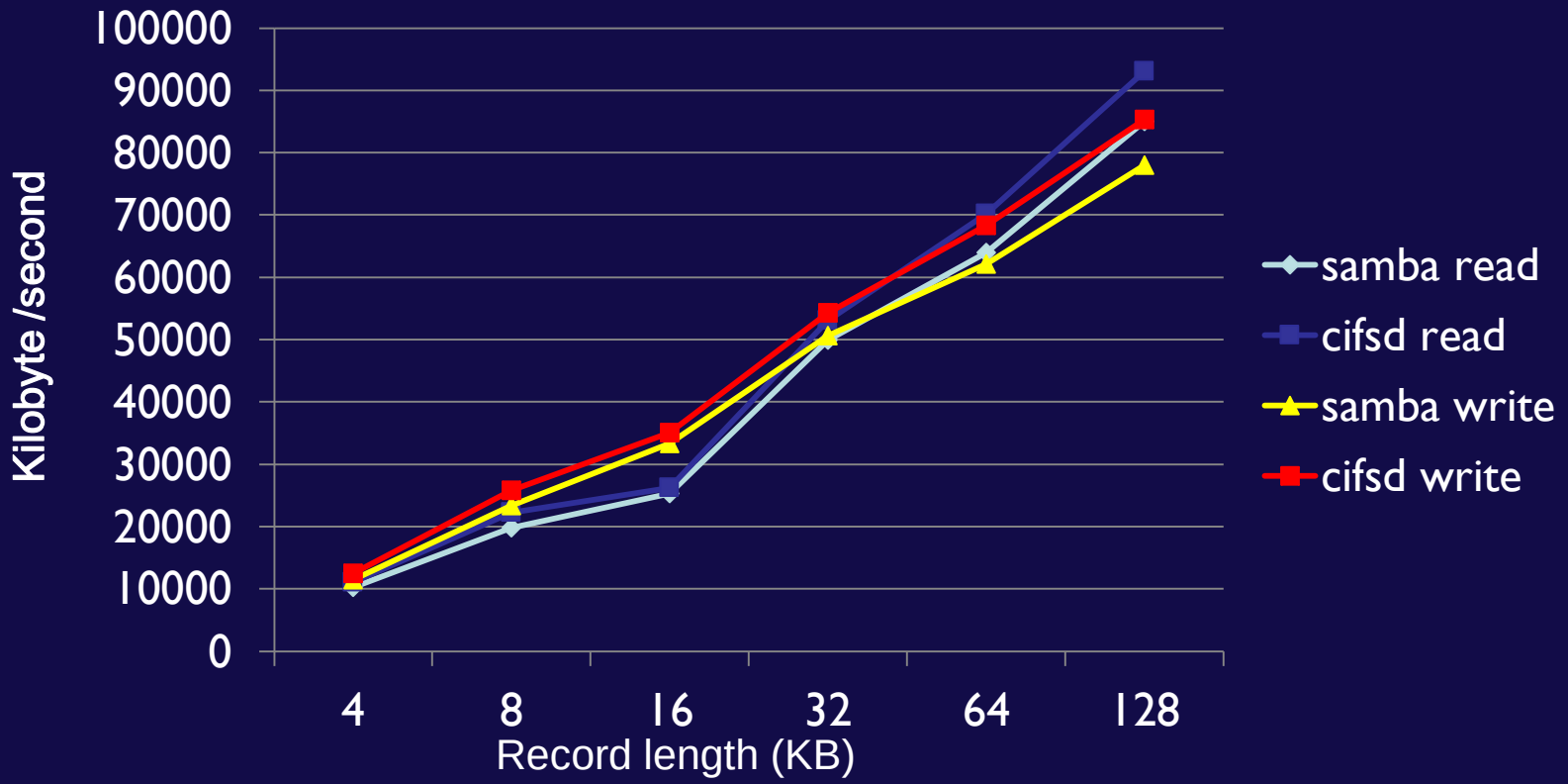


PERFORMANCE

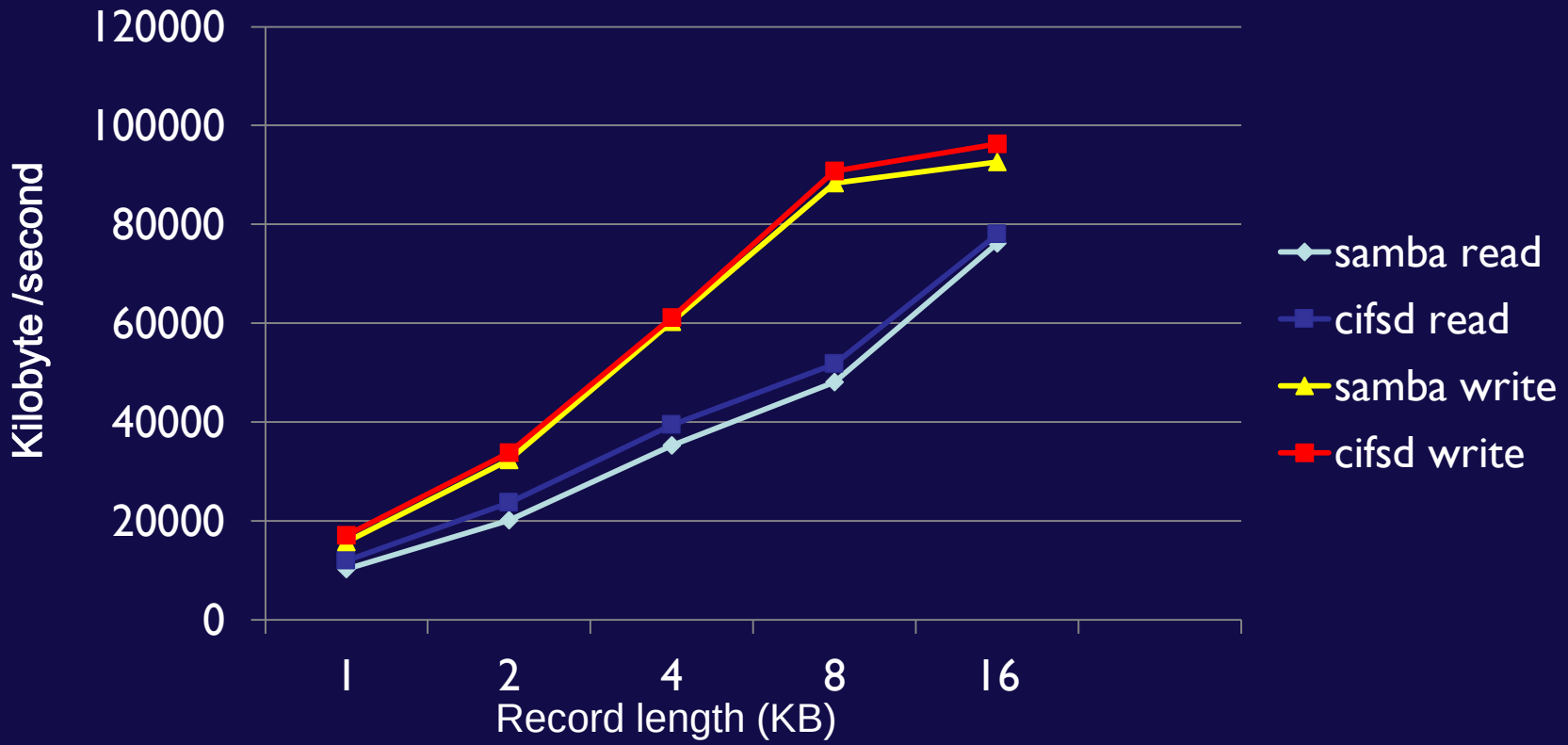
Test Environment

- Direct connection on TWO PC
- Tools : IOZONE, fileop, bench-oplock
- Client : CIFS Client(SMB3.11)
- Share : tmpfs
- SMB Server : samba 4.7.6 and cifsd

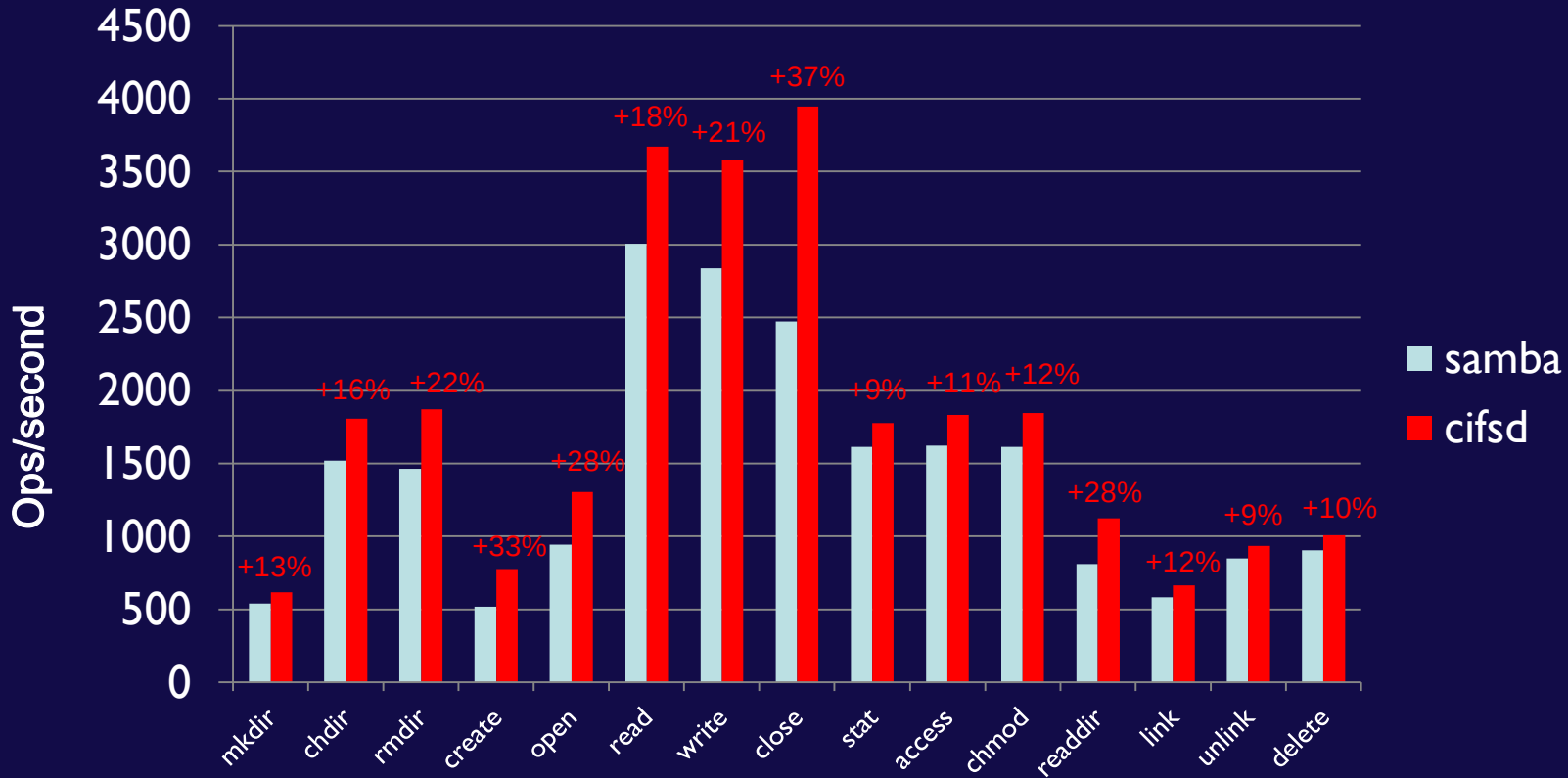
Single Read/Write iozone throughput



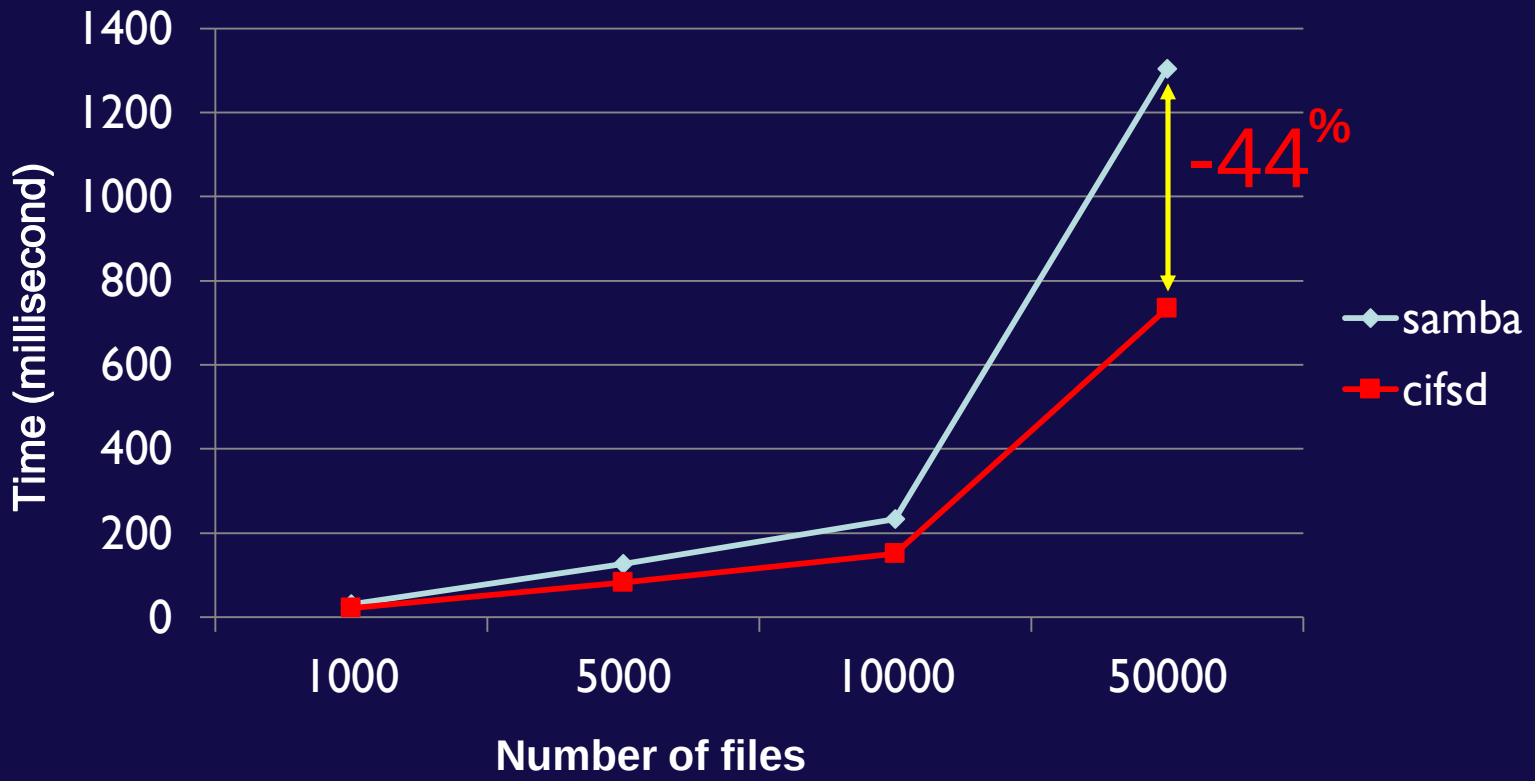
Multi Read/Write iozone throughput



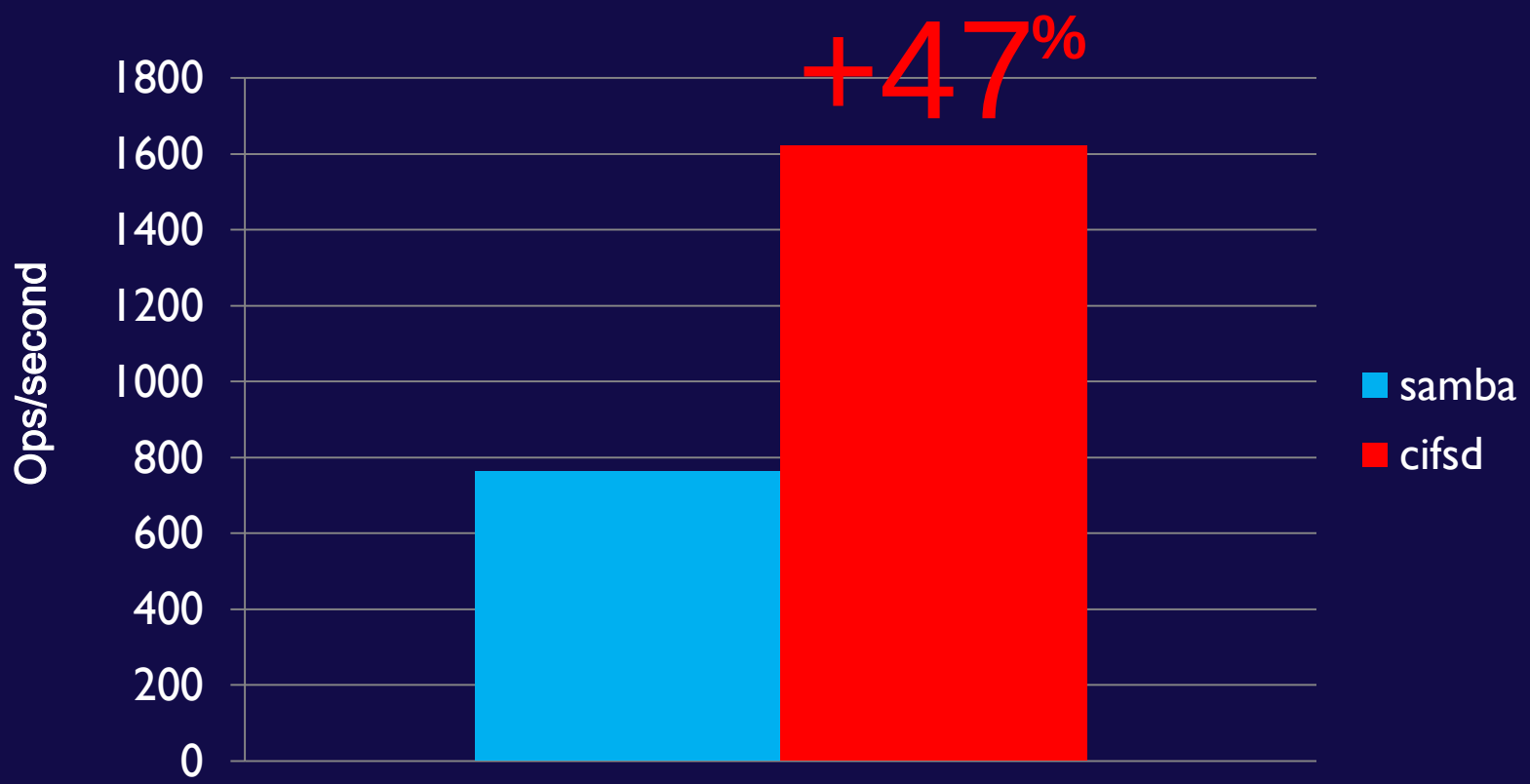
Fileop throughput



File lookup time(ls –l)

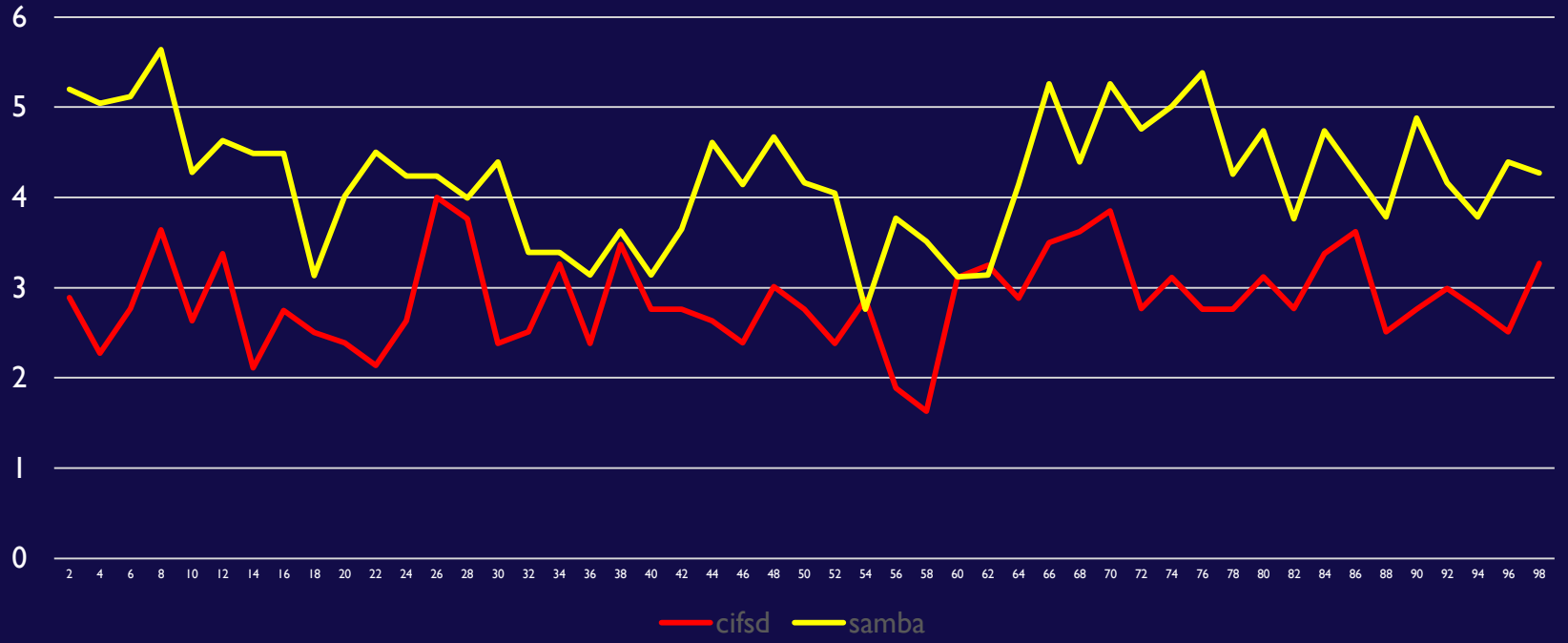


Bench oplock

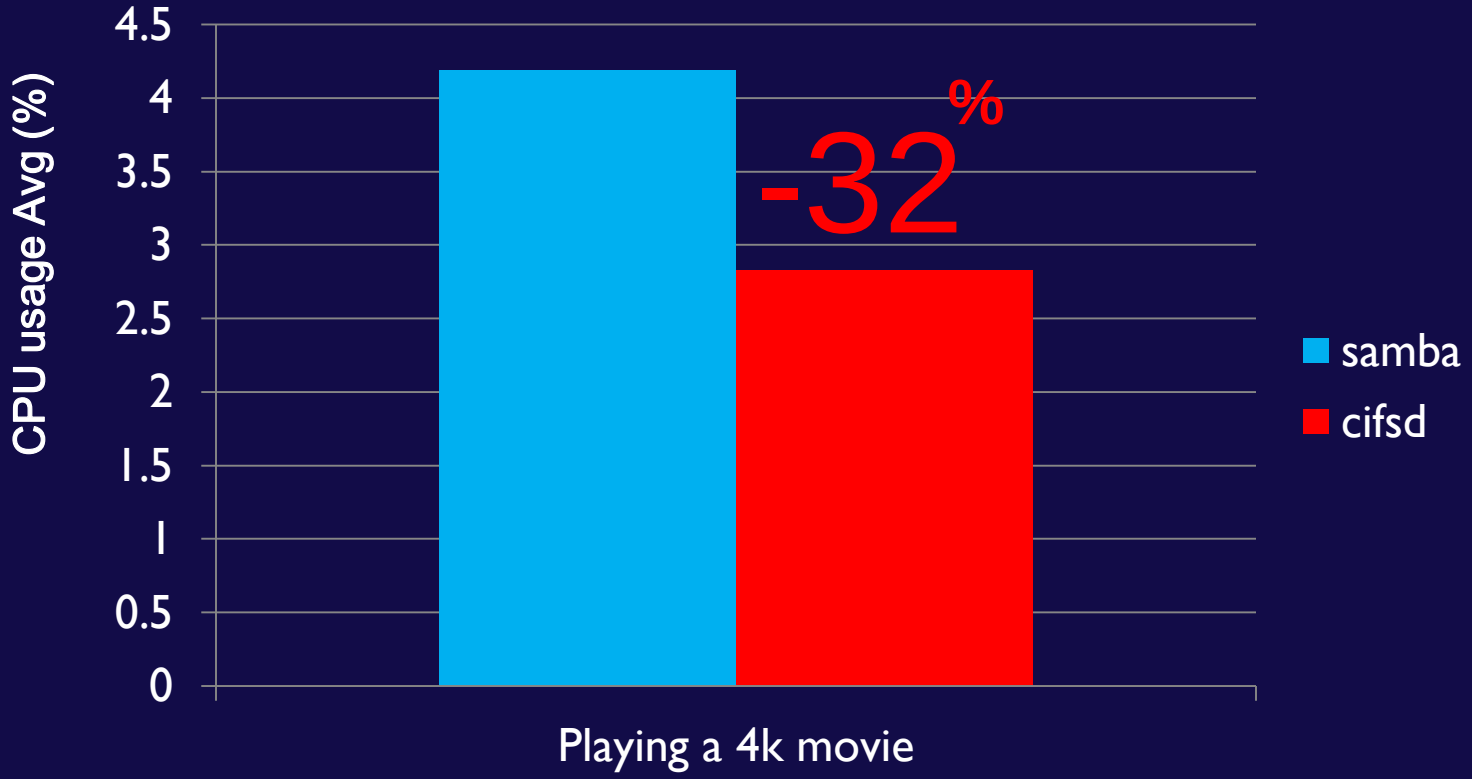


CPU Usage

Playing a 4K movie
mpstat –p ALL 2



CPU Usage



Compatibility

SMB CLIENT VERSIONS	CIFSD SUPPORTED
Windows XP (SMB 1.0)	✓
Windows Vista (SMB 2 .0)	✓
Windows 7 (SMB 2.1)	✓
Windows 8 (SMB 3.0)	✓
Windows 10 (SMB 3.1.1)	✓
MacOS(~ High Sierra)	✓
Ubuntu File Explorer	✓
Linux CIFS Client(linux v5.3-rc6)	✓

Stability

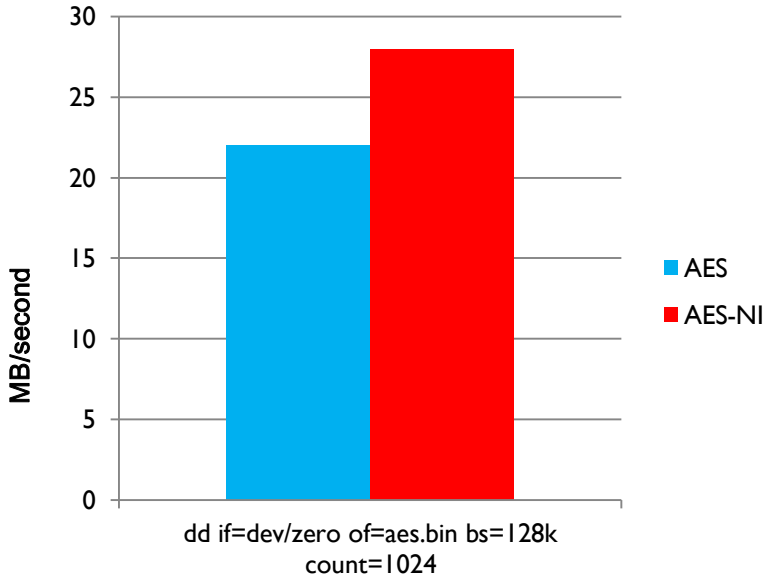
- XFSTESTS
  - Total : 82, PASS : 82
  - Detail test result link
- SMB TORTURE
  - Total : 311, PASS : 205, FAIL : 11, N/S : 95
  - Detail test result link

Catching up cifs's new features

- SMB Direct
- Posix extension mode bit
- GCM encryption
- rsize and wsize increase
- Test Automation

Hardware Acceleration

- Performance advantage through HW acceleration
- Accelerate encryption(AES-GCM) performance with AES-NI support in kernel



SMB over RDMA

SMB Direct

- New transport protocol for supporting SMB over RDMA
  - Introduced in SMB 3.0 (Windows Server 2012)
- Operates over Reliable RDMA transports
  - Infiniband, RoCE, iWARP
- Defines transport framing for SMB exchange

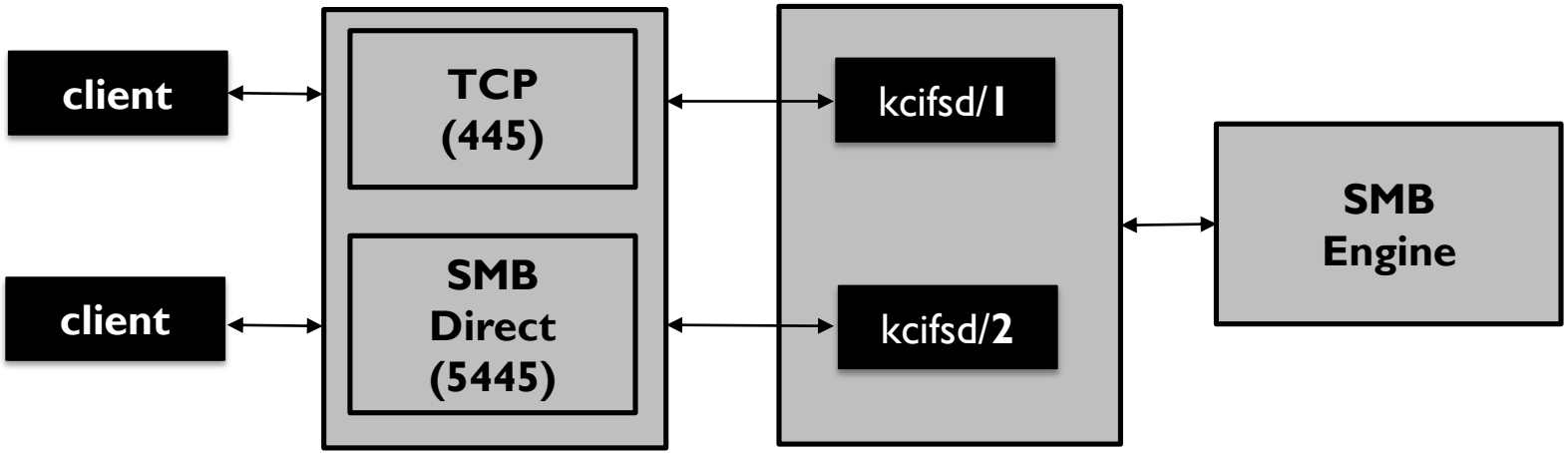
SMB Direct

- Negotiate capabilities
  - Version, message size, credits
- Supports datagram-type send / receive for the exchange of SMB messages
  - RDMA send / receive
  - Fragmentation / Reassembly
  - Credits management

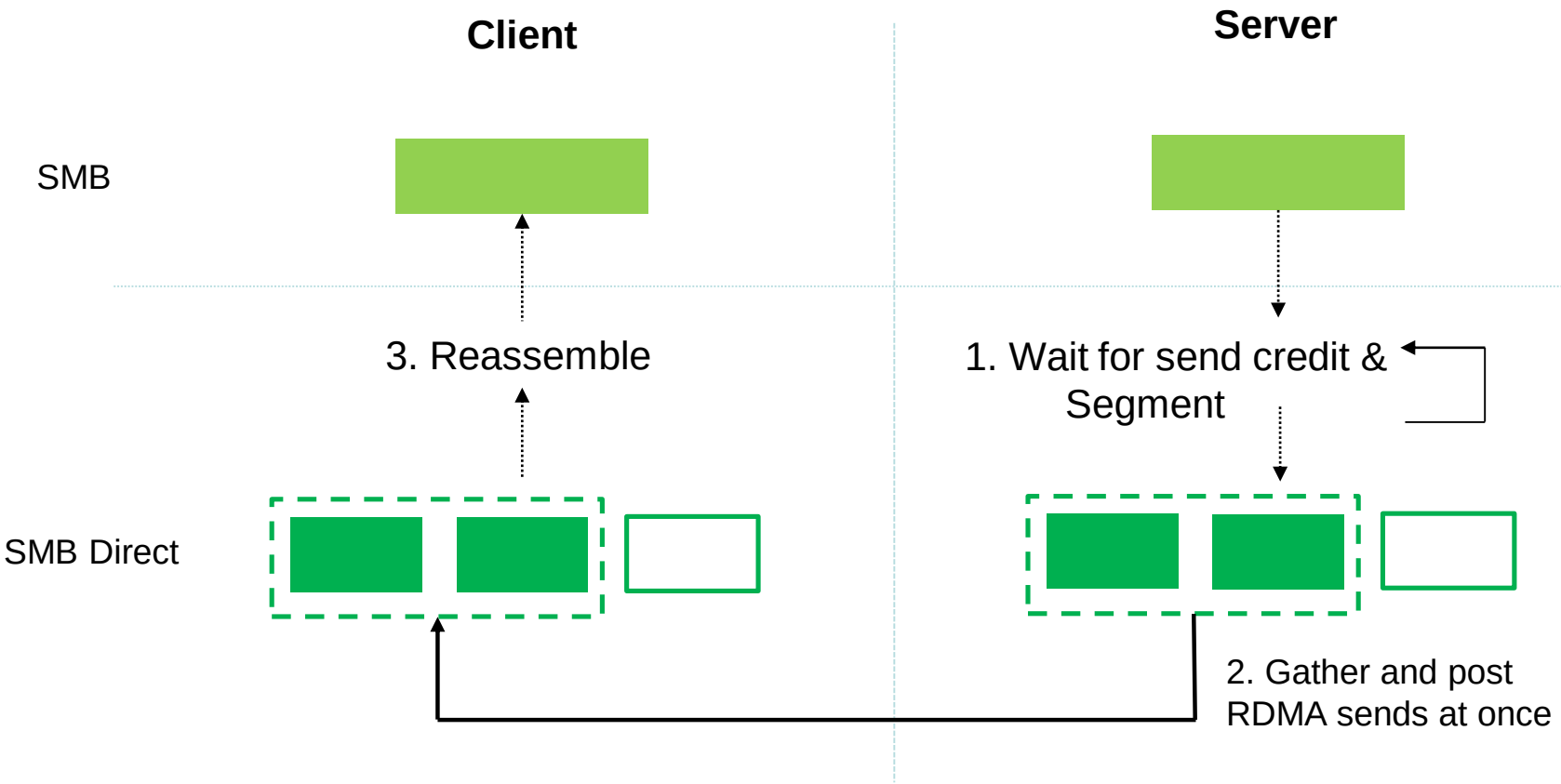
SMB Direct

- Supports remote direct data placement for moving data between memory buffers on each peer
  - RDMA read / write
  - Memory registration and advertisement of the location to another peer
  - Remote memory invalidation (SMB 3.02)

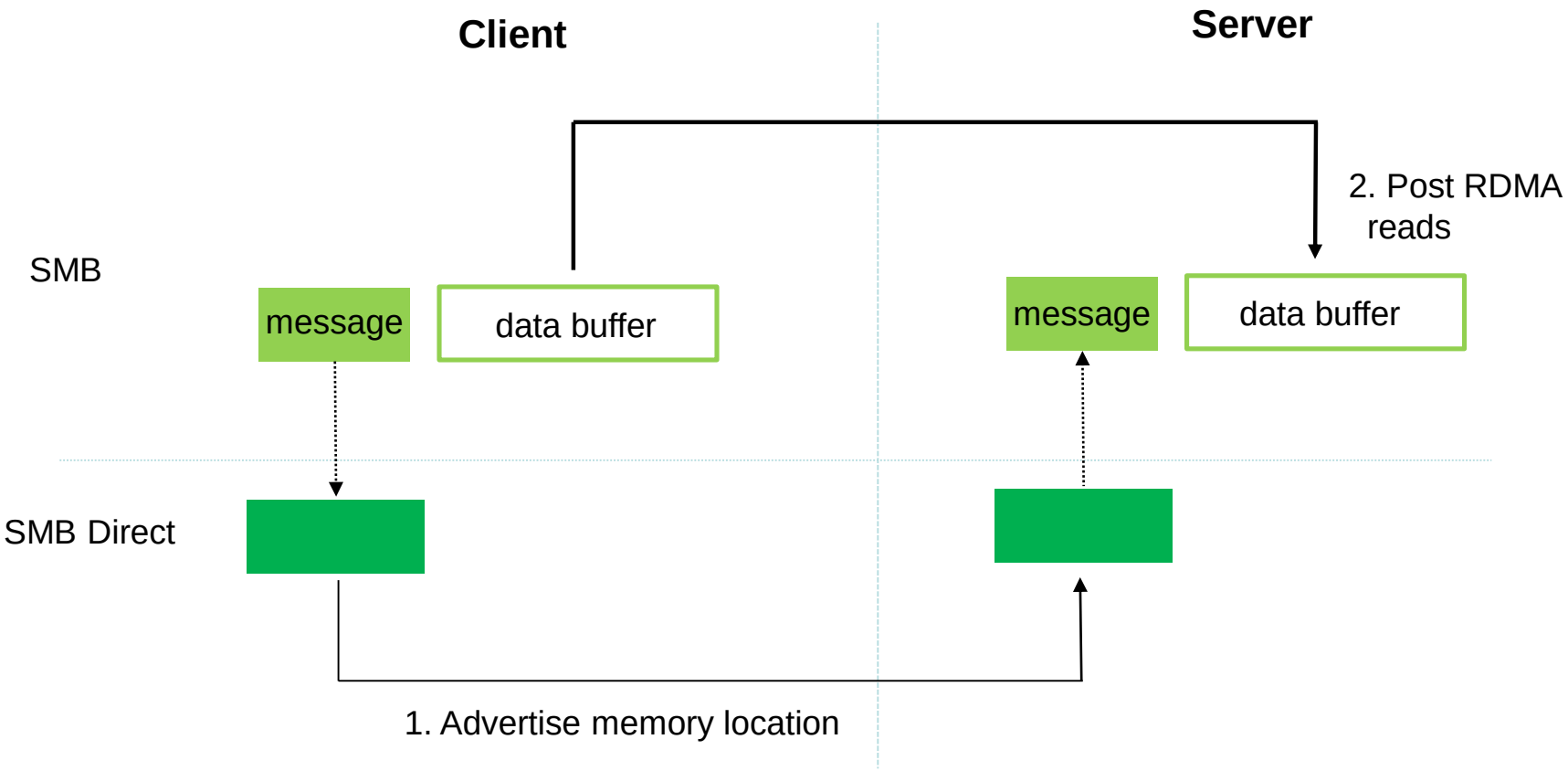
SMB Direct and TCP



Send SMB3 Response



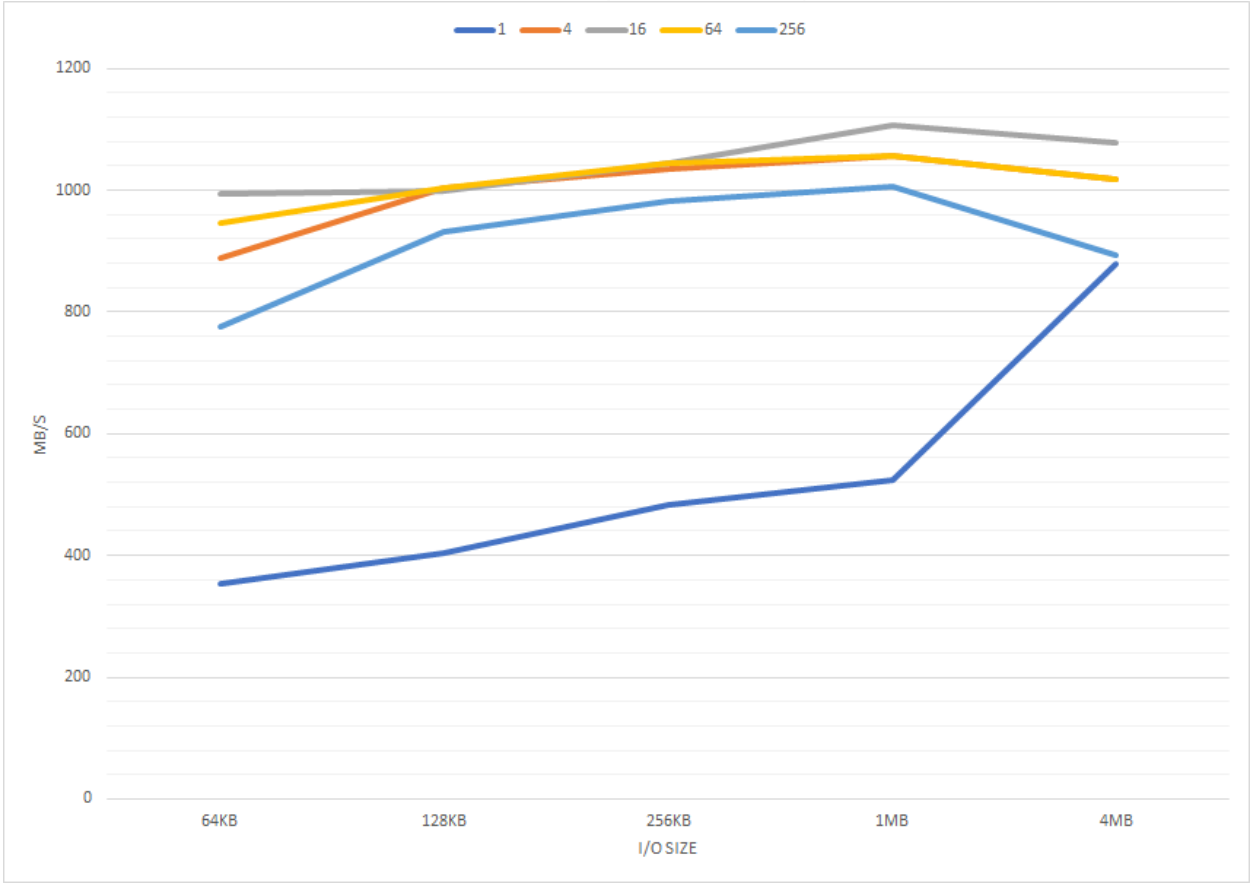
Transfer File Data



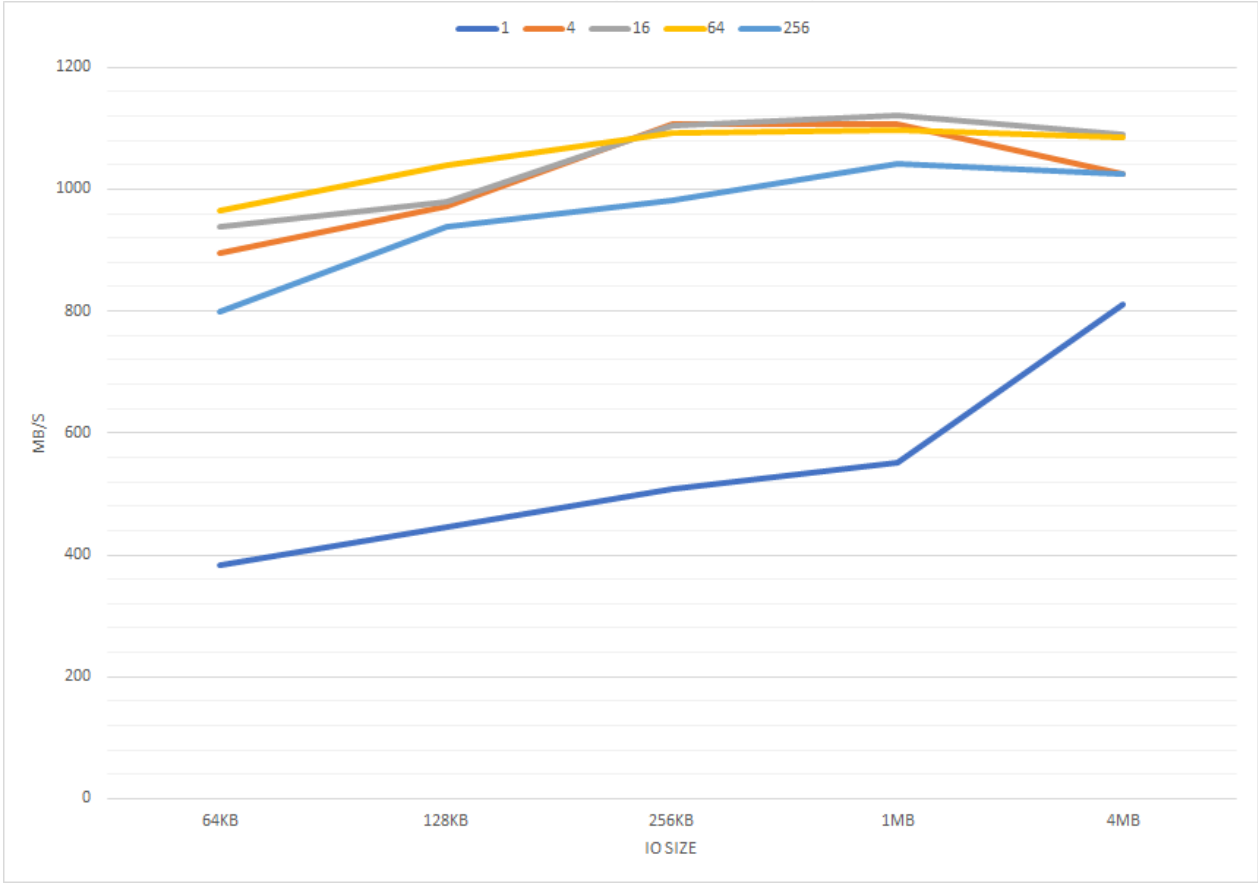
Test Environment

- Client / Server PC
  - Intel i7-3770k 3.90GHz
  - 16GB RAM
  - Mellanox **ConnectX-2 10G**
- Direct connection on two PC
- Linux kernel v5.2 cifs client (SMB 3.02)
- fio

Write Throughput



Read Throughput



Plans

- Windows Protocol Testsuite
- SMB2 notify
- WinACL support
- Check compatibility with different NICs
- Support Multichannel


The end, Thank you!
Any question?