Nginx Event Mechanism

# 概述

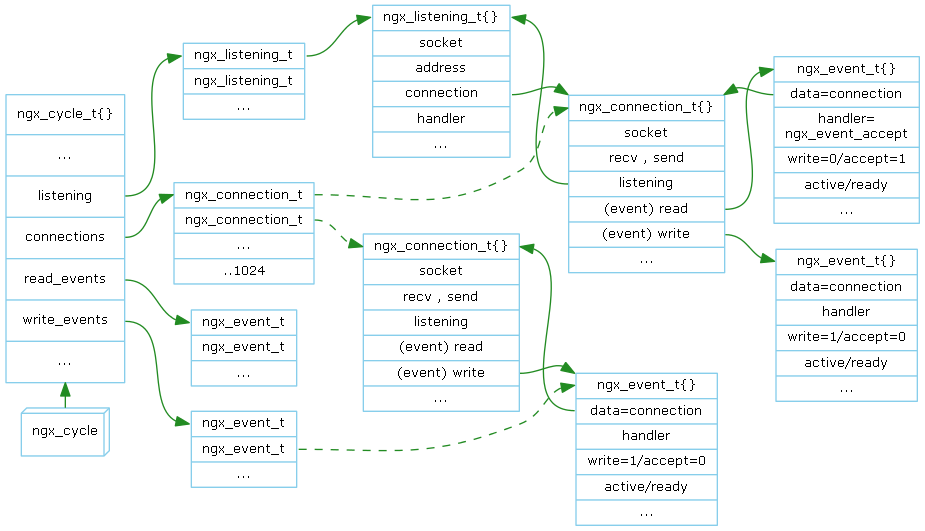
暂无

# 模块

* ngx\_event\_core\_module
* ngx\_select\_module
* ngx\_epoll\_module

# 文件句柄

## ngx\_listening\_t



用于管理Listen Socket，它由配置所创建。

* socket

文件描述符，等同于fd。

* address

略

* connection

指向ngx\_connection\_t，管理该Listen Socket的Connection对象。

* handler

当该Listen Socket收到连接请求并接受后的处理函数(回调)。

## ngx\_connection\_t

用于管理Socket的Connection对象，包括Listen Socket和Monitor Socket。这里关注Connection对事件和回调函数的管理。

* socket

文件描述符，等同于fd。

* recv send

收发函数指针，间接调用系统API recv和send。

* listening

反指向ngx\_listening\_t。如果socket不是Listen Socket，则listening指向相应的ngx\_listening\_t。

* read

指向读事件 ngx\_event\_t。

* write

指向写事件 ngx\_event\_t。

## ngx\_event\_t

抽象一个事件。在Nginx中事件是发生在文件描述符(file descriptor, socket)上的事件。比如socket有数据收到。

* data

指向ngx\_connection\_t对象。

* handler

事件发生后的处理函数(回调)。

* write accept

write表示是可写事件。

accept表示是可接受事件。可接受事件实际是Listen Socket上的可读事件。

write和accept都为0时，表示是可读事件。

* active

表示文件描述符已经送入Monitor。

* ready

表示事件已经发生。

## ngx\_cycle\_t

全局变量ngx\_cycle。生命周期内的数据管理。

* listening

ngx\_listening\_t数组，管理所有的Listen Sockets。

* connectons

预分配的ngx\_connection\_t数组资源。

* read\_events

预分配的可读事件ngx\_event\_t数组资源。

* write\_events

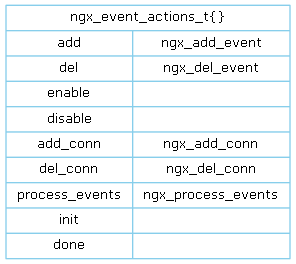
预分配的可写事件ngx\_event\_t数组资源。

# I/O多路复用模型

Nginx支持多种IO多路复用模型select、epoll等。 select是传统的模型，epoll是linux 2.6开始支持的更好的模型。

## 统一接口

* add



把ngx\_event\_t送入Monitor，ngx\_add\_event是一个宏定义的别名。

* del

把ngx\_event\_t从Monitor删除，ngx\_del\_event是一个宏定义的别名。

* enable

略

* disable

略

* add\_conn

把ngx\_connection\_t的read event和write event送入Monitor，ngx\_add\_conn是一个宏定义的别名。

* del\_conn

把ngx\_connection\_t的read event和write event从Monitor删除，ngx\_del\_conn是一个宏定义的别名。

* process\_events

等待事件发生，并调用ngx\_event\_t的handler处理回调函数或把ngx\_event\_t加入postpone队列，延迟处理。

ngx\_process\_events是一个宏定义的别名。

* init

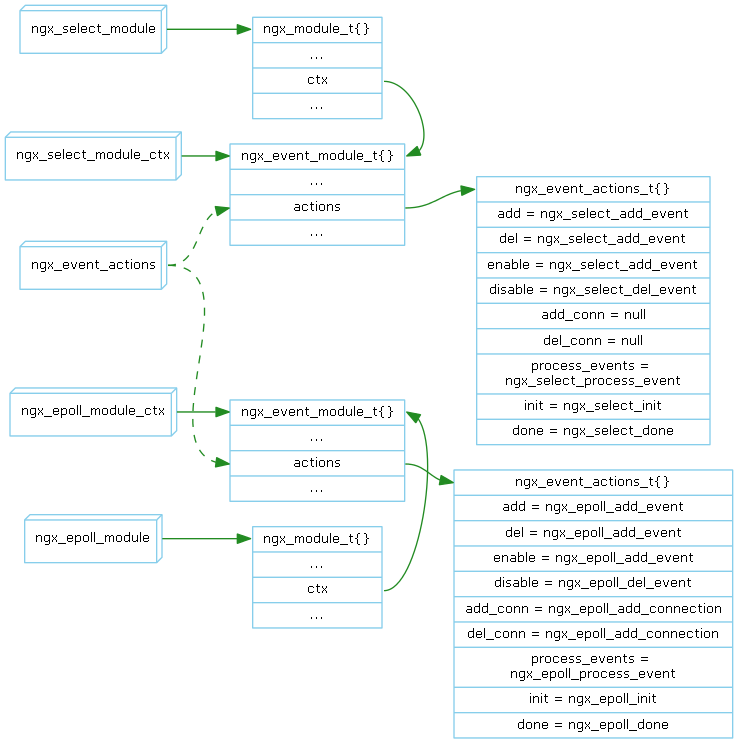
初始化I/O多路复用模型。

* done

略

## 数据组织

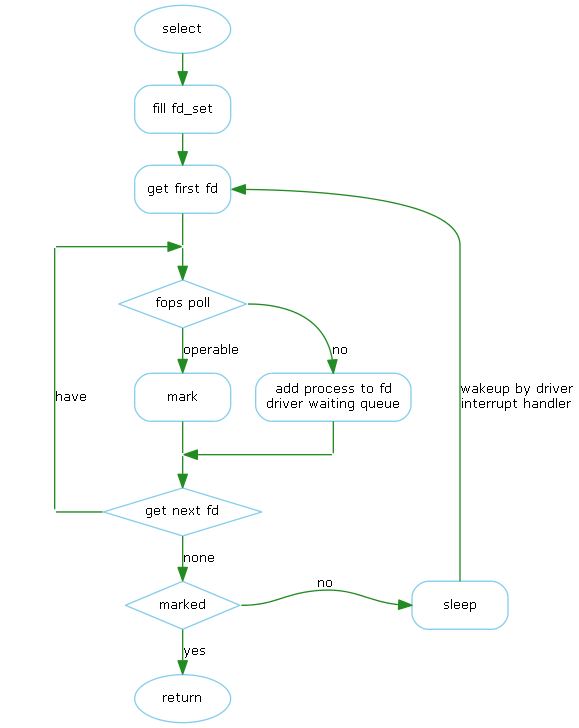
各个复用模型的接口实现由ngx\_module\_t和ngx\_event\_module\_t.actions管理。



全局变更ngx\_event\_actions指向(复制)了选用的接口实现(init函数)。如图。也就是说：

不管选用的是select模型还是epoll模型，调用ngx\_event\_actions接口即可。

## select机制



1. 从用户空间复制文件描述符集合到内核空间。
2. 遍历集合，调用poll（虚拟文件系统的一个接口）查看是否可操作（可读、可写、异常）。

2.1) 文件描述符可操作，加入可读、可写、异常集合。

2.2) 文件描述符不可操作，将进程挂入文件描述符对应的文件（设备）的等待队列。

1. 如果有任何一个文件描述符上可操作，则可操作的文件描述符集合从select返回。
2. 如果没有任何文件描述符上可操作，则进入sleep状态等待。
3. 任何一个文件描述符对应的文件（设备）中断发生，中断处理函数唤醒进程。
4. 进程从sleep返回，并进入2)继续运行。

上述流程图和过程是select的内部工作原理流程。调用者过程如下：

1. 复制文件描述符集合
2. 调用select
3. 遍历可操作集合和事件集合

可以看出它有三个方面的不足：

1. 每次调用时要重复地从用户空间复制文件描述符集合到内核空间。
2. 每次调用要把进程放入各个文件描述符的等待队列，返回前又从各个等待队列中删除。
3. 每次调用时要重复地遍历文件描述符(轮询)。

## epoll机制

针对select的三个不足，epoll机制作了改进，提供了三个接口函数：epoll\_create、epoll\_ctl、epoll\_wait。

1. 调用epoll\_create在内核创建一个对象EP。
2. 调用epoll\_ctl加入文件描述符。

2.1) 将文件描述符和用户参数加入EP的容器，以便删除。

2.2) 调用poll查看是否可操作，如果可操作则将文件描述符放入EP的一个数组RD。

2.3) 并将进程挂入文件描述符对应的文件（设备）的等待队列。

2.4) 向文件描述符对应的文件（设备）Driver注册回调函数挂入的等待队列。

1. 调用epoll\_wait

3.1) 如果RD不为空，则将RD返回。

3.2) 如果RD为空，则进入sleep状态等待。

1. （数据到达网卡后）文件描述符对应的文件（设备）中断发生，调用中断处理函数。
2. 中断处理函数调用回调函数，将文件描述符放入RD，并唤醒进程
3. 进程从sleep返回，并将RD从epoll\_wait返回。
4. 调用者从RD中得到文件描述符和用户参数，从用户参数可又得到ngx\_event\_t以及ngx\_connection\_t。

由于EP的存在，所以不需重复地从用户空间复制文件描述符集合到内核空间（解决了select第一个不足），

也不需要重复地将进程放入各个文件描述符的等待队列然后删除（解决了select第二个不足）。

epoll\_wait内部以及调用者不再需要遍历集合（解决了select第三个不足）。

# 超时管理

## 时间管理

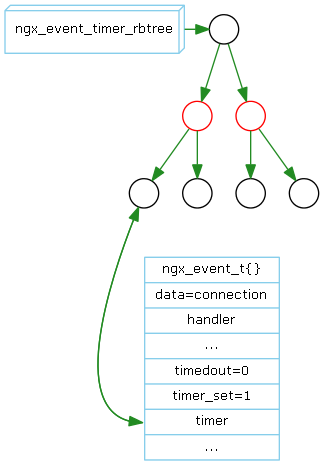
通常调用系统API gettimeofday来获得当前的时间值。由于系统调用开销问题，Nginx努力避免频繁调用gettimeofday。

Nginx以一定策略在select、epoll\_wait等阻塞函数返回之后调用gettimeofday，

并将获得的时间值保存到全局变量ngx\_cached\_time。此后读ngx\_cached\_time来获得当前时间。

这种方法会带来时间误差，但Nginx对时间精度要求并不高，这种方法是可接受的、高效的。

## 事件管理



Nginx使用红黑树来管理超时事件，并以超时时刻为KEY从早到晚排列。即最早要超时的事件在树的最左结点。

全局变量ngx\_event\_timer\_rbtree用于管理这棵红黑树。

事件ngx\_event\_t中有三个字段用于超时：

* timedout

事件是否已经超时。

* timer\_set

事件是否已经加入红黑树。

* timer

红黑树结点，将ngx\_event\_t挂进红黑树。已知它的指针可以用offsetof来算出ngx\_event\_t的地址。

## 扫描时机

### 事件触发扫描

超时事件是由排序的红黑树管理的，很容易（快）得到最早要超时的事件，即可算出第一个要超时事件距离当前的时间差。

以上述的时间差作为函数超时值调用select或epoll\_wait阻塞函数。select或epoll\_wait因事件发生或等待超时而返回后，

调用gettimeofday更新ngx\_cached\_time。如果当前时间ngx\_cached\_time大于上一次扫描红黑树的时间，则扫描红黑树，处理超时事件。

当事件很多时，select或epoll\_wait因为等到事件而频繁返回，导致频繁调用gettimeofday，和频繁扫描红黑树。

（TODO：定量分析）

于是就开发了定时触发扫描。

### 定时触发扫描

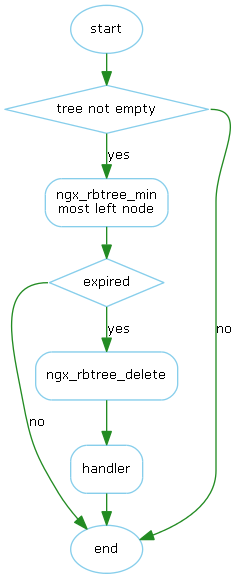
设置一个系统定时器(默认为100ms)。调用无超时的select或epoll\_wait阻塞函数。

如果select或epoll\_wait因事件发生而返回后，不调用gettimeofday更新ngx\_cached\_time，也就不会扫描红黑树。

如果select或epoll\_wait因定时器信号(SIGALRM)触发而中断返回后(SIGALRM信号处理函数中置标识)，

调用gettimeofday更新ngx\_cached\_time，当前时间ngx\_cached\_time一定大于上一次扫描红黑树的时间，扫描红黑树。

## 扫描过程



1. 红黑树没有结点，则结束扫描。
2. 如果最左结点没有超时，则结束扫描。
3. 如果最左结点超时，则把它从红黑树中删除，并调用回调处理函数。
4. 返回1)

扫描时，只需判断最左结点是否超时，而不需要扫描整棵树。

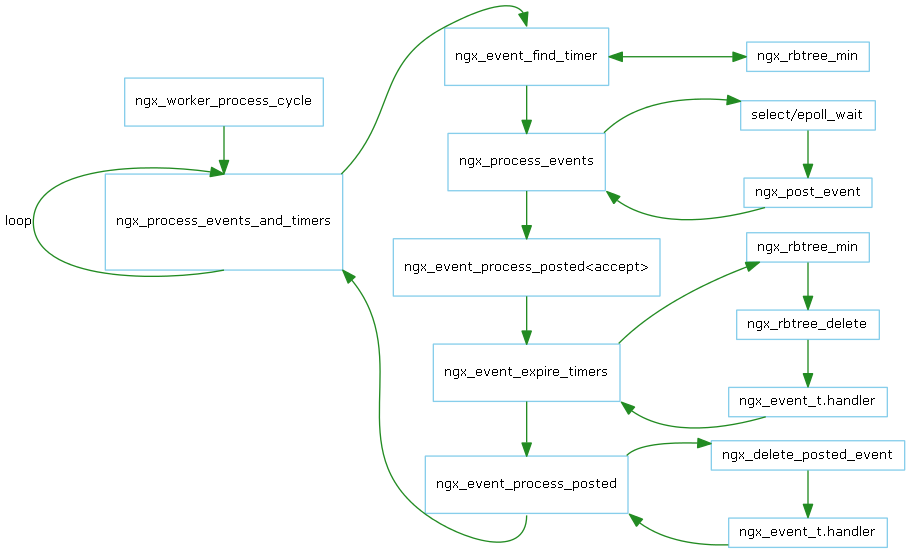
# 延迟处理

通过select或epoll\_wait收到事件后，可以把事件ngx\_event\_t先挂入ngx\_posted\_accept\_events和ngx\_posted\_events队列。

顾名思义，ngx\_posted\_accept\_events挂可接受事件，ngx\_posted\_events挂可读和可写事件。

事件全收下后，先处理ngx\_posted\_accept\_events队列中的事件，再处理超时，最后处理ngx\_posted\_events队列中的事件。

# 事件引擎



# Prototype

<https://github.com/lingjf/nginx_event.git>

# Latest revision

https://github.com/lingjf/nginx\_analyse/blob/master/doc/