Contents
Ch 0 Intro
Welcome to the page
About ...
Syllabus ...
Homework ...
Discussions ...
Exam info ...
Lecture slide guides
Class info
Ch 0 OS background
About the course ...
Prerequisites ...
Things you need
Policy ...

Devices

By Dr. Jon Turner (jonathan)
Written in c, Word (Updated March 10, 2011)

> We live in a world of things, and our only connection with them is that we know how to manipulate them or to consume them.
> — Clark Fussen (1930–1988), The Tank Monkey, chapter 1

Introduction

Ultimately, a computer is useful only if it can access devices and communicate with people and objects. Common devices include keyboards, displays, disks, mice, printers, scanners, and keyboards. They also include devices that connect to other computers, such as network interfaces and modems. Even the CPU and memory can be considered devices, since they are managed by the OS and accessed by user programs. The OS allows us to treat a convenient interface to the device and to hide device-specific details within the driver (for instance, the kernel outside of the device drivers shouldn't care whether your computer has a Microsoft mouse or a Logitech mouse). A device driver accesses the device controller to control the device itself. A device controller is the hardware (and/or firmware) that controls the operation of the device.

Device drivers implement mechanisms, not policy. A mechanism specifies the ways of interacting with the device. A policy defines who can access the device and in what way they can control it. The policy will be handled at higher layers in the operating system and the device driver has only to implement the operations needed to interface with the device.

Very broadly, the device component of an operating system contains three components:

1. Generic device driver code: these are abstract interfaces for devices and code for managing instances of and references to devices.
2. Drivers for specific devices: this is code written for interfacing with specific devices. This group also includes the code for bus drivers, such as discovering components on a USB bus and sending and receiving data on the bus.
3. The buffer cache: this manages the caching and I/O to block-oriented devices such as disks.

Modularity

A goal in managing devices is that I/O operations should have a consistent interface for different devices. This is known as device independence. Device independence implies that higher levels of the operating system, including user programs, can access devices in a uniform manner. For example, a program that reads from a USB memory stick should also be able to read from a floppy disk. Similarly, a program if the computer is to go to a USB memory key rather than to a disk. This isn't always possible since devices may have very distinct capabilities but it is desirable to maximize the number of common interfaces.

Drivers are modular and may often be dynamically loaded into or removed from the kernel at any time. Under the Linux operating system, one can use the lsmod command to add a module and the rmmod command to remove it. A dynamic loader finds unresolved symbols within the module in the symbol table of the running kernel.

Each module has a function, called module_init, that the kernel calls when the module is first loaded to allow the module initialize itself and register itself. Similarly, a module can have a function called cleanup_module that is called just before the module is removed from the kernel. The kernel also keeps a use count for each device in use. A module can be removed only if the use count is 0.

While we'll focus on the kernel anymore here, I/O is important in ints that users drivers can be implemented at the user level. For example, the X windows server is a user level process that offers graphic resources to it clients. It runs on top of a normal level device driver that gives it access to the frame buffer. Client applications connect to this user level server to perform communications with the device.

Categories of I/O

There are four broad categories of I/O:

- block devices
- character devices
- file systems
- network (sockets)

Of these, character and block devices will appear in the file system name space (we'll see how shortly).

Block devices

Block devices provide unnamed access to the underlying hardware. Block devices are devices that we can have a file system. They support addressable block-oriented I/O (e.g., read block number, write block number) and allow programs to read or write any block on the device. That is, you can read the same block number over and over again and get the same contents. Examples of block devices are: USB memory keys, disks, CDs, and DVDs.

Because blocks are addressable and persistent, block I/O lends itself to caching frequently used blocks in memory. The buffer cache is a pool of kernel memory that is allocated to hold frequently used blocks from block devices. By using the buffer cache, we can minimize the number of I/O requests that actually require a device I/O operation. the buffer cache also allows applications to read/write from/to the device as a stream of bytes or arbitrary-sized blocks. For example, suppose that a disk has a minimum block size of 512 bytes (as most disks do). If you want to modify 10 bytes on a disk, you need to read the entire 512 byte block (or two blocks if the 10 bytes straddle a block boundary), store the data somewhere in the buffer cache, modify the 10 bytes, and write the modified block(s) out again.

File system

The file system determines how data is organized on a block device in order to present higher-level software with a hierarchy of directories and files, with access permissions for both.

A file system is not a device driver. It's a software driver within the operating system that maps between low-level and high-level data structures.

Network device

The network access device is a packet, rather than stream, oriented device. It is not visible in the file system but is accessible through the socket interface. We will examine this when we look at networking. The underlying network device may be either hardware (e.g., an ethernet controller) or software (e.g., a loopback driver). Higher level networking software will be agnostic to this.

Character device

Character devices provide unstructured access to the underlying hardware. In some ways, they are a catch-all for anything that is not a block or network device. Examples of character devices include:

- Traditional character-stream devices, such as a terminal multiplexer, modem, printer, scanner, or mouse.
- The frame buffer. This is considered a character device even though it has its own buffer management policies and custom interfaces
- Sound devices, I²C bus controllers, and other bus interfaces.

In addition to supporting "character" devices, the character device drivers are also used with block devices to bypass the buffer cache and provide raw I/O operations directly to the user's program address space. For example, this feature is used for programs that dump and check file systems. When accessing block devices via this interface, the program has to abide by the restrictions of the hardware and ensure that the I/O is a multiple of the device's block size.

For terminal devices (virtual terminals as well as old-style serial port terminals), it may make sense to process input from these devices in one of two modes:

Raw mode (cooked)
: Input characters are entered by the system as they are received (e.g. speed) and collected in a buffer until a newline is received. Then the entire line is made available to the process that is waiting on a read operation. The user may rerun/replay once in this mode and use special characters to start/stop output, suspend a process, or delete the current input line.

raw mode
: Characters are made available to the process as soon as they are received. This is the only mode that makes sense for a non-terminal device such as a mouse. For a terminal, this mode is needed if a screen editor wishes to process every keystroke that is gen.

These modes were created in the era of dumb terminals so provide users with rudimentary line-buffering capabilities without having a real system call return as soon as a single character was typed. Even shells tend to support in place command editing and must revert to raw mode. The time that you mostly come in contact with a line discipline is whenever you use a terminal/window and run programs that request like, entered input (e.g., via Open).

Blocking, Non-Blocking, Asynchronous I/O, and Buffering

The system call interface allows us to either wait for data to be:

Blocking I/O
: means that the user process will block (wait) until the I/O is complete. If you're reading 1,500 bytes from a device, the read will block until the data is ready. The solaris system call to often used with non-blocking I/O to check whether data is ready for a device.

Non-blocking I/O
: means that an I/O system will would not put the process to sleep waiting on I/O. Instead, the process would get a status stating that the I/O is not ready. If you're reading 1,500 bytes from a device and only 12 bytes are ready, your read request will return only the 12 bytes.

Asynchronous I/O
: is similar to non-blocking I/O in that the process would not be put to sleep. It's a two-stage operation: a request for a read or write operation is first made, and returns immediately. Later on, the process is notified that the operation is complete.

In the kernel, the buffer cache provides an asynchronous I/O interface to block devices. A write to the buffer cache does not necessarily translate to an immediate write to the disk. This is good because it reduces the amount of I/O needed to the device in case that same block gets modified again in the future. However, if the system crashes in a dive off spontaneously before all modified blocks are written, that data is then lost. To minimize data loss, the operating system will periodically force a flush of data that's cached in the buffer cache. On BSD systems, a user process, update, calls the sync system call every 30 seconds to flush data. On Linux, the operation is done by a process called kupdate; the kernel tunables daemon.

Buffered I/O
: is where the kernel makes a copy of the data on a write operation from the user process to a temporary buffer. Similarly, for a read operation, the kernel reads data into a buffer first and then copies the requested data to the user process.

Buffering is clearly an overhead since it requires copying blocks of data. However, it has a number of benefits:

- It deals with block boundaries. Data may be read in chunks safely you want to read it as a byte stream. A buffer provides a place to store the chunks while giving you the stream of bytes. This specific example is called a data checker, since it provides a way to get a constant bandwidth stream with bursty input.
- Buffering allows user data to be modified without affecting the data that is actually being written to or read from the device. This is especially critical for asynchronous write operations. If you modify a buffer after you requested that a write but before the data was copied to the device, the device gets the modified data. By buffering, you can provide a point-in-time snapshot of the data.
- Buffering can improve performance by caching data in memory for the purpose of the buffer cache. It's often the case that the same blocks will be accessed from the disk over and over again. If they are cached then we don't have to perform the disk operation — and we save a lot of time.
- Buffers help with alignment. As we mentioned in the above examples, if you want to read 10 bytes but a disk forces you to read at least 512 bytes, you need to have a place to store the extra data. If you want to write 10 bytes, you'll need to read 512 bytes into a buffer, modify the 10 bytes, and then write the results back to the disk.

Devices as files

On many operating systems (POSIX), ones such as Linux, UNIX, BSD, OS X, as well as NT, die-hard Windows systems), devices are presented via the File I/O interface. Since we read and write to files, why not use the same mechanisms for devices? Certain operations may not make sense, such as a seek function on a character device, but all functions don't need to be fully supported to present the abstract interface.

On Linux, each character device driver initializes itself upon being loaded, registers a cdev structure that describes the device, and implements a set of functions in a file_operations structure:

```
struct file_operations {
    struct module *owner;
    loff_t (*llseek) (struct file *, loff_t, int);
    ssize_t (*read) (struct file *, char __user *, size_t, loff_t *);
    ssize_t (*write) (struct file *, const char __user *, size_t, loff_t *);
    ssize_t (*aio_read) (struct kiocb *, const struct iovec *, unsigned long, loff_t);
    ssize_t (*aio_write) (struct kiocb *, const struct iovec *, unsigned long, loff_t);
    int (*readdir) (struct file *, void *, filldir_t);
    unsigned int (*poll) (struct file *, struct poll_table_struct *);
    int (*ioctl) (struct inode *, struct file *, unsigned int, unsigned long);
    int (*mmap) (struct file *, struct vm_area_struct *);
    int (*open) (struct inode *, struct file *);
    int (*flush) (struct file *, fl_owner_t id);
    int (*release) (struct inode *, struct file *);
    int (*fsync) (struct file *, struct dentry *, int datasync);
    int (*aio_fsync) (struct kiocb *, int datasync);
    int (*fasync) (int, struct file *, int);
    ...
};
```

These functions are called by the kernel to transform generic input/output requests to device-specific operations. The Tenlo. like as file" interface abstraction applies only to character devices.

Each device driver implements a fixed set of entry points. It defines whether the device has a block or character interface (or multiple interfaces). Block device interfaces appear in a block device table. Character device interfaces appear in a character device table.

Block devices are used to implement file systems and are accessed via the operating system's buffer cache. A block device driver initializes and registers a gendisk (generic disk) structure and implements functions defined in block_device_operations.

A network device driver initializes and registers a net_device (network device) structure and implements functions defined in net_device_ops.

Locating devices

Within the kernel, a device is identified by two numbers:

Major number
: identifies the specific device driver and is an index into the device table (either the block or character table).

Minor device number
: This number is interpreted within the device driver. It identifies which, of possibly several, specific devices the request is for. For example, a major number may identify a SATA disk driver. A minor number will identify a specific disk on the SATA interface.

How do user programs locate devices? There are a couple of approaches:

1. Use an explicit namespace for devices that is unique from files. This is the approach taken by MS-DOS and supported by the Win32 API. Devices are given names such as \\.\A:, \\.\PhysicalDrive0, etc. The latter cannot be used as part of a file name and a device that hosts a file system will have the device name identified as part of the full pathname of the file (e.g., \\.\Volume{guid}\data.ogg).
2. Incorporate device names into the file system namespace. A device is a file with no contents (data) but whose metadata (descriptive information about a file) identifies it as a device, given the device type (block or character) and provides the major and minor numbers. On POSIX systems, devices may be located anywhere in the file system but are traditionally placed under the /dev directory. Thus special files can be created with the kernel system call or device command. When the file is opened, it is recognized as a device special file and high-level file operations are passed to the functions defined in the file_operations structure of the character device driver or to the common block device interface for block devices.

The Windows NT architecture (NT, 2000, Vista, Windows 7, etc) uses a similar concept, although the device names do not live on the disk file system. When a device driver is loaded, it is registered by name with the Object Manager, which is responsible for presenting a name space to applications. Device names have a hierarchical namespace that is maintained by the Object Manager and integrated with the namespace of files and other objects. For example, \Device\Serial0 identifies the first serial port. One snag is that the Win32 API requires the old, style MS-DOS device names. These names reside in the \?? directory in the Object Manager's namespace. The names are visible to Win32 programs and are symbolic links to the Windows NT device names.

In summary, the Big Idea was to use the file system interface as an abstract interface for both file and device I/O.

Device driver structure

Character (and raw block) devices include these entry points and must implement the file_operations functions mentioned earlier:

- open: open the device
- close: close the device
- ioctl: device I/O control operation
- mmap: map the device offset into a memory location
- flush: flush I/O buffers
- poll: control/use the device
- select: poll the device for I/O readiness
- llseek: change device offset
- aio: driver-assisted operation

Block devices must implement these functions:

- open: prepare for I/O – called for each open system call on a block device (e.g. when a file system is mounted)
- request: read or write blocks – this is the main function associated with a block device. It takes a request and starts/requests to the buffer cache. If the block is not there, then it requests it from the device. The strategy routine finds a list of I/O requests (nreadwr, defined in the drv_ow_physio_dly_initiate_request) it implements the code to move data from the device to the block (read) and from the block to the device (write).
- flush: called after the final close using the device terminates.
- size: get partition size — the capacity of the device.

Execution contexts and Interrupt handling

When kernel code, such a device driver (or anything else, for that matter), is executed, it is done so in one of three contexts:

Interrupt context
: Interrupt context is invoked by the spontaneous change in the flow of execution when a hardware interrupt takes place. Any actions performed in this context cannot block because there is no process that requested the action that could be put to sleep and scheduled to wake up when an operation's is complete.

User context
: The system entering in user context is invoked by a user thread when it makes a system call (or generates a trap). For example, when the user calls the read system call, the wrapper function for the system call executes a trap instruction that jumps to a well-known entry point in the kernel and switches the processor to run in kernel (supervisor) mode. However, the memory map is still unchanged and the kernel is aware that the thread from which the request came (on Linux, the global variable current contains a pointer to current task_struct that identifies the calling thread). Because of this, the thread (or process) may block on a semaphore, I/O, or copy operation. In this sense, we can think of the thread just happens to be executing kernel code.

Kernel context
: The kernel itself may schedule worker threads just like it schedules user processes. For all practical purposes, these no different than any thread that are associated with user processes: they just happen to run in kernel mode. Because they are scheduled entities, even though they have no relation to any user thread, they may block on I/O or any other operation that would cause a thread to block. When the condition that made the thread block does not exist anymore, the thread is put on the ready queue and rescheduled by the scheduler. Note: system call processing takes place in kernel mode in the user context; not in the kernel context.

Device drivers do not get hooked directly when a system interrupt occurs. Instead, they register themselves with the kernel's interrupt handler when they initialize themselves. This is known as a callback. Hooking is a general term for allowing software to intercept messages or events. It also allows multiple device drivers to share the same interrupt if necessary. The interrupt handler performs the following functions:

1. Save all registers
2. Set up the correct segments
3. Get interrupt service routine in driver with the appropriate unit number (ID of device that generated the interrupt)
4. Restore registers
5. Return from interrupt

The process of responding to interrupts or interrupt (or, in the general case, any message or event flow) is known as hooking. By separating interrupt handling from the driver, the driver does not have to deal with saving or restoring registers or dealing with any specifics of the interrupt handling mechanism. Moreover, interrupts may be chained and the operating system can keep global statistics on interrupts.

Because an interrupt is, by its very nature, an interruption of the current workflow on the system, and the interrupt service routines in RAM quickly. In many systems, interrupts are disabled while an interrupt is being serviced, so the tasks more outgoing to finish quickly. Also, since the interrupt context cannot block, we have to make sure that the interrupt service routine does not perform any potentially blocking actions, such as waiting on I/O or a semaphore.

To ensure speedy interrupt servicing, interrupt handling is generally split into two parts:

- The top half is the part that is executed by the interrupt handler. It performs the device's interrupt handler (in Linux, via request_irq) and is called whenever an interrupt is detected. Its goal is to finish as quickly as possible. It saves any data in a buffer or queue (e.g., grabs an incoming packet from the ethernet card), schedules a bottom half and exits.
- The bottom half of the interrupt handler (also known as the work queue or tasklet/thread) is the part that is scheduled by the top half for later execution. This is where the real work of the interrupt servicing is done. The bottom half is run in the kernel context and is scheduled by the kernel's process scheduler. Because of this, interrupts are enabled and the thread can perform blocking operations. On Linux 2.6, the kernel provides: workitem and work queues for dispatching bottom half threads. Linux creates one work queue thread per processor (with names such as events/0 and events/1).

I/O Queues

One of the most efficient means of communication between the top and bottom halves of an interrupt handler. A device queue table keeps track of a list of devices and the current status of work device (e.g., idle or busy). Each device has a I/O queue attached to it. Whenever an I/O request is received, it is placed on a device's queue for processing.

When an input or output operation completes, the device controller generates an interrupt, which causes the interrupt handler to call the top half of the device driver's interrupt service routine. The interrupt service routine removes the appropriate request from the device's queue, notifies the requester that the request completed, and starts the next request from the queue.

Figure 1: I/O queues

Frameworks and a unified device/driver model

Kernel-like, the Linux kernel distinguishes between classes of device drivers sorted well. Drivers were primarily disk interfaces (block devices) or serial terminals (character devices). As more devices were added to machines, the character driver became the catch-all driver for any kind of device. Non-I/O related commands were implemented as custom commands that are called via the ioctl system call.

Frameworks were created to formalize a set of interfaces for drivers of specific types. For instance, The ALSA framework is a kernel-level API for audio. Instead of having each audio chipset create its own custom interfaces, they can all be written to conform to ALSA APIs. Frameworks define common parts and common interfaces for the same types of devices (e.g., ALSA sets, TTY serial, SCSI core, frameBuffer devices). All these devices are still seen as normal devices to users and treated as such by the kernel. A framework just defines a common interface and a set of operations that the device must implement (e.g., a frameBuffer operations. ALSA audio operations).

For example, a frameBuffer driver has to implement functions defined in struct fb_ops, which is a set of frameBuffer-specific functions and include functions such as:

```
int (*fb_open)(), int (*fb_release)(), ssize_t (*fb_read)(), ssize_t (*fb_write)(), int (*fb_blank)(), int (*fb_pan_display)(), void (*fb_fillrect)(), void (*fb_copyarea)(), void (*fb_imageblit)(),
int (*ioctl)(), void (*fb_rotate)(), int (*fb_sync)(), int (*fb_get_caps)(), ...
```

It also has to allocate a fb_info structure with a call to framebuffer_alloc(), set the .fbops field to the operation structure, and register the framebuffer device with register_framebuffer().

The Linux 2.6 kernel also tried to unify the relationship between devices, drivers, and buses, with the unified device/driver model. A bus driver is a driver for interacting with each communication bus that supports multiple devices. Example bus drivers are USB, PCI, IPS, MMC, and I2C. It is responsible for:

- registering the bus type
- registering adapter (interface drivers) (USB controllers, SPI controllers, ...). These are devices that are capable of detecting and providing access to devices connected to the bus
- providing a driver interface (USB drivers, I²C drivers, SPI drivers)
- matching device drivers against discovered devices.

I/O Scheduling and Disks

From the 1960s through around 2010, the magnetic disk has been a crucial device in many computers. Much of the structure of the buffer cache and block drivers has been designed with the disk in mind. For high-storage density, random access, and low cost, it's still hard to beat disk device, although flash memory is making strong inroads.

A disk for data storage, is a disc-shaped piece of metal, typically 2.5 inches on laptops and 3.5 inches on larger systems (in the shown pace, 3–10 inches were common once). It is coated with magnetic material. Several disks are sometimes mounted on one spindle, An individual disk is known as a platter.

Data is recorded on the surface of a disk on a series of concentric tracks. Each track is divided into a set of fixed-length blocks called sectors (which are separated from each other by inter-record gaps). A sector typically stores from 512 bytes to several kilobytes of data. A hard disk or drive disk on a stack on the disk cycle typically at 7,200 rpm, but other common speeds are 5,400, 10,400, 10,800, and 15,000 rpm). There is one read/write head per recording surface. A cylinder is the set of tracks available vertically (each track is under its own head). Since there is only one set of heads per cylinder, the disk drive must choose on or seek to across cylinders. A disk is a random-access device meaning that you can request any sector to be read at any time, but the access time for the data depends on several factors:

1. How much does the disk arm have to move to get to the requested cylinder (seek time)?
2. How much time does it take for the desired sector to rotate under the head (rotational latency)?
3. How fast can data be transferred as the sector passes underneath the head (transfer rate)? For a 4,200 RPM disk, a disk rate 1,000 MB/sec is a typical sustained transfer rate.

The time for a disk to rotate to the right position may seem insignificant since it's spinning at 7,200 rpm, but let's consider what this time means to the CPU. If a disk spins at 7,200 rpm, there is spins at 120 revolutions per second, taking 8.3 milliseconds to rotate once. The average latency is half of time, or 4.2 milliseconds. A 3.2 GHz Intel Core 2 processor can perform approximately 76,160 million instructions per second. Therefore, the CPU can perform approximately 318 million instructions in the time that it takes the disk to spin so that the proper sector will be positioned under the disk read/write head. Seek times are even longer. As far as processors are concerned, disks are — and always have been — slow.

Disk Scheduling

As with other slow devices, requests for disk service are stored in a queue. By the time a disk has completed one data transfer operation there may be several more requests in the queue. The question then arises is that of which order should these requests be processed. There are several algorithms that can be used to sort disk requests.

First come, first served

The simplest thing to do is to service each disk request on the queue in the order that it arrived. The drawback is that this scheme does not try to optimize seek time. Consider having one request for data on cylinder 20, the next request for data on cylinder 1000, and the third request for data on cylinder 21. The disk head will have to move across the disk surface to service the request on cylinder 1000 before it can service the request on cylinder 21. Servicing the requests in the order 20, 21, 1000 would minimize the amount of time that the disk head has to move. This first come, first served algorithm does not take this into consideration and will seek to cylinder 20, then to cylinder 1000, then to 21, and then back to 1000. We find ourselves spending most of our time seeking back and forth rather than transferring data.

Elevator algorithm (SCAN)

In the elevator algorithm (also called SCAN), we keep the disk head moving in inner cylinders, processing queued requests in order of increasing cylinder numbers until there are no more requests in that direction. The head will then start seeking to outer cylinders, processing any queued requests in order of decreasing cylinder numbers until the disk reaches the highest track number. After that we reverse again. This behavior is analogous to that of an elevator.

Figure 2: Elevator algorithm

With this algorithm, we need to know the current head position and direction of the disk head (figure 2). LOS is scheduled in the sequence of the current direction. For example, if we're at track 50 and the last request we serviced was track 45, then we would next service track 43, then 39, and so on until we reach the track with the lowest request in that direction. We then reverse direction and service the next request in the opposite direction. Any new requests that are behind the current position are queued for servicing on the next sweep in the opposite direction. After the head reaches the lowest track number, it reverses direction again. We service the next request in order of increasing track number, moving toward the highest track number. Any new requests that are behind the current position are queued for servicing on the next sweep in the opposite direction.

LOOK

LOOK is a common variant of the elevator algorithm. Instead of seeking to the very end unless there is a request at the end, we only seek as far as the highest request in the current direction. Then we start seeking in the opposite direction, processing any queued requests in order of decreasing cylinder numbers until there are no more requests in that direction. After that we reverse again. This behavior is analogous to that of an elevator that doesn't go to the top floor unless there is a request to go there.

Circular SCAN (C-SCAN)

With SCAN and LOOK, tracks in the middle of the disk get a statistical advantage in decreased average latency. In an attempt to provide more uniform wait times, the Circular SCAN algorithm only schedules requests when the head is moving in one direction. Once the end of the disk is reached, the head seeks to the beginning without servicing any I/O.

Circular-LOOK (C-LOOK)

This is like C-SCAN but instead of seeking to the start of the disk, we seek to the lowest track with scheduled I/O.

Shortest Seek Time First (SSTF)

To minimize average response time for processes, we were able to use a shortest job first scheduling algorithm. We just attempt a similar strategy with disk scheduling: the queue of current requests can be sorted by cylinder number and requests can be scheduled in order of proximity to the current head position. This is known as shortest seek time first (SSTF). Unfortunately, this algorithm can lead to starvation. Requests that come from remote parts/peripherals or "below" requests may keep entering in and be scheduled ahead of the less desirable requests. SSF also may create a lot of disk seeking and why on the head may seek back and forth with ever-wider swings as it services the more distant requests.

Examples: Linux disk scheduling options

Linux 2.6 gives you a choice of using one of several disk-scheduling algorithms. Let's take brief look at what they offer.

Completely Fair Queuing (CFQ)
: This is the default scheduler on most distributions. Its goal is to distribute I/O equally among all I/O requests. It supports two types of requests. Synchronous requests go to per- process queues with time-slices allocated per queue. Asynchronous requests are batched into queues by priority levels.

Deadline
: Each request has a deadline. Requests are serviced using a C-SCAN algorithm. However, if a deadline for any I/O operation is threatened, skip to that request immediately. This approach helps with real-time performance and gives priority to real-time processes. Otherwise, it's just a C-SCAN algorithm.

NOOP
: This is a super-simple scheduler. It's just a FIFO queue. This approach has minimal CPU overhead but assumes that the block device is intelligent and capable of resolving/ordering operations internally.

Anticipatory
: Anticipatory scheduling introduces a delay before dispatching I/O to try to aggregate and/or reorder requests to improve locality and reduce disk seek. The hope is that, with the delay, there's a high likelihood of additional nearby I/O requests coming in. After the delay, the scheduler will then continue to use a C-SCAN algorithm to service requests. Anticipatory scheduling benefits servers, then have it. If a request arrives, then just apply a C-SCAN algorithm. The approach is fair. While it has as support for real-time and may result in higher I/O latency, it works surprisingly well in benchmarks.

Smarter disks

Many disk scheduling algorithms were designed for disks where the operating system would address data blocks by cylinder, head, sector and also had a good idea of where the disk head was at any time. That is no longer the case. Disks now present a logical address space that looks like a flat memory that is used to hold frequently used blocks, as prefetch blocks, and to reassociate write requests. For example, consumer-grade Western Digital Caviar Black drives have read-ahead buffers and 64 MB caches.

Disks support logical block addressing (LBA), presenting their storage as a contiguous set of block numbers: as the actual location of a block is just an advanced guess (we guess that contiguous blocks are quicker to access than distant blocks). In addition, disks have:

- Automatic bad block mapping, which can mess up algorithms that assume blocks are laid out sequentially. When a disk is formatted, the formatting process leaves spare sectors on a track for remapping bad blocks.
- Native Command Queuing. Both SATA and SCSI protocols allow the drive to queue and try prioritize disk requests. Up to 256 commands can be queued up.
- The presence of a sizable cache on the disk allows frequently accessed blocks to be stored in the disk's RAM. Read-ahead can be enabled to optimize for sequential I/O. This causes the disk, if it has nothing else to do, to start successive blocks following a recited request into its cache.
- Some disks are a hybrid design, with a mix of non-volatile flash memory and disk storage. This creates bigger caches as well as persistent caches: the disk can acknowledge a write without actually writing it to magnetic storage.

Moreover, solid state disks are increasingly popular alternatives to magnetic storage. NAND flash memory still provides as with a block- based I/O interface but there is no seek latency whatsoever, so there is no advantages to reordering requests. Also, because flash memory can be written to only a limited number of times before it wears out, the goal is to even out the number of writes across all the cells. Algorithms will need to be cognizant of the amount of writes they do so to give more block and lead to consider wear-leveling, where block writes can be distributed among all the blocks of storage.

References

- The Linux Kernel Module Programming Guide. Peter Jay Salzman, Michael Burian, and Ori Pomerantz. Linux Documentation Project, May 18, 2007.
- Anatomy & Physiology of an Ioctl(). Patrick Brady. 2008 Google I/O Session Videos and Slides.
- Google I/O sessions, May 19—20, 2008.
- Linux Device Drivers, 3rd Edition: Interfacing Linux's and the 2.6 Kernel. O. John Shakelford. Real Pub Magazine
- www.ieee. Technique Hugo & Philip Cordeiro. Fortalexa Technologies
- MPI: Code-enabling Device Naming and Symbolic Links. Article ID: 191143, Microsoft Support
- Windows NT Device Driver Development. Peter G. Viscarola, W. Anthony Mason
- Kernel Korner – The New Block Queue Interface in the 2.6 Kernel. By Robert Love. Linux Journal, November 2003
- Writing Device Drivers. Copyright 1999, 2000, Scarlet and for its affiliates (formerly Sun).
- Linux Kernel Development, 3rd Edition. By Robert Love. Addison-Wesley Professional – Prefix Linux Club
- Linux Kernel Architecture for Device Drivers. Thomas Petazzoni, Free Electrons
- LDD3 – A Conveyance Implementation of Ioctls. Greg Kroah-Hartman. IBM Corp. Linux Technology Center. Proceedings of the Linux Symposium. July 21–24, 2003.

Unless otherwise noted or cited, these materials are licensed under Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License. All rights reserved.
This site is protected from unauthorized copyright and other commercial uses. Any part of this site may not be copied, reproduced, stored in, distributed in, or transmitted in any form or by any means whatsoever (electronic, photocopying, or otherwise) without the prior written consent of the copyright holder.
Any code is released under the BSD license.
Any unauthorized use of this site may violate copyright laws, trademark laws, the laws of privacy and publicity, and communications regulations and statutes.

Build: 0.0.1.0.20.20221 (Wed Jan 19 2022 16:20:20 GMT-0600 (Central Standard Time))