HTTP-FUSE Xenoppix

http://unit.aist.go.jp/itri/knoppix/http-fuse/xen/index-en.html

Kuniyasu Suzaki [1] , Toshiki Yagi [1], Kengo Iijima [1],
Kenji Kitagawa [2], Shuichi Tashiro [3]
[1] National Institute of Advanced Industrial Science and Technology (AIST),
[2] Alhpa Systems Inc,
[3] Information-Technology Promotion Agency (IPA)

Outline

- Purpose of HTTP-FUSE Xenoppix
- HTTP-FUSE CLOOP
  - Network/Storage transparent Virtual Block Device
- Current Status & Performance
  - Some OSes (Plan9, NetBSD, and Linux) boot with 6MB bootable CD-ROM
- Discussions, Conclusions
- DEMO

Purpose of HTTP-FUSE Xenoppix

- New OS Circulation Methods
  - (Basically) Internet Thin Client for anonymous user
    - Client PC has *Virtual Boot Loader* only
      - Traditional methods (PXE Client boot):
        - » Kernel and miniroot are obtained by TFTP
        - » Root File System is obtained by NFS
      - Our methods:
        - » HTTP is used instead of TFTP and NFS.
    - It makes easy to try other Distributions (Gentto, Fedora,etc) and other OSes (Plan9, ***BSD, OpenSolaris,etc)
  - It can be saved on a local HD
    - Unfortunately current version is not normal OS installation yet. It acts as a VM OS image.

As a Virtual Boot Loader

- Xenoppix works
  - KNOPPIX boots as a Domain0 (Host) OS.
    - “AutoConfig of KNOPPIX” is advanced at device-detect and driver-setup.
    - Prepare Xen (Dom0) environment on anonymous PC
  - “Xen” prepares virtual machine environment.
    - Para-virtualization enables to boot some OSes on x86 PC
    - Full-virtualization enables to boot any OSes on Intel VT and AMD SVM.

What we developed

- Most difficult point is to get Disk Image (Root File System)
  - Existing Network File Systems & Network Block Devices are not good, because they suppose special server software and special port.

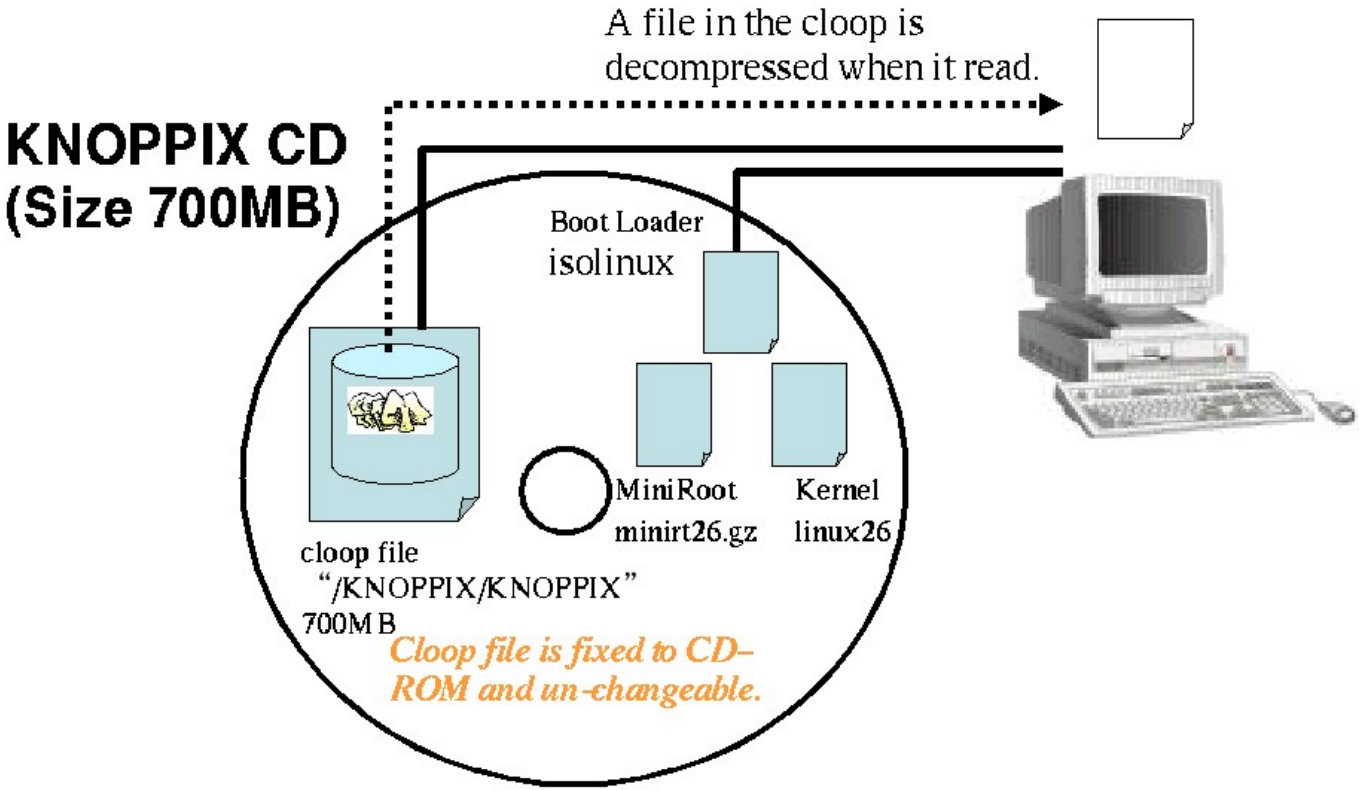
- We developed a Virtual Block Device
  - HTTP-FUSE CLOOP
    - Network/Storage Transparent
      - Handle network (dis/re)-connection
    - The requirement on server is to distribute files only.
      - It is based on HTTP because HTTP is popular and powerful file distribution methods

Related Block Device/File System

	Port	Connection	Proxy/Cache	Mountable (Direct Execution)	Performance
HTTP-FUSE (Block Device)	80	LESS	○	○	× Some techniques cover
iSCSI (Block Device)	3260	FULL	×	○	○
NBD (Block Device)	1077	FULL	×	○	○
NFS (File System)	2049, 2050	FULL	△	○	○
AFS (File System)	7000,7001,7002,7003,7004,7005,7006,7007,7008,7009	FULL	○	○	○
SFS (File System)	4	FULL	×	○	○
WebDAV (File Sharing System)	80	LESS	○	×	△

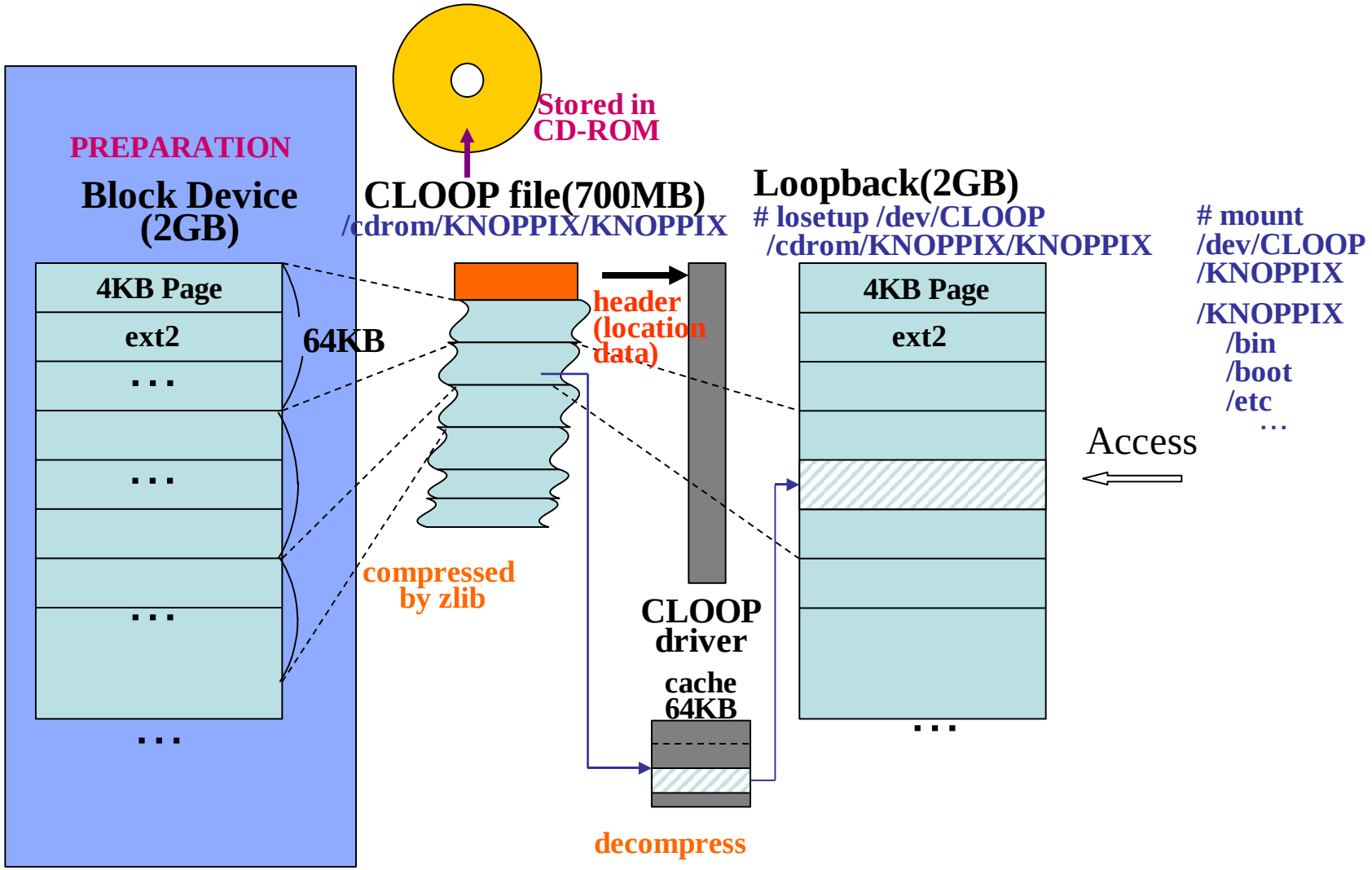
HTTP-FUSE CLOOP

- HTTP-FUSE CLOOP is based on CLOOP (Compressed Loopback device) which is used on KNOPPIX.
- CLOOP enables to pack 2.0GB contents (Root File System) to 700MB CD-ROM.



CLOOP (Compressed Loopback)

- Each 64KB block is compressed by “zlib” and saved a CLOOP file.
  - The total size of a block device becomes between 1/2 and 1/3.
- CLOOP file has a header for block location information.
  - CLOOP driver searches a relevant block using the information.
- CLOOP driver has a cache.
  - Decompressed block is stored and reused if the next access fits to it.


PREPARATION
Block Device
(2GB)
4KB Page
ext2
64KB
Stored in
CD-ROM
CLOOP file(700MB)
/cdrom/KNOPPIX/KNOPPIX
header
(location
data)
compressed
by zlib
CLOOP
driver
cache
64KB
decompress
Loopback(2GB)
# losetup /dev/CLOOP
/cdrom/KNOPPIX/KNOPPIX
4KB Page
ext2
Access
mount
/dev/CLOOP
/KNOPPIX
/KNOPPIX
/bin
/boot
/etc

Drawback of CLOOP

- CLOOP is good for CD bootable Linux but …

- Still large file
  - it has whole block data, even if the data are compressed.
- It doesn't allow partial update.
  - We have to rebuild whole CLOOP file, even if 1 bit update.
- The problems are caused by *integration of "whole block data" and "location information"*.

- HTTP-FUSE CLOOP
  - Large loopback file is divided to many small block files.
  - separates block data and location information.

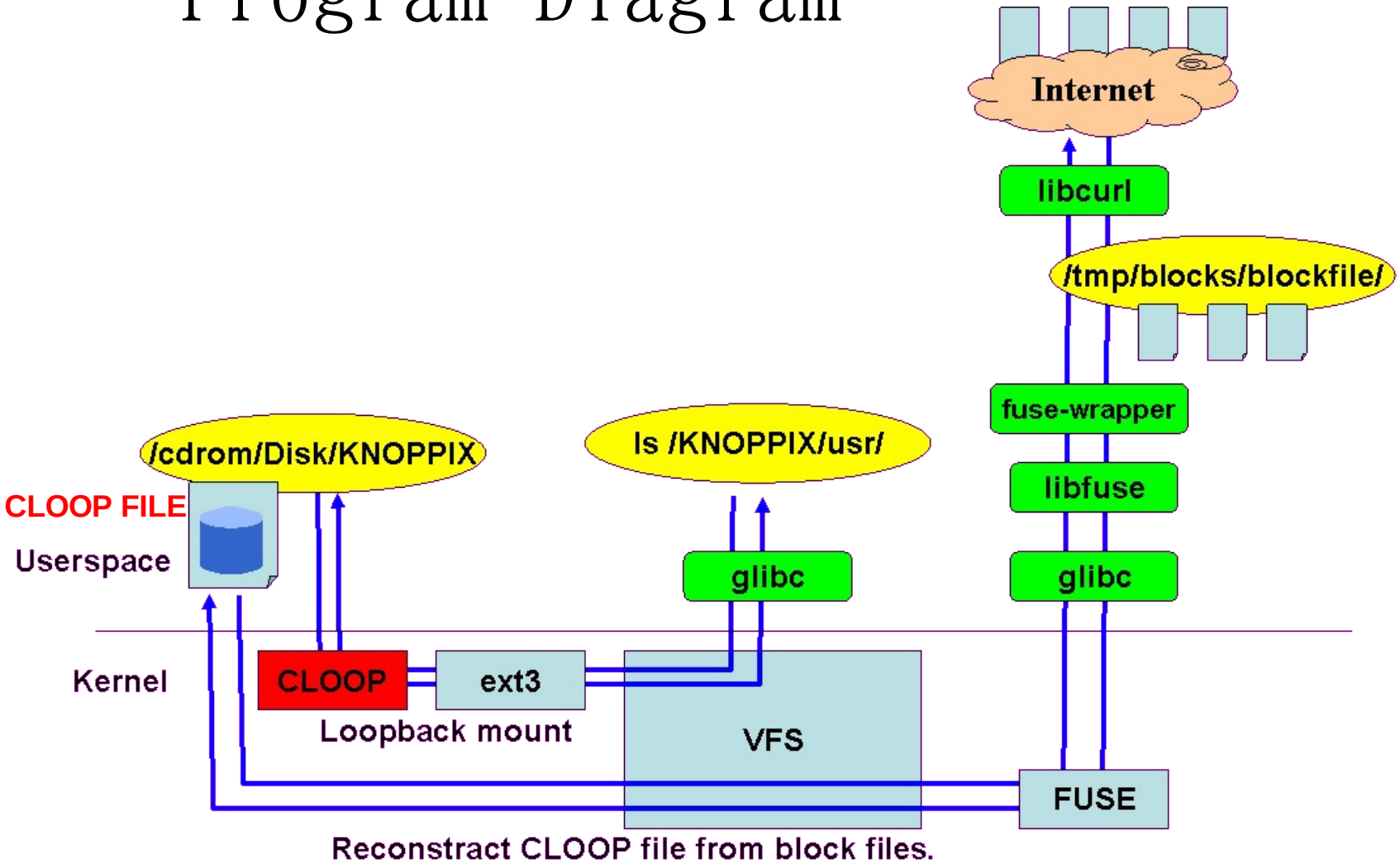
HTTP-FUSE CLOOP (1/2)

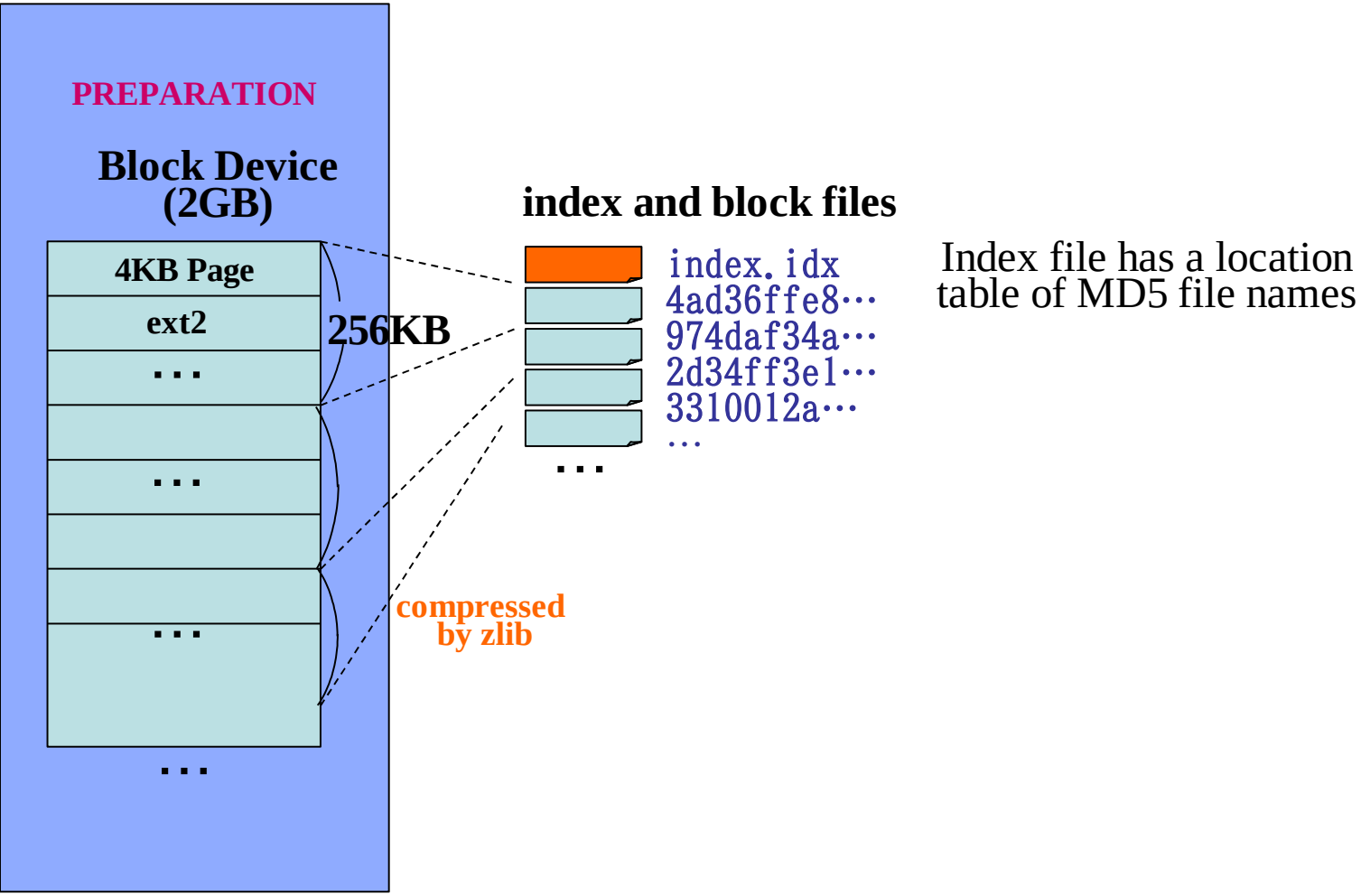
- Each compressed block is saved to each block file.
  - Current block size is 256KB because 64KB is too small.
    - Utilize TCP window size(64KB)
    - Compressed 256KB block data becomes about 100KB.
- Instead of a header of CLOOP file, “index” file works as location information of block files.
- Block file name is a MD5 value of its contents.
  - If there is a same contests blocks, they are held together a same name file and reduce total file space.
  - The basic idea is resemble to “Venti of Plan9”
- Block files are reconstructed to a CLOOP file by FUSE wrapper.
  - FUSE is a User-land File System.
    - http://fuse.sf.net

HTTP-FUSE CLOOP (2/2)

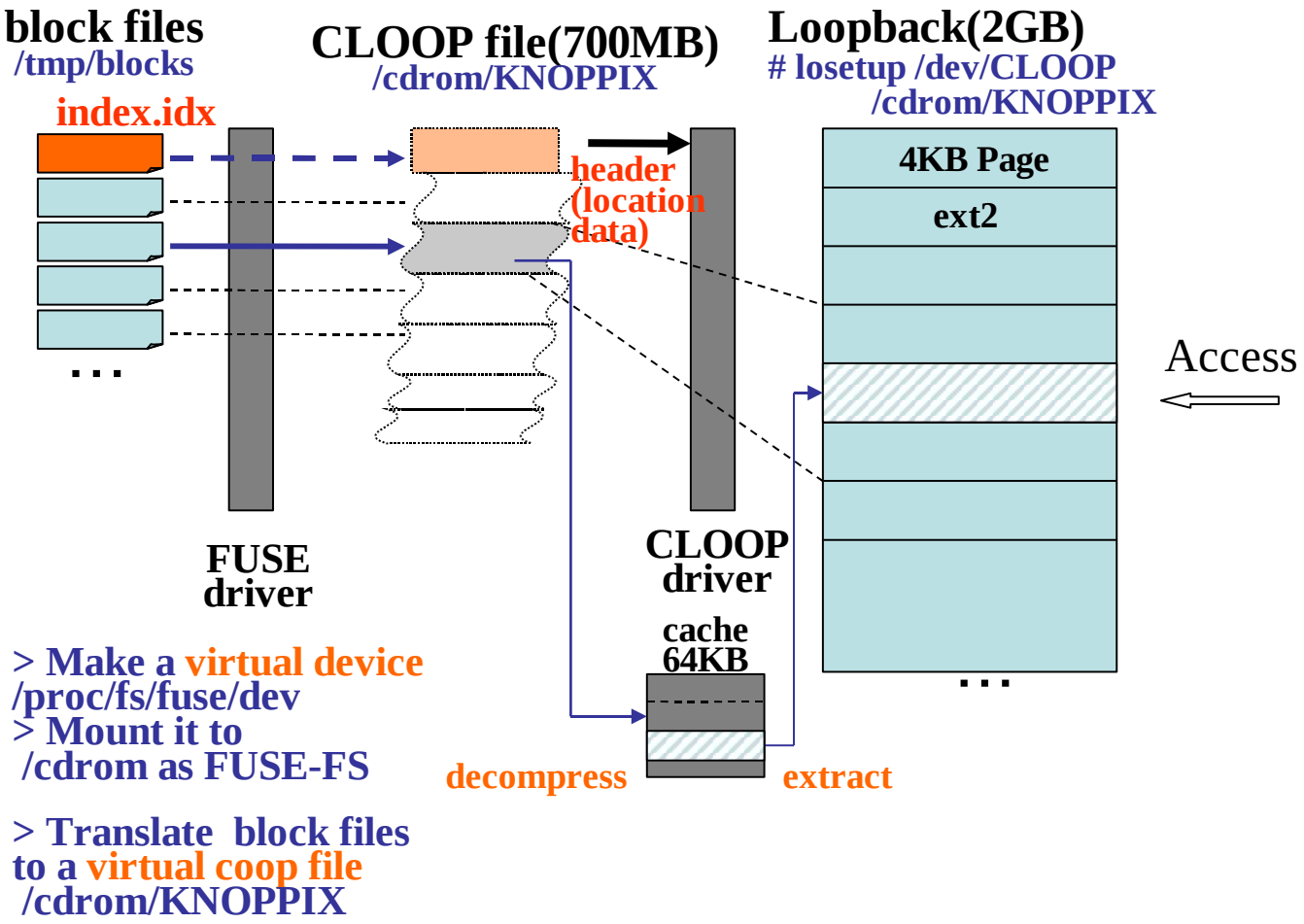
- When a file is updated or created on a existing block device, the relevant block files are newly created with new MD5 name. The “index” file are also renewed.
  - The file system on HTTP-FUSE CLOOP have to be updatable. ISO9660 is not suitable.
  - Old block files are reusable.
- HTTP for file deliver
  - Most popular and well designed. Web hosting is inexpensive.
  - 80 port is usually opened.
    - Other network block devices use special port which is usually closed.
- Block files are network/storage transparent.
  - Block files are cached and reused on local storage.
  - If necessary block files are stored in a local storage, network connection is not necessary.

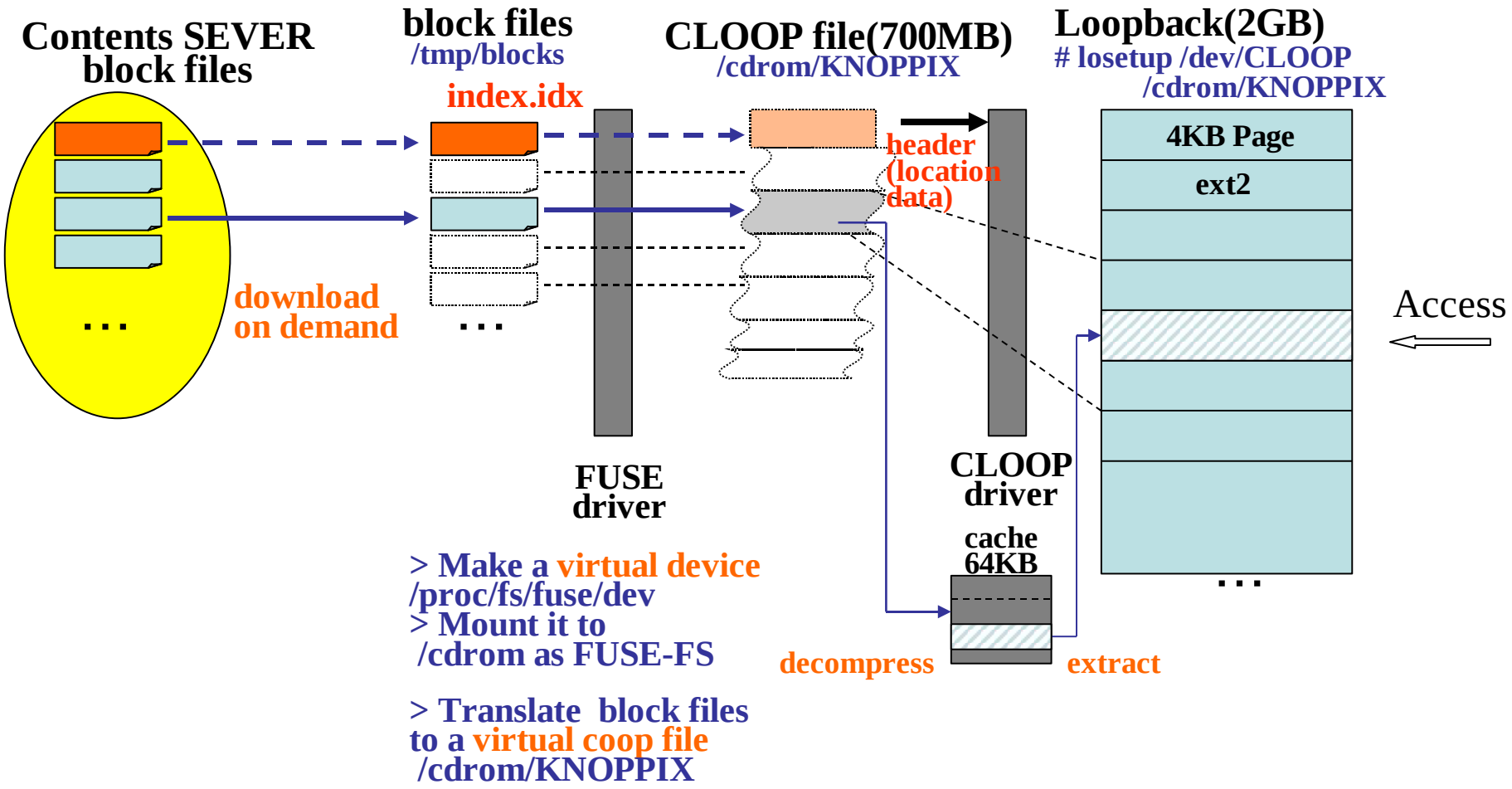
HTTP-FUSE CLOOP Program Diagram

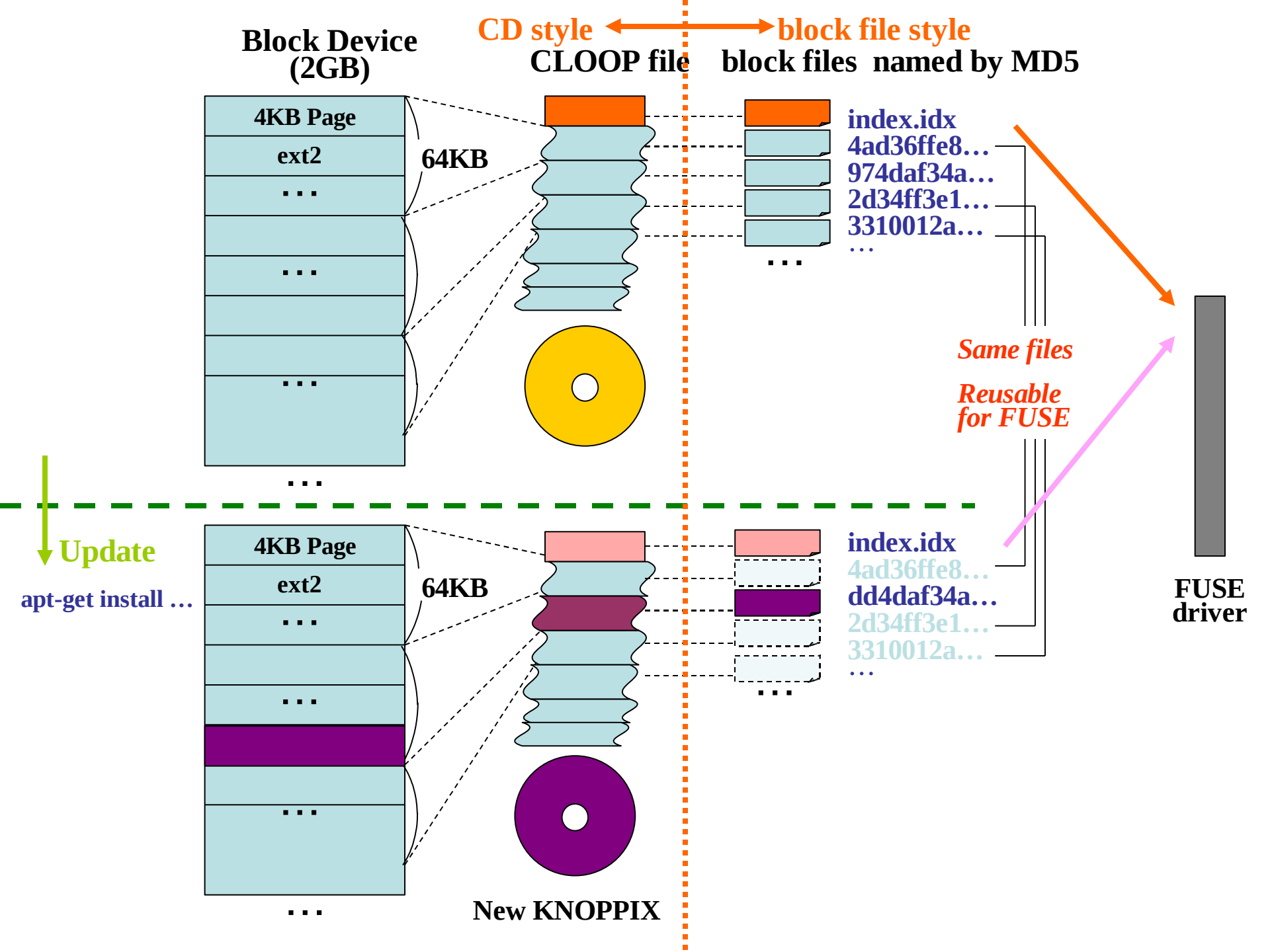




Index file has a location table of MD5 file names


block files
/tmp/blocks
index.idx
CLOOP file(700MB)
/cdrom/KNOPPIX
header
(location
data)
Loopback(2GB)
# losetup /dev/CLOOP
/cdrom/KNOPPIX
4KB Page
ext2
Access
FUSE
driver
CLOOP
driver
cache
64KB
decompress
extract
> Make a virtual device
/proc/fs/fuse/dev
> Mount it to
/cdrom as FUSE-FS
> Translate block files
to a virtual coop file
/cdrom/KNOPPIX


Contents SEVER
block files
block files
/tmp/blocks
index.idx
CLOOP file(700MB)
/cdrom/KNOPPIX
Loopback(2GB)
# losetup /dev/CLOOP
/cdrom/KNOPPIX
header
(location
data)
4KB Page
ext2
download
on demand
Access
FUSE
driver
CLOOP
driver
cache
64KB
decompress
extract
> Make a virtual device
/proc/fs/fuse/dev
> Mount it to
/cdrom as FUSE-FS
> Translate block files
to a virtual coop file
/cdrom/KNOPPIX

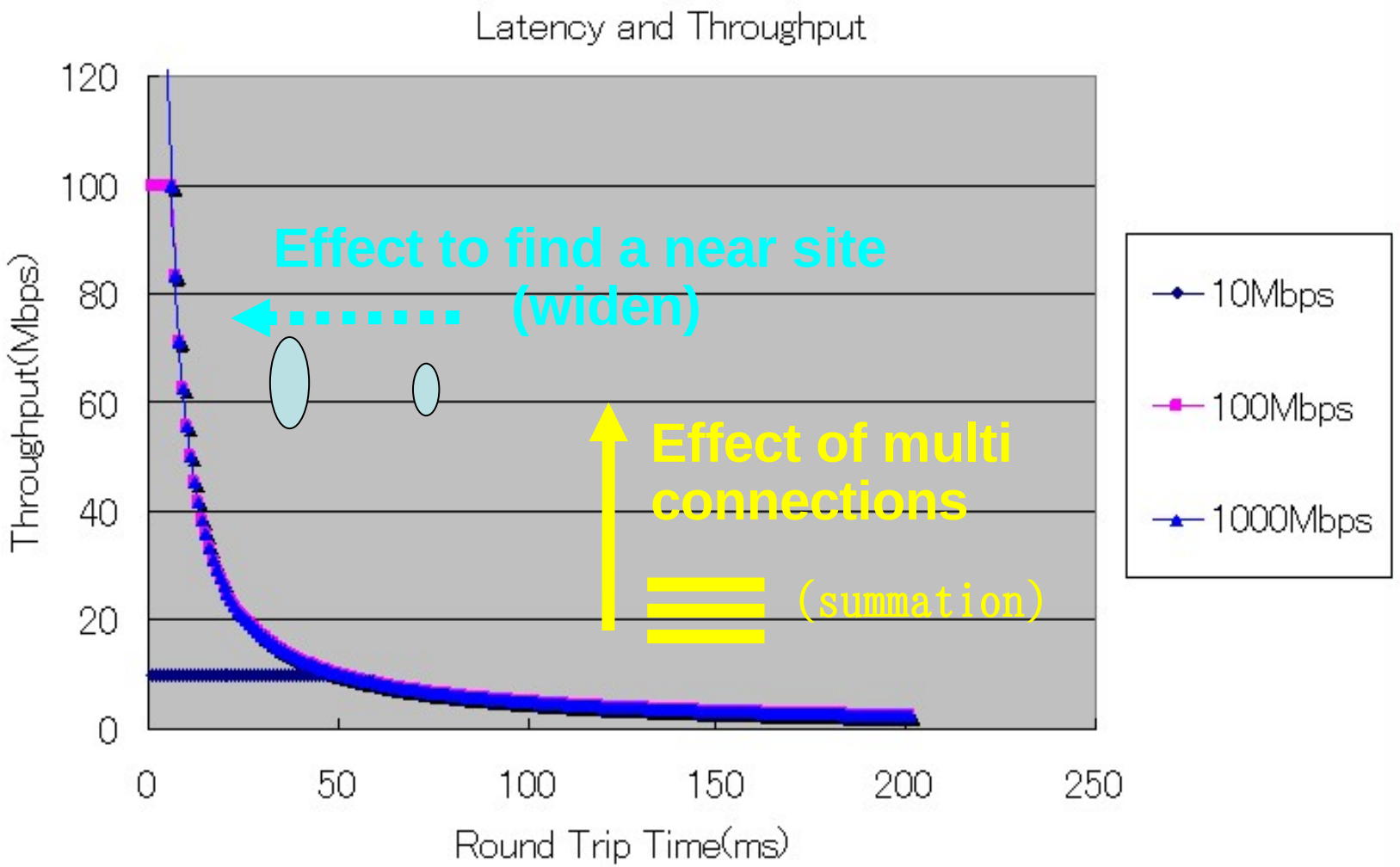

Block Device
(2GB)
CD style
block file style
CLOOP file
block files named by MD5
4KB Page
ext2
64KB
index.idx
4ad36ffe8…
974daf34a…
2d34ff3e1…
3310012a…
Same files
Reusable
for FUSE
FUSE
driver
Update
apt-get install …
dd4daf34a…
New KNOPPIX

Drawback of HTTP-FUSE CLOOP

- Vulnerable for network latency
- Fragmentation caused by Mismatch of block size between file system and block device.

Drawback of HTTP-FUSE CLOOP (Network Latency)

- Virtual block device can't accept multiple read requests.
  - HTTP-FUSE CLOOP has to download block files on demand.
  - Ext2 -> fuse-CLOOP -> HTTP(Intetnet) -> Block File
  - It makes vulnerable for network latency and become narrow throughput.



TCP window size is 64KB and 1 connection .

10Mbps : x < 50 y=10, x >= 50, y= (50/x)*10

100Mbps: x < 5 y=100, x>=5, y=(5/x)*100

1000Mbps: x< 0.5 y=1000, x>=0.5, y=(0.5/x)*1000

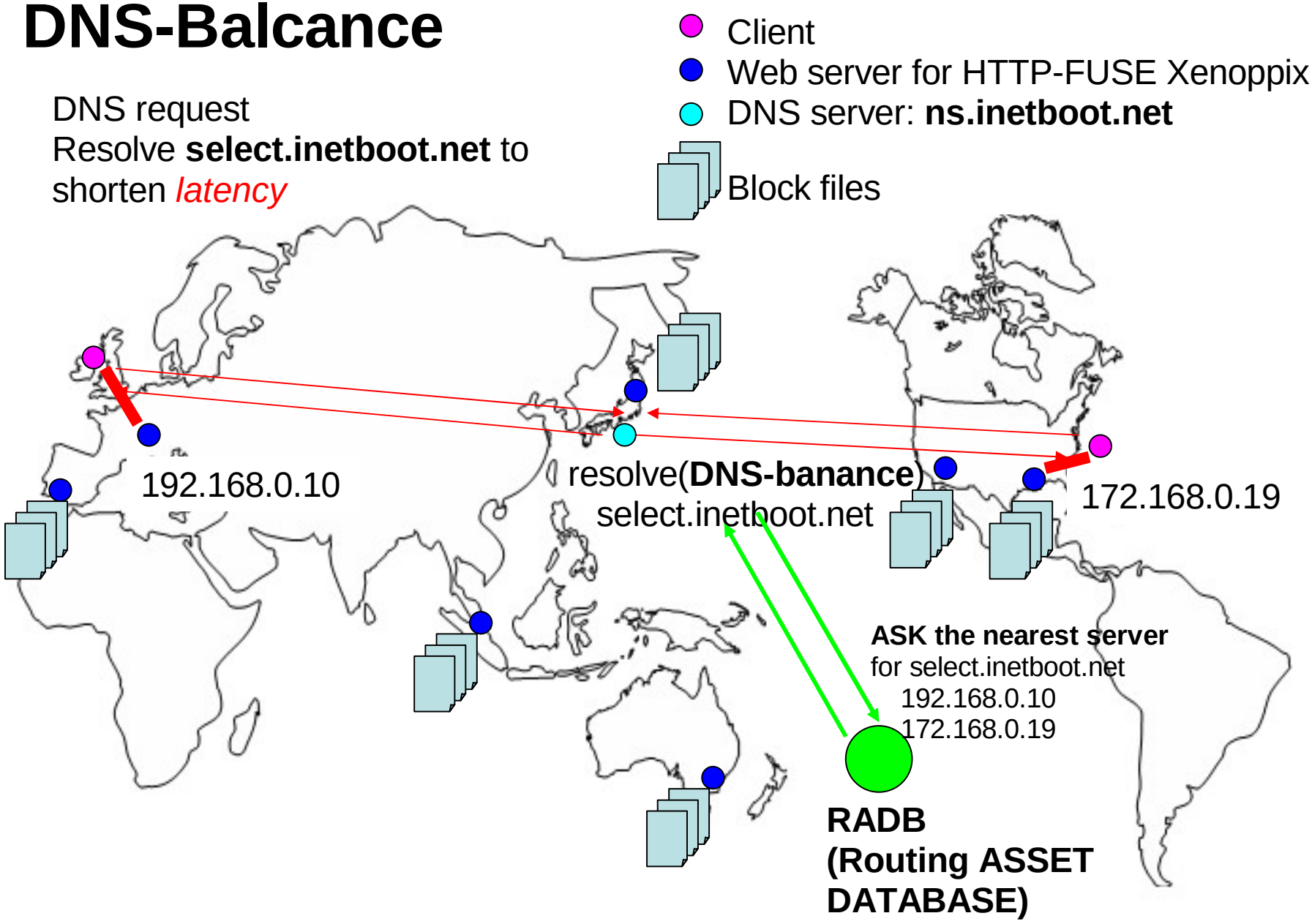
Current HTTP sites

- Web Hosting Service is reasonable.
  - 5GB/mount from 10$

Care for network latency (Find near download site)

- Client Side Solution: netselect
  - netselect measures the latency of candidates servers and finds nearest one.
    - http://www.worldvisions.ca/~apenwarr/netselect/
- Server Side Solution: DNS-Balance
  - DNS-Balance is a kind of name resolver which suggests near mirror servers with routing information offered by RADB.net
    - http://openlab.jp/dns_balance/dns_balance.html

DNS-Balcance

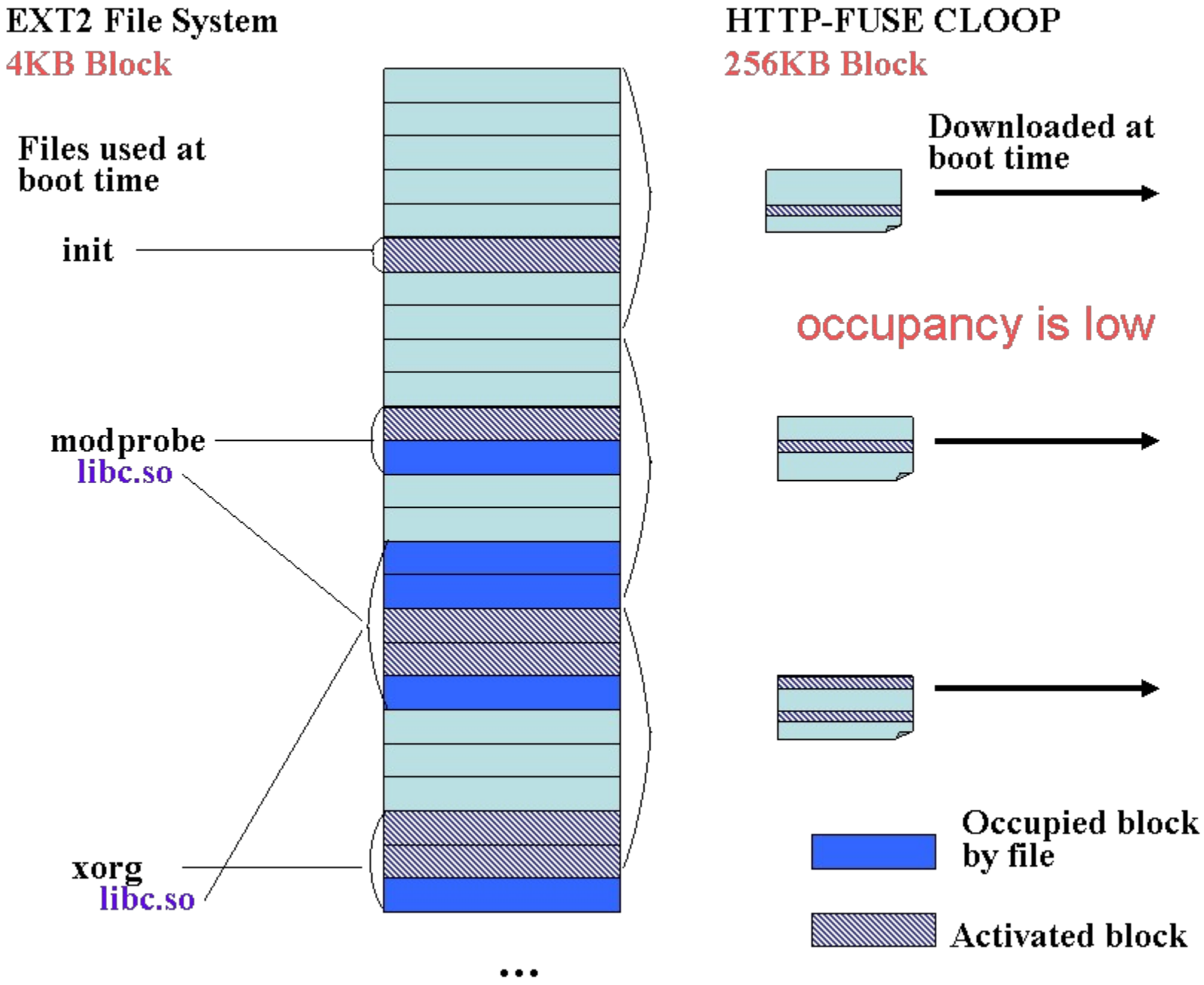


Care for network latency (Multiple download connections)

- DLAHEAD (DownLoad AHEAD)
  - Take a profile of downloaded block files at boot time.
  - The block files are downloaded in advance with extra download connections.
  - Effective and easy to implement
- Software RAID using MD (Developing)
  - Striping virtual block device

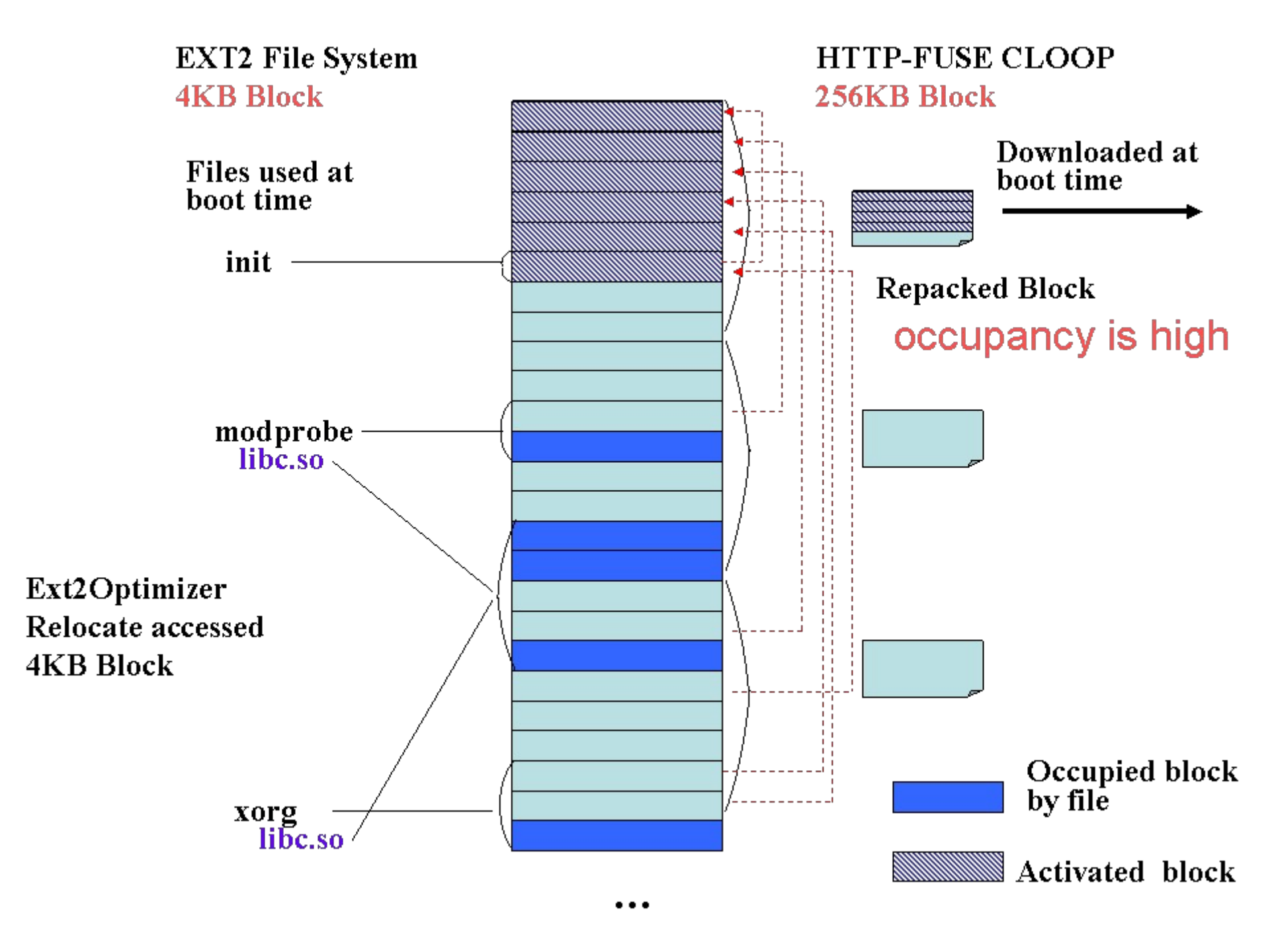
Drawback of HTTP-FUSE CLOOP (Fragmentation)

- Caused by block size mismatch
  - HTTP-FUSE CLOOP 256KB
  - File System (ext2) 4KB


EXT2 File System
4KB Block
Files used at boot time
init
modprobe
libc.so
xorg
libc.so
...
HTTP-FUSE CLOOP
256KB Block
Downloaded at boot time
occupancy is low
Occupied block by file
Activated block

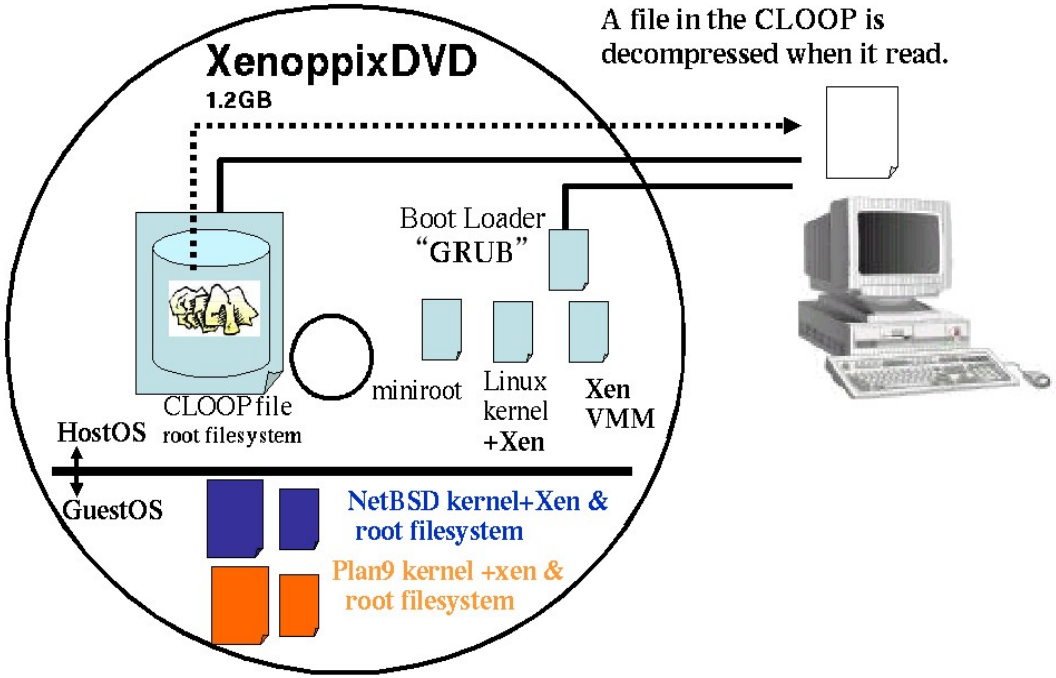
Care for fragmentation

- We developed “ext2optimizer” to repack ext2 data block.
  - Take a profile of *activated data blocks* on ext2.
  - Relocate the data blocks to be packed in fewer block files.


EXT2 File System
4KB Block
HTTP-FUSE CLOOP
256KB Block
Files used at boot time
Downloaded at boot time
init
Repacked Block
occupancy is high
modprobe
libc.so
Ext2Optimizer
Relocate accessed
4KB Block
xorg
libc.so
Occupied block by file
Activated block
...

Current Status

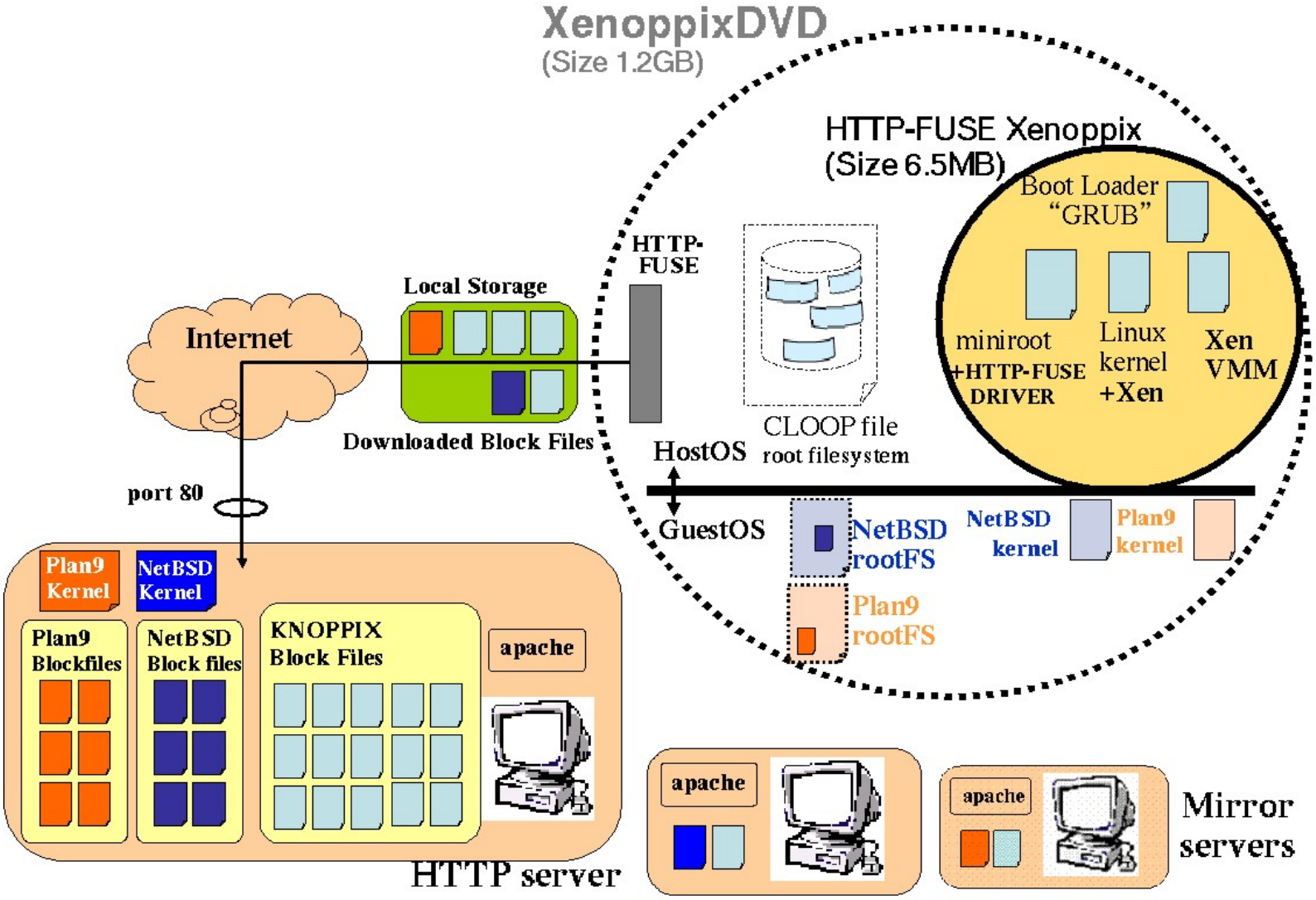
- Current HTTP-FUSE Xenoppix is based on Xenoppix 20051212.
  - KNOPPIX 4.0.2 (kernel 2.6.12) as Dom0
  - Xen 2.0.6
    - Para-virtualized OS: Plan9, NetBSD as DomU



Contents of Xenoppix DVD

Host OS (Domain0)	Linux	Xen VMM 2.0.6	0.12MB
		kernel 2.6.12 with Xen patch	1.3MB
		mini root	0.89MB
		Root File System (CLOOP)	870MB
Guest OS (DomainU)	NetBSD	kernel with Xen patch	1.9MB
		Disk Image (LOOP)	387MB
	Plan9	kernel with Xen patch	1.7MB
		Disk Image (LOOP)	337MB

Apply HTTP-FUSE to Xenoppix



Block files of HTTP-FUSE Xenoppix

Original loopback file	Number of block files	Size of block file	Amount of files
Domain0 (KNOPPIX) 680MB,CLOOP	7,483	Max: 262,230 Min: 277 Ave: 94,740	680MB
DomainU (NetBSD) 387MB,LOOP	1,559	Max: 253,977 Min: 277 Ave: 86,642	130MB
DomainU (Plan9) 337MB, LOOP	1,346	Max: 262,230 Min: 277 Ave: 73,161	94MB

Performance

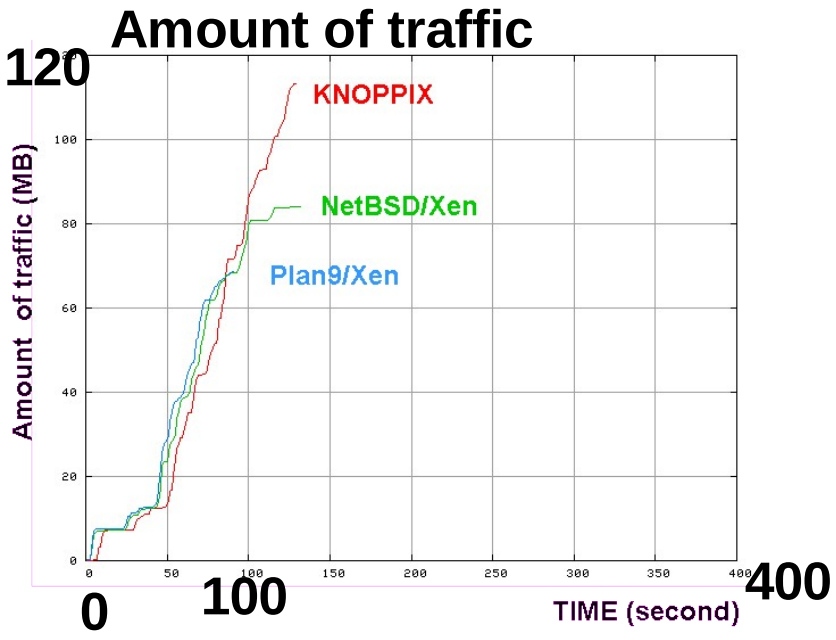
- Machines Environment
  - Server: Dell PowerEdge 1600SC (Xeon 2.8Ghz, 1G NIC, 4GB Memory)
    - Apache2
  - Client: IBM ThinkPAD T23 (PenIII 1GHz, 100M NIC, 1GB Memory)
- Measure Boot time of
  - KNOPPIX as Dom0 (Dom0+X+ KDE)
  - Pan9 as DomU (Dom0+X+ Plan9onVNC)
  - NetBSD as DomU (Dom0+X+ NetBSDonVNC)
- DLAHEAD is targeted for Dom0 and X because to cut extra download on each case.

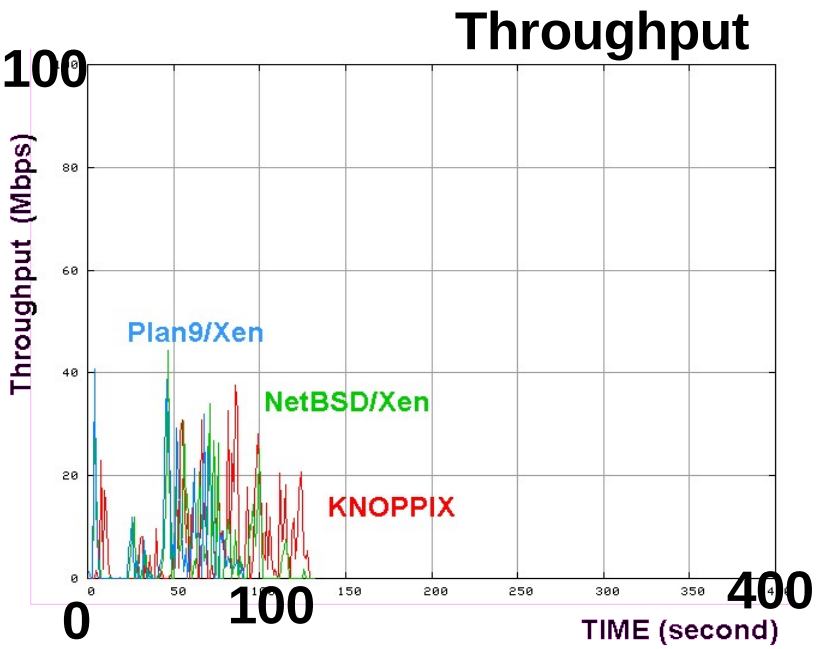
Boot Time

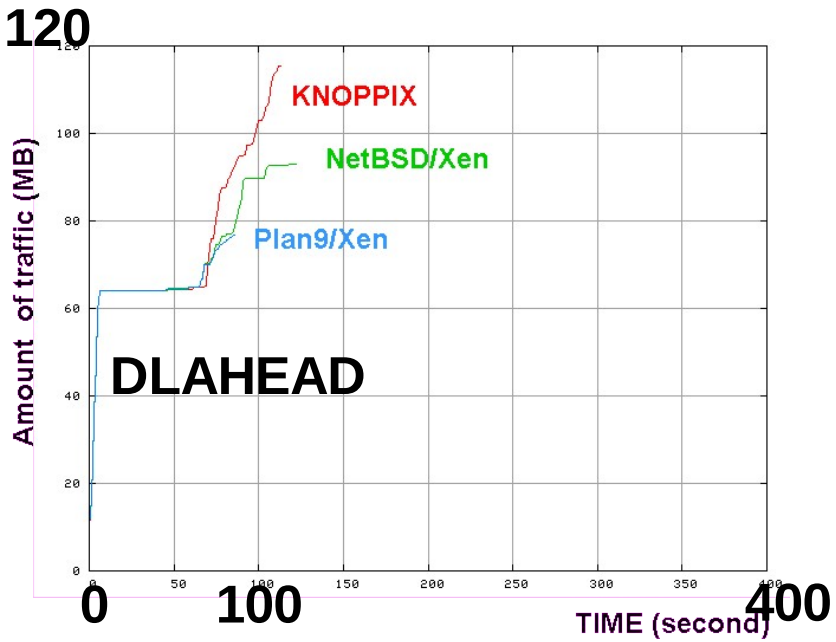
	Xenopix DVD	LAN Environment (No Latency)		Internet Environment (100msec latency)	
			DLAHEAD		DHALEAD
KNOPPIX	184	173 94%	157(16, 9%) 85%	432 235%	282(150, 35%) 153%
NetBSD on Xen	162	176 108%	166(10, 6%) 102%	384 237%	231(153, 40%) 143%
Plan9 on Xen	127	135 106%	130(5, 4%) 102%	340 268%	200(140, 41%) 157%

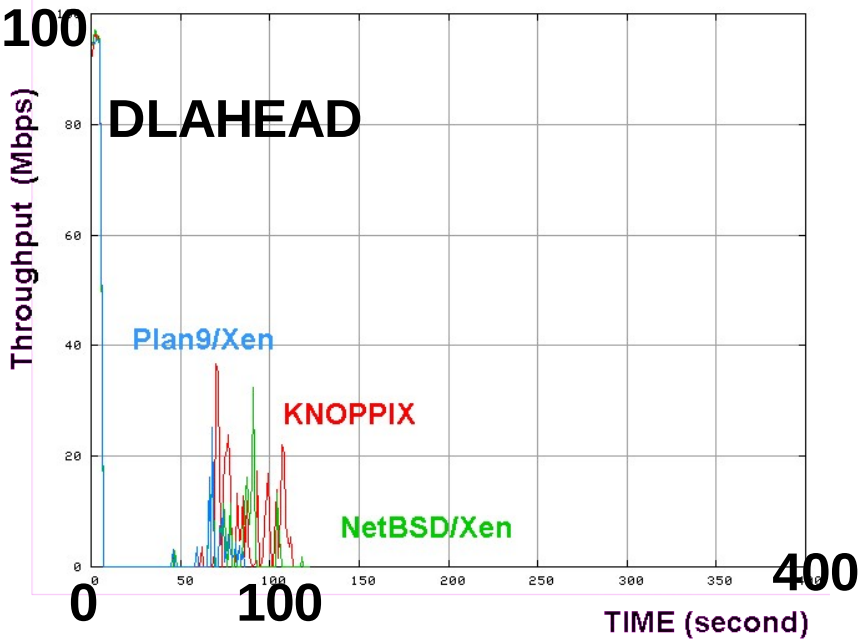
Boot Time(Sec). Upper part shows boot time and lower part shows percentage compared to boot time of Xenoppix DVD. The value in parenthesis shows time and percent shortened by DLAHEAD.

LAN Environment

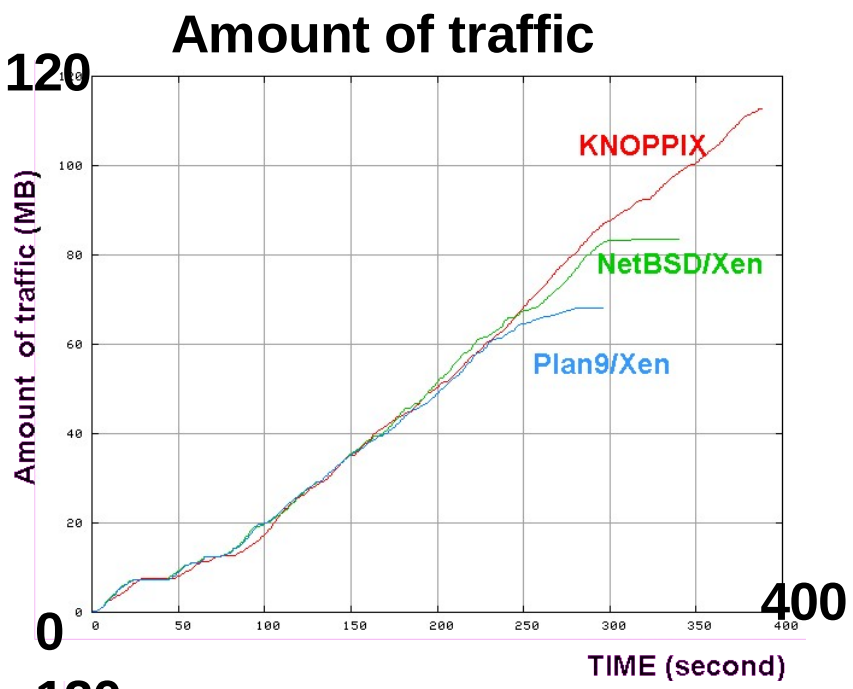


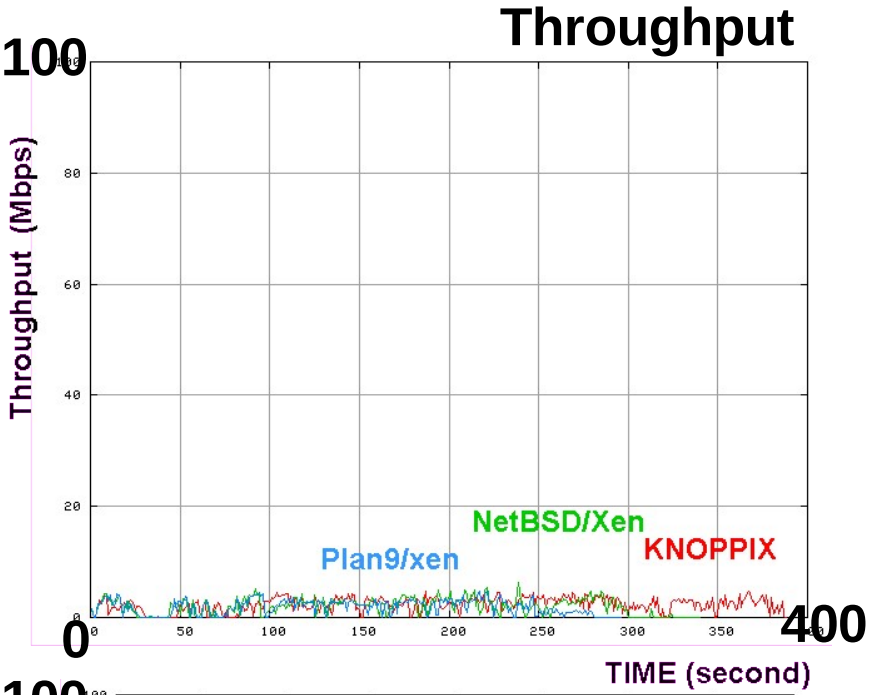


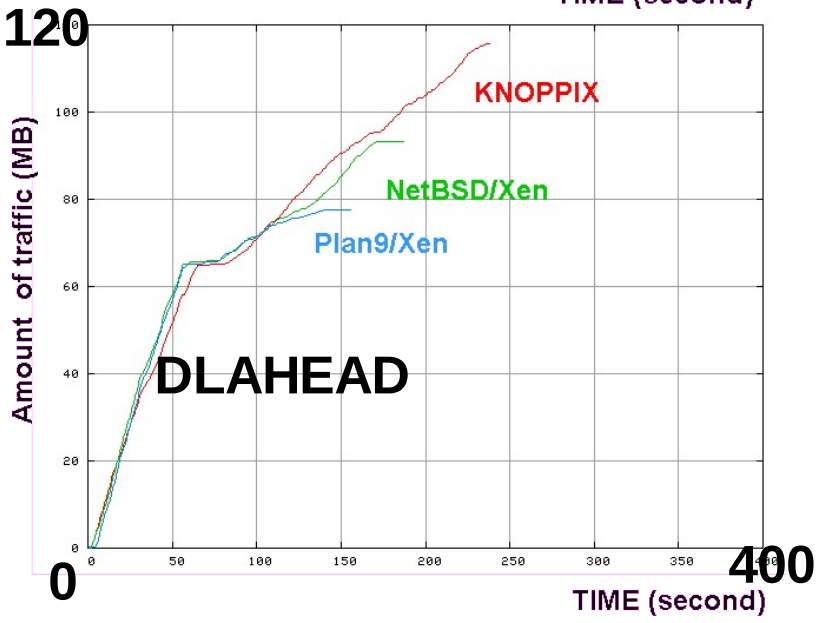


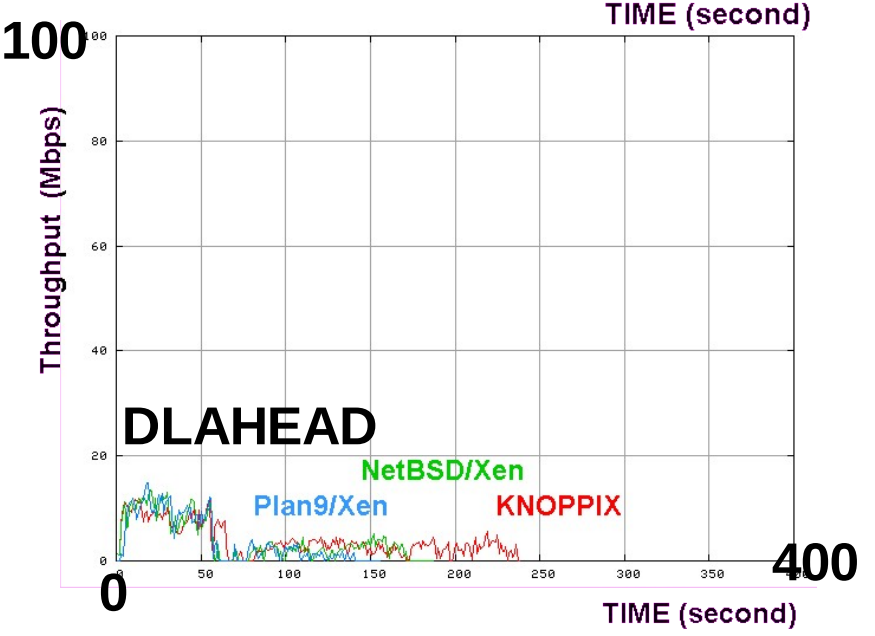


Internet Environment (100msec latency)









Future work

- FAST BOOT
  - Software RAID, Ext2Optimizer for applications
- Runtime Block File Creation
  - It allows Customization of HTTP-FUSE Xenoppix on a client
- Trusted Boot
  - Trusted GRUB + TPM module
  - Block files are measured by their MD5 file name.
- Live migration
- More OSes (OpenSolaris, MINIX, etc)

Discussion

- VMWare?
- Most important thing for Internet Thin Client is *periodic secure update*.
  - Current update is not so frequently.
- Another important thing is to *keep OLD OS image*.
  - Unfortunately Web disk space is limited.

Conclusions

- We wanted to develop Virtual Boot Loader for Internet Thin Client and developed HTTP-FUSE CLLOP.
  - Current HTTP-FUSE Xenoppix can boot KNOPPIX, Plan9 and NetBSD with 6MB bootable CD-ROM.
- We will adapt Xen3 to HTTP-FUSE Xenopix to use full-virtualization. It will add more bootable OSes.
- DEMO?