# nginx平台初探(100%)

## 初探nginx架构(100%)

众所周知，nginx性能高，而nginx的高性能与其架构是分不开的。那么nginx究竟是怎么样的呢？这一节我们先来初识一下nginx框架吧。

nginx在启动后，在unix系统中会以daemon的方式在后台运行，后台进程包含一个master进程和多个worker进程。我们也可以手动地关掉后台模式，让nginx在前台运行，并且通过配置让nginx取消master进程，从而可以使nginx以单进程方式运行。很显然，生产环境下我们肯定不会这么做，所以关闭后台模式，一般是用来调试用的，在后面的章节里面，我们会详细地讲解如何调试nginx。所以，我们可以看到，nginx是以多进程的方式来工作的，当然nginx也是支持多线程的方式的，只是我们主流的方式还是多进程的方式，也是nginx的默认方式。nginx采用多进程的方式有诸多好处，所以我就主要讲解nginx的多进程模式吧。

刚才讲到，nginx在启动后，会有一个master进程和多个worker进程。master进程主要用来管理worker进程，包含：接收来自外界的信号，向各worker进程发送信号，监控worker进程的运行状态，当worker进程退出后(异常情况下)，会自动重新启动新的worker进程。而基本的网络事件，则是放在worker进程中来处理了。多个worker进程之间是对等的，他们同等竞争来自客户端的请求，各进程互相之间是独立的。一个请求，只可能在一个worker进程中处理，一个worker进程，不可能处理其它进程的请求。worker进程的个数是可以设置的，一般我们会设置与机器cpu核数一致，这里面的原因与nginx的进程模型以及事件处理模型是分不开的。nginx的进程模型，可以由下图来表示：



在nginx启动后，如果我们要操作nginx，要怎么做呢？从上文中我们可以看到，master来管理worker进程，所以我们只需要与master进程通信就行了。master进程会接收来自外界发来的信号，再根据信号做不同的事情。所以我们要控制nginx，只需要通过kill向master进程发送信号就行了。比如kill -HUP pid，则是告诉nginx，从容地重启nginx，我们一般用这个信号来重启nginx，或重新加载配置，因为是从容地重启，因此服务是不中断的。master进程在接收到HUP信号后是怎么做的呢？首先master进程在接到信号后，会先重新加载配置文件，然后再启动新的worker进程，并向所有老的worker进程发送信号，告诉他们可以光荣退休了。新的worker在启动后，就开始接收新的请求，而老的worker在收到来自master的信号后，就不再接收新的请求，并且在当前进程中的所有未处理完的请求处理完成后，再退出。当然，直接给master进程发送信号，这是比较老的操作方式，nginx在0.8版本之后，引入了一系列命令行参数，来方便我们管理。比如，./nginx -s reload，就是来重启nginx，./nginx -s stop，就是来停止nginx的运行。如何做到的呢？我们还是拿reload来说，我们看到，执行命令时，我们是启动一个新的nginx进程，而新的nginx进程在解析到reload参数后，就知道我们的目的是控制nginx来重新加载配置文件了，它会向master进程发送信号，然后接下来的动作，就和我们直接向master进程发送信号一样了。

现在，我们知道了当我们在操作nginx的时候，nginx内部做了些什么事情，那么，worker进程又是如何处理请求的呢？我们前面有提到，worker进程之间是平等的，每个进程，处理请求的机会也是一样的。当我们提供80端口的http服务时，一个连接请求过来，每个进程都有可能处理这个连接，怎么做到的呢？首先，每个worker进程都是从master进程fork过来，在master进程里面，先建立好需要listen的socket（listenfd）之后，然后再fork出多个worker进程。所有worker进程的listenfd会在新连接到来时变得可读，为保证只有一个进程处理该连接，所有worker进程在注册listenfd读事件前抢accept\_mutex，抢到互斥锁的那个进程注册listenfd读事件，在读事件里调用accept接受该连接。当一个worker进程在accept这个连接之后，就开始读取请求，解析请求，处理请求，产生数据后，再返回给客户端，最后才断开连接，这样一个完整的请求就是这样的了。我们可以看到，一个请求，完全由worker进程来处理，而且只在一个worker进程中处理。

那么，nginx采用这种进程模型有什么好处呢？当然，好处肯定会很多了。首先，对于每个worker进程来说，独立的进程，不需要加锁，所以省掉了锁带来的开销，同时在编程以及问题查找时，也会方便很多。其次，采用独立的进程，可以让互相之间不会影响，一个进程退出后，其它进程还在工作，服务不会中断，master进程则很快启动新的worker进程。当然，worker进程的异常退出，肯定是程序有bug了，异常退出，会导致当前worker上的所有请求失败，不过不会影响到所有请求，所以降低了风险。当然，好处还有很多，大家可以慢慢体会。

上面讲了很多关于nginx的进程模型，接下来，我们来看看nginx是如何处理事件的。

有人可能要问了，nginx采用多worker的方式来处理请求，每个worker里面只有一个主线程，那能够处理的并发数很有限啊，多少个worker就能处理多少个并发，何来高并发呢？非也，这就是nginx的高明之处，nginx采用了异步非阻塞的方式来处理请求，也就是说，nginx是可以同时处理成千上万个请求的。想想apache的常用工作方式（apache也有异步非阻塞版本，但因其与自带某些模块冲突，所以不常用），每个请求会独占一个工作线程，当并发数上到几千时，就同时有几千的线程在处理请求了。这对操作系统来说，是个不小的挑战，线程带来的内存占用非常大，线程的上下文切换带来的cpu开销很大，自然性能就上不去了，而这些开销完全是没有意义的。

为什么nginx可以采用异步非阻塞的方式来处理呢，或者异步非阻塞到底是怎么回事呢？我们先回到原点，看看一个请求的完整过程。首先，请求过来，要建立连接，然后再接收数据，接收数据后，再发送数据。具体到系统底层，就是读写事件，而当读写事件没有准备好时，必然不可操作，如果不用非阻塞的方式来调用，那就得阻塞调用了，事件没有准备好，那就只能等了，等事件准备好了，你再继续吧。阻塞调用会进入内核等待，cpu就会让出去给别人用了，对单线程的worker来说，显然不合适，当网络事件越多时，大家都在等待呢，cpu空闲下来没人用，cpu利用率自然上不去了，更别谈高并发了。好吧，你说加进程数，这跟apache的线程模型有什么区别，注意，别增加无谓的上下文切换。所以，在nginx里面，最忌讳阻塞的系统调用了。不要阻塞，那就非阻塞喽。非阻塞就是，事件没有准备好，马上返回EAGAIN，告诉你，事件还没准备好呢，你慌什么，过会再来吧。好吧，你过一会，再来检查一下事件，直到事件准备好了为止，在这期间，你就可以先去做其它事情，然后再来看看事件好了没。虽然不阻塞了，但你得不时地过来检查一下事件的状态，你可以做更多的事情了，但带来的开销也是不小的。所以，才会有了异步非阻塞的事件处理机制，具体到系统调用就是像select/poll/epoll/kqueue这样的系统调用。它们提供了一种机制，让你可以同时监控多个事件，调用他们是阻塞的，但可以设置超时时间，在超时时间之内，如果有事件准备好了，就返回。这种机制正好解决了我们上面的两个问题，拿epoll为例(在后面的例子中，我们多以epoll为例子，以代表这一类函数)，当事件没准备好时，放到epoll里面，事件准备好了，我们就去读写，当读写返回EAGAIN时，我们将它再次加入到epoll里面。这样，只要有事件准备好了，我们就去处理它，只有当所有事件都没准备好时，才在epoll里面等着。这样，我们就可以并发处理大量的并发了，当然，这里的并发请求，是指未处理完的请求，线程只有一个，所以同时能处理的请求当然只有一个了，只是在请求间进行不断地切换而已，切换也是因为异步事件未准备好，而主动让出的。这里的切换是没有任何代价，你可以理解为循环处理多个准备好的事件，事实上就是这样的。与多线程相比，这种事件处理方式是有很大的优势的，不需要创建线程，每个请求占用的内存也很少，没有上下文切换，事件处理非常的轻量级。并发数再多也不会导致无谓的资源浪费（上下文切换）。更多的并发数，只是会占用更多的内存而已。 我之前有对连接数进行过测试，在24G内存的机器上，处理的并发请求数达到过200万。现在的网络服务器基本都采用这种方式，这也是nginx性能高效的主要原因。

我们之前说过，推荐设置worker的个数为cpu的核数，在这里就很容易理解了，更多的worker数，只会导致进程来竞争cpu资源了，从而带来不必要的上下文切换。而且，nginx为了更好的利用多核特性，提供了cpu亲缘性的绑定选项，我们可以将某一个进程绑定在某一个核上，这样就不会因为进程的切换带来cache的失效。像这种小的优化在nginx中非常常见，同时也说明了nginx作者的苦心孤诣。比如，nginx在做4个字节的字符串比较时，会将4个字符转换成一个int型，再作比较，以减少cpu的指令数等等。

现在，知道了nginx为什么会选择这样的进程模型与事件模型了。对于一个基本的web服务器来说，事件通常有三种类型，网络事件、信号、定时器。从上面的讲解中知道，网络事件通过异步非阻塞可以很好的解决掉。如何处理信号与定时器？

首先，信号的处理。对nginx来说，有一些特定的信号，代表着特定的意义。信号会中断掉程序当前的运行，在改变状态后，继续执行。如果是系统调用，则可能会导致系统调用的失败，需要重入。关于信号的处理，大家可以学习一些专业书籍，这里不多说。对于nginx来说，如果nginx正在等待事件（epoll\_wait时），如果程序收到信号，在信号处理函数处理完后，epoll\_wait会返回错误，然后程序可再次进入epoll\_wait调用。

另外，再来看看定时器。由于epoll\_wait等函数在调用的时候是可以设置一个超时时间的，所以nginx借助这个超时时间来实现定时器。nginx里面的定时器事件是放在一棵维护定时器的红黑树里面，每次在进入epoll\_wait前，先从该红黑树里面拿到所有定时器事件的最小时间，在计算出epoll\_wait的超时时间后进入epoll\_wait。所以，当没有事件产生，也没有中断信号时，epoll\_wait会超时，也就是说，定时器事件到了。这时，nginx会检查所有的超时事件，将他们的状态设置为超时，然后再去处理网络事件。由此可以看出，当我们写nginx代码时，在处理网络事件的回调函数时，通常做的第一个事情就是判断超时，然后再去处理网络事件。

我们可以用一段伪代码来总结一下nginx的事件处理模型：

while (true) {  
 for t in run\_tasks:  
 t.handler();  
 update\_time(&now);  
 timeout = ETERNITY;  
 for t in wait\_tasks: /\* sorted already \*/  
 if (t.time <= now) {  
 t.timeout\_handler();  
 } else {  
 timeout = t.time - now;  
 break;  
 }  
 nevents = poll\_function(events, timeout);  
 for i in nevents:  
 task t;  
 if (events[i].type == READ) {  
 t.handler = read\_handler;  
 } else { /\* events[i].type == WRITE \*/  
 t.handler = write\_handler;  
 }  
 run\_tasks\_add(t);  
}

好，本节我们讲了进程模型，事件模型，包括网络事件，信号，定时器事件。

## nginx基础概念(100%)

### connection

在nginx中connection就是对tcp连接的封装，其中包括连接的socket，读事件，写事件。利用nginx封装的connection，我们可以很方便的使用nginx来处理与连接相关的事情，比如，建立连接，发送与接受数据等。而nginx中的http请求的处理就是建立在connection之上的，所以nginx不仅可以作为一个web服务器，也可以作为邮件服务器。当然，利用nginx提供的connection，我们可以与任何后端服务打交道。

结合一个tcp连接的生命周期，我们看看nginx是如何处理一个连接的。首先，nginx在启动时，会解析配置文件，得到需要监听的端口与ip地址，然后在nginx的master进程里面，先初始化好这个监控的socket(创建socket，设置addrreuse等选项，绑定到指定的ip地址端口，再listen)，然后再fork出多个子进程出来，然后子进程会竞争accept新的连接。此时，客户端就可以向nginx发起连接了。当客户端与服务端通过三次握手建立好一个连接后，nginx的某一个子进程会accept成功，得到这个建立好的连接的socket，然后创建nginx对连接的封装，即ngx\_connection\_t结构体。接着，设置读写事件处理函数并添加读写事件来与客户端进行数据的交换。最后，nginx或客户端来主动关掉连接，到此，一个连接就寿终正寝了。

当然，nginx也是可以作为客户端来请求其它server的数据的（如upstream模块），此时，与其它server创建的连接，也封装在ngx\_connection\_t中。作为客户端，nginx先获取一个ngx\_connection\_t结构体，然后创建socket，并设置socket的属性（ 比如非阻塞）。然后再通过添加读写事件，调用connect/read/write来调用连接，最后关掉连接，并释放ngx\_connection\_t。

在nginx中，每个进程会有一个连接数的最大上限，这个上限与系统对fd的限制不一样。在操作系统中，通过ulimit -n，我们可以得到一个进程所能够打开的fd的最大数，即nofile，因为每个socket连接会占用掉一个fd，所以这也会限制我们进程的最大连接数，当然也会直接影响到我们程序所能支持的最大并发数，当fd用完后，再创建socket时，就会失败。nginx通过设置worker\_connectons来设置每个进程支持的最大连接数。如果该值大于nofile，那么实际的最大连接数是nofile，nginx会有警告。nginx在实现时，是通过一个连接池来管理的，每个worker进程都有一个独立的连接池，连接池的大小是worker\_connections。这里的连接池里面保存的其实不是真实的连接，它只是一个worker\_connections大小的一个ngx\_connection\_t结构的数组。并且，nginx会通过一个链表free\_connections来保存所有的空闲ngx\_connection\_t，每次获取一个连接时，就从空闲连接链表中获取一个，用完后，再放回空闲连接链表里面。

在这里，很多人会误解worker\_connections这个参数的意思，认为这个值就是nginx所能建立连接的最大值。其实不然，这个值是表示每个worker进程所能建立连接的最大值，所以，一个nginx能建立的最大连接数，应该是worker\_connections \* worker\_processes。当然，这里说的是最大连接数，对于HTTP请求本地资源来说，能够支持的最大并发数量是worker\_connections \* worker\_processes，而如果是HTTP作为反向代理来说，最大并发数量应该是worker\_connections \* worker\_processes/2。因为作为反向代理服务器，每个并发会建立与客户端的连接和与后端服务的连接，会占用两个连接。

那么，我们前面有说过一个客户端连接过来后，多个空闲的进程，会竞争这个连接，很容易看到，这种竞争会导致不公平，如果某个进程得到accept的机会比较多，它的空闲连接很快就用完了，如果不提前做一些控制，当accept到一个新的tcp连接后，因为无法得到空闲连接，而且无法将此连接转交给其它进程，最终会导致此tcp连接得不到处理，就中止掉了。很显然，这是不公平的，有的进程有空余连接，却没有处理机会，有的进程因为没有空余连接，却人为地丢弃连接。那么，如何解决这个问题呢？首先，nginx的处理得先打开accept\_mutex选项，此时，只有获得了accept\_mutex的进程才会去添加accept事件，也就是说，nginx会控制进程是否添加accept事件。nginx使用一个叫ngx\_accept\_disabled的变量来控制是否去竞争accept\_mutex锁。在第一段代码中，计算ngx\_accept\_disabled的值，这个值是nginx单进程的所有连接总数的八分之一，减去剩下的空闲连接数量，得到的这个ngx\_accept\_disabled有一个规律，当剩余连接数小于总连接数的八分之一时，其值才大于0，而且剩余的连接数越小，这个值越大。再看第二段代码，当ngx\_accept\_disabled大于0时，不会去尝试获取accept\_mutex锁，并且将ngx\_accept\_disabled减1，于是，每次执行到此处时，都会去减1，直到小于0。不去获取accept\_mutex锁，就是等于让出获取连接的机会，很显然可以看出，当空余连接越少时，ngx\_accept\_disable越大，于是让出的机会就越多，这样其它进程获取锁的机会也就越大。不去accept，自己的连接就控制下来了，其它进程的连接池就会得到利用，这样，nginx就控制了多进程间连接的平衡了。

ngx\_accept\_disabled = ngx\_cycle->connection\_n / 8  
 - ngx\_cycle->free\_connection\_n;  
  
if (ngx\_accept\_disabled > 0) {  
 ngx\_accept\_disabled--;  
  
} else {  
 if (ngx\_trylock\_accept\_mutex(cycle) == NGX\_ERROR) {  
 return;  
 }  
  
 if (ngx\_accept\_mutex\_held) {  
 flags |= NGX\_POST\_EVENTS;  
  
 } else {  
 if (timer == NGX\_TIMER\_INFINITE  
 || timer > ngx\_accept\_mutex\_delay)  
 {  
 timer = ngx\_accept\_mutex\_delay;  
 }  
 }  
}

好了，连接就先介绍到这，本章的目的是介绍基本概念，知道在nginx中连接是个什么东西就行了，而且连接是属于比较高级的用法，在后面的模块开发高级篇会有专门的章节来讲解连接与事件的实现及使用。

### request

这节我们讲request，在nginx中我们指的是http请求，具体到nginx中的数据结构是ngx\_http\_request\_t。ngx\_http\_request\_t是对一个http请求的封装。 我们知道，一个http请求，包含请求行、请求头、请求体、响应行、响应头、响应体。

http请求是典型的请求-响应类型的的网络协议，而http是文本协议，所以我们在分析请求行与请求头，以及输出响应行与响应头，往往是一行一行的进行处理。如果我们自己来写一个http服务器，通常在一个连接建立好后，客户端会发送请求过来。然后我们读取一行数据，分析出请求行中包含的method、uri、http\_version信息。然后再一行一行处理请求头，并根据请求method与请求头的信息来决定是否有请求体以及请求体的长度，然后再去读取请求体。得到请求后，我们处理请求产生需要输出的数据，然后再生成响应行，响应头以及响应体。在将响应发送给客户端之后，一个完整的请求就处理完了。当然这是最简单的webserver的处理方式，其实nginx也是这样做的，只是有一些小小的区别，比如，当请求头读取完成后，就开始进行请求的处理了。nginx通过ngx\_http\_request\_t来保存解析请求与输出响应相关的数据。

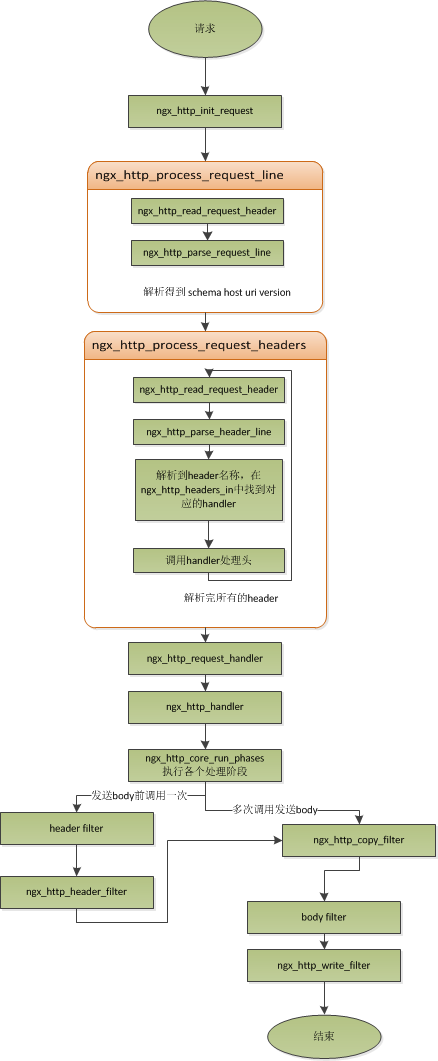
那接下来，简要讲讲nginx是如何处理一个完整的请求的。对于nginx来说，一个请求是从ngx\_http\_init\_request开始的，在这个函数中，会设置读事件为ngx\_http\_process\_request\_line，也就是说，接下来的网络事件，会由ngx\_http\_process\_request\_line来执行。从ngx\_http\_process\_request\_line的函数名，我们可以看到，这就是来处理请求行的，正好与之前讲的，处理请求的第一件事就是处理请求行是一致的。通过ngx\_http\_read\_request\_header来读取请求数据。然后调用ngx\_http\_parse\_request\_line函数来解析请求行。nginx为提高效率，采用状态机来解析请求行，而且在进行method的比较时，没有直接使用字符串比较，而是将四个字符转换成一个整型，然后一次比较以减少cpu的指令数，这个前面有说过。很多人可能很清楚一个请求行包含请求的方法，uri，版本，却不知道其实在请求行中，也是可以包含有host的。比如一个请求GET <http://www.taobao.com/uri> HTTP/1.0这样一个请求行也是合法的，而且host是www.taobao.com，这个时候，nginx会忽略请求头中的host域，而以请求行中的这个为准来查找虚拟主机。另外，对于http0.9版来说，是不支持请求头的，所以这里也是要特别的处理。所以，在后面解析请求头时，协议版本都是1.0或1.1。整个请求行解析到的参数，会保存到ngx\_http\_request\_t结构当中。

在解析完请求行后，nginx会设置读事件的handler为ngx\_http\_process\_request\_headers，然后后续的请求就在ngx\_http\_process\_request\_headers中进行读取与解析。ngx\_http\_process\_request\_headers函数用来读取请求头，跟请求行一样，还是调用ngx\_http\_read\_request\_header来读取请求头，调用ngx\_http\_parse\_header\_line来解析一行请求头，解析到的请求头会保存到ngx\_http\_request\_t的域headers\_in中，headers\_in是一个链表结构，保存所有的请求头。而HTTP中有些请求是需要特别处理的，这些请求头与请求处理函数存放在一个映射表里面，即ngx\_http\_headers\_in，在初始化时，会生成一个hash表，当每解析到一个请求头后，就会先在这个hash表中查找，如果有找到，则调用相应的处理函数来处理这个请求头。比如:Host头的处理函数是ngx\_http\_process\_host。

当nginx解析到两个回车换行符时，就表示请求头的结束，此时就会调用ngx\_http\_process\_request来处理请求了。ngx\_http\_process\_request会设置当前的连接的读写事件处理函数为ngx\_http\_request\_handler，然后再调用ngx\_http\_handler来真正开始处理一个完整的http请求。这里可能比较奇怪，读写事件处理函数都是ngx\_http\_request\_handler，其实在这个函数中，会根据当前事件是读事件还是写事件，分别调用ngx\_http\_request\_t中的read\_event\_handler或者是write\_event\_handler。由于此时，我们的请求头已经读取完成了，之前有说过，nginx的做法是先不读取请求body，所以这里面我们设置read\_event\_handler为ngx\_http\_block\_reading，即不读取数据了。刚才说到，真正开始处理数据，是在ngx\_http\_handler这个函数里面，这个函数会设置write\_event\_handler为ngx\_http\_core\_run\_phases，并执行ngx\_http\_core\_run\_phases函数。ngx\_http\_core\_run\_phases这个函数将执行多阶段请求处理，nginx将一个http请求的处理分为多个阶段，那么这个函数就是执行这些阶段来产生数据。因为ngx\_http\_core\_run\_phases最后会产生数据，所以我们就很容易理解，为什么设置写事件的处理函数为ngx\_http\_core\_run\_phases了。在这里，我简要说明了一下函数的调用逻辑，我们需要明白最终是调用ngx\_http\_core\_run\_phases来处理请求，产生的响应头会放在ngx\_http\_request\_t的headers\_out中，这一部分内容，我会放在请求处理流程里面去讲。nginx的各种阶段会对请求进行处理，最后会调用filter来过滤数据，对数据进行加工，如truncked传输、gzip压缩等。这里的filter包括header filter与body filter，即对响应头或响应体进行处理。filter是一个链表结构，分别有header filter与body filter，先执行header filter中的所有filter，然后再执行body filter中的所有filter。在header filter中的最后一个filter，即ngx\_http\_header\_filter，这个filter将会遍历所有的响应头，最后需要输出的响应头在一个连续的内存，然后调用ngx\_http\_write\_filter进行输出。ngx\_http\_write\_filter是body filter中的最后一个，所以nginx首先的body信息，在经过一系列的body filter之后，最后也会调用ngx\_http\_write\_filter来进行输出(有图来说明)。

这里要注意的是，nginx会将整个请求头都放在一个buffer里面，这个buffer的大小通过配置项client\_header\_buffer\_size来设置，如果用户的请求头太大，这个buffer装不下，那nginx就会重新分配一个新的更大的buffer来装请求头，这个大buffer可以通过large\_client\_header\_buffers来设置，这个large\_buffer这一组buffer，比如配置4 8k，就是表示有四个8k大小的buffer可以用。注意，为了保存请求行或请求头的完整性，一个完整的请求行或请求头，需要放在一个连续的内存里面，所以，一个完整的请求行或请求头，只会保存在一个buffer里面。这样，如果请求行大于一个buffer的大小，就会返回414错误，如果一个请求头大小大于一个buffer大小，就会返回400错误。在了解了这些参数的值，以及nginx实际的做法之后，在应用场景，我们就需要根据实际的需求来调整这些参数，来优化我们的程序了。

处理流程图：



以上这些，就是nginx中一个http请求的生命周期了。我们再看看与请求相关的一些概念吧。

#### keepalive

当然，在nginx中，对于http1.0与http1.1也是支持长连接的。什么是长连接呢？我们知道，http请求是基于TCP协议之上的，那么，当客户端在发起请求前，需要先与服务端建立TCP连接，而每一次的TCP连接是需要三次握手来确定的，如果客户端与服务端之间网络差一点，这三次交互消费的时间会比较多，而且三次交互也会带来网络流量。当然，当连接断开后，也会有四次的交互，当然对用户体验来说就不重要了。而http请求是请求应答式的，如果我们能知道每个请求头与响应体的长度，那么我们是可以在一个连接上面执行多个请求的，这就是所谓的长连接，但前提条件是我们先得确定请求头与响应体的长度。对于请求来说，如果当前请求需要有body，如POST请求，那么nginx就需要客户端在请求头中指定content-length来表明body的大小，否则返回400错误。也就是说，请求体的长度是确定的，那么响应体的长度呢？先来看看http协议中关于响应body长度的确定：

1. 对于http1.0协议来说，如果响应头中有content-length头，则以content-length的长度就可以知道body的长度了，客户端在接收body时，就可以依照这个长度来接收数据，接收完后，就表示这个请求完成了。而如果没有content-length头，则客户端会一直接收数据，直到服务端主动断开连接，才表示body接收完了。
2. 而对于http1.1协议来说，如果响应头中的Transfer-encoding为chunked传输，则表示body是流式输出，body会被分成多个块，每块的开始会标识出当前块的长度，此时，body不需要通过长度来指定。如果是非chunked传输，而且有content-length，则按照content-length来接收数据。否则，如果是非chunked，并且没有content-length，则客户端接收数据，直到服务端主动断开连接。

从上面，我们可以看到，除了http1.0不带content-length以及http1.1非chunked不带content-length外，body的长度是可知的。此时，当服务端在输出完body之后，会可以考虑使用长连接。能否使用长连接，也是有条件限制的。如果客户端的请求头中的connection为close，则表示客户端需要关掉长连接，如果为keep-alive，则客户端需要打开长连接，如果客户端的请求中没有connection这个头，那么根据协议，如果是http1.0，则默认为close，如果是http1.1，则默认为keep-alive。如果结果为keepalive，那么，nginx在输出完响应体后，会设置当前连接的keepalive属性，然后等待客户端下一次请求。当然，nginx不可能一直等待下去，如果客户端一直不发数据过来，岂不是一直占用这个连接？所以当nginx设置了keepalive等待下一次的请求时，同时也会设置一个最大等待时间，这个时间是通过选项keepalive\_timeout来配置的，如果配置为0，则表示关掉keepalive，此时，http版本无论是1.1还是1.0，客户端的connection不管是close还是keepalive，都会强制为close。

如果服务端最后的决定是keepalive打开，那么在响应的http头里面，也会包含有connection头域，其值是\"Keep-Alive\"，否则就是\"Close\"。如果connection值为close，那么在nginx响应完数据后，会主动关掉连接。所以，对于请求量比较大的nginx来说，关掉keepalive最后会产生比较多的time-wait状态的socket。一般来说，当客户端的一次访问，需要多次访问同一个server时，打开keepalive的优势非常大，比如图片服务器，通常一个网页会包含很多个图片。打开keepalive也会大量减少time-wait的数量。

#### pipe

在http1.1中，引入了一种新的特性，即pipeline。那么什么是pipeline呢？pipeline其实就是流水线作业，它可以看作为keepalive的一种升华，因为pipeline也是基于长连接的，目的就是利用一个连接做多次请求。如果客户端要提交多个请求，对于keepalive来说，那么第二个请求，必须要等到第一个请求的响应接收完全后，才能发起，这和TCP的停止等待协议是一样的，得到两个响应的时间至少为2\*RTT。而对pipeline来说，客户端不必等到第一个请求处理完后，就可以马上发起第二个请求。得到两个响应的时间可能能够达到1\*RTT。nginx是直接支持pipeline的，但是，nginx对pipeline中的多个请求的处理却不是并行的，依然是一个请求接一个请求的处理，只是在处理第一个请求的时候，客户端就可以发起第二个请求。这样，nginx利用pipeline减少了处理完一个请求后，等待第二个请求的请求头数据的时间。其实nginx的做法很简单，前面说到，nginx在读取数据时，会将读取的数据放到一个buffer里面，所以，如果nginx在处理完前一个请求后，如果发现buffer里面还有数据，就认为剩下的数据是下一个请求的开始，然后就接下来处理下一个请求，否则就设置keepalive。

#### lingering\_close

lingering\_close，字面意思就是延迟关闭，也就是说，当nginx要关闭连接时，并非立即关闭连接，而是先关闭tcp连接的写，再等待一段时间后再关掉连接的读。为什么要这样呢？我们先来看看这样一个场景。nginx在接收客户端的请求时，可能由于客户端或服务端出错了，要立即响应错误信息给客户端，而nginx在响应错误信息后，大分部情况下是需要关闭当前连接。nginx执行完write()系统调用把错误信息发送给客户端，write()系统调用返回成功并不表示数据已经发送到客户端，有可能还在tcp连接的write buffer里。接着如果直接执行close()系统调用关闭tcp连接，内核会首先检查tcp的read buffer里有没有客户端发送过来的数据留在内核态没有被用户态进程读取，如果有则发送给客户端RST报文来关闭tcp连接丢弃write buffer里的数据，如果没有则等待write buffer里的数据发送完毕，然后再经过正常的4次分手报文断开连接。所以,当在某些场景下出现tcp write buffer里的数据在write()系统调用之后到close()系统调用执行之前没有发送完毕，且tcp read buffer里面还有数据没有读，close()系统调用会导致客户端收到RST报文且不会拿到服务端发送过来的错误信息数据。那客户端肯定会想，这服务器好霸道，动不动就reset我的连接，连个错误信息都没有。

在上面这个场景中，我们可以看到，关键点是服务端给客户端发送了RST包，导致自己发送的数据在客户端忽略掉了。所以，解决问题的重点是，让服务端别发RST包。再想想，我们发送RST是因为我们关掉了连接，关掉连接是因为我们不想再处理此连接了，也不会有任何数据产生了。对于全双工的TCP连接来说，我们只需要关掉写就行了，读可以继续进行，我们只需要丢掉读到的任何数据就行了，这样的话，当我们关掉连接后，客户端再发过来的数据，就不会再收到RST了。当然最终我们还是需要关掉这个读端的，所以我们会设置一个超时时间，在这个时间过后，就关掉读，客户端再发送数据来就不管了，作为服务端我会认为，都这么长时间了，发给你的错误信息也应该读到了，再慢就不关我事了，要怪就怪你RP不好了。当然，正常的客户端，在读取到数据后，会关掉连接，此时服务端就会在超时时间内关掉读端。这些正是lingering\_close所做的事情。协议栈提供 SO\_LINGER 这个选项，它的一种配置情况就是来处理lingering\_close的情况的，不过nginx是自己实现的lingering\_close。lingering\_close存在的意义就是来读取剩下的客户端发来的数据，所以nginx会有一个读超时时间，通过lingering\_timeout选项来设置，如果在lingering\_timeout时间内还没有收到数据，则直接关掉连接。nginx还支持设置一个总的读取时间，通过lingering\_time来设置，这个时间也就是nginx在关闭写之后，保留socket的时间，客户端需要在这个时间内发送完所有的数据，否则nginx在这个时间过后，会直接关掉连接。当然，nginx是支持配置是否打开lingering\_close选项的，通过lingering\_close选项来配置。 那么，我们在实际应用中，是否应该打开lingering\_close呢？这个就没有固定的推荐值了，如Maxim Dounin所说，lingering\_close的主要作用是保持更好的客户端兼容性，但是却需要消耗更多的额外资源（比如连接会一直占着）。

这节，我们介绍了nginx中，连接与请求的基本概念，下节，我们讲基本的数据结构。

## 基本数据结构(99%)

nginx的作者为追求极致的高效，自己实现了很多颇具特色的nginx风格的数据结构以及公共函数。比如，nginx提供了带长度的字符串，根据编译器选项优化过的字符串拷贝函数ngx\_copy等。所以，在我们写nginx模块时，应该尽量调用nginx提供的api，尽管有些api只是对glibc的宏定义。本节，我们介绍string、list、buffer、chain等一系列最基本的数据结构及相关api的使用技巧以及注意事项。

### ngx\_str\_t(100%)

在nginx源码目录的src/core下面的ngx\_string.h|c里面，包含了字符串的封装以及字符串相关操作的api。nginx提供了一个带长度的字符串结构ngx\_str\_t，它的原型如下：

typedef struct {  
 size\_t len;  
 u\_char \*data;  
} ngx\_str\_t;

在结构体当中，data指向字符串数据的第一个字符，字符串的结束用长度来表示，而不是由\'\0\'来表示结束。所以，在写nginx代码时，处理字符串的方法跟我们平时使用有很大的不一样，但要时刻记住，字符串不以\'\0\'结束，尽量使用nginx提供的字符串操作的api来操作字符串。 那么，nginx这样做有什么好处呢？首先，通过长度来表示字符串长度，减少计算字符串长度的次数。其次，nginx可以重复引用一段字符串内存，data可以指向任意内存，长度表示结束，而不用去copy一份自己的字符串(因为如果要以\'\0\'结束，而不能更改原字符串，所以势必要copy一段字符串)。我们在ngx\_http\_request\_t结构体的成员中，可以找到很多字符串引用一段内存的例子，比如request\_line、uri、args等等，这些字符串的data部分，都是指向在接收数据时创建buffer所指向的内存中，uri，args就没有必要copy一份出来。这样的话，减少了很多不必要的内存分配与拷贝。 正是基于此特性，在nginx中，必须谨慎的去修改一个字符串。在修改字符串时需要认真的去考虑：是否可以修改该字符串；字符串修改后，是否会对其它的引用造成影响。在后面介绍ngx\_unescape\_uri函数的时候，就会看到这一点。但是，使用nginx的字符串会产生一些问题，glibc提供的很多系统api函数大多是通过\'\0\'来表示字符串的结束，所以我们在调用系统api时，就不能直接传入str->data了。此时，通常的做法是创建一段str->len + 1大小的内存，然后copy字符串，最后一个字节置为\'\0\'。比较hack的做法是，将字符串最后一个字符的后一个字符backup一个，然后设置为\'\0\'，在做完调用后，再由backup改回来，但前提条件是，你得确定这个字符是可以修改的，而且是有内存分配，不会越界，但一般不建议这么做。 接下来，看看nginx提供的操作字符串相关的api。

#define ngx\_string(str) { sizeof(str) - 1, (u\_char \*) str }

ngx\_string(str)是一个宏，它通过一个以\'\0\'结尾的普通字符串str构造一个nginx的字符串，鉴于其中采用sizeof操作符计算字符串长度，因此参数必须是一个常量字符串。

#define ngx\_null\_string { 0, NULL }

定义变量时，使用ngx\_null\_string初始化字符串为空字符串，字符串的长度为0，data为NULL。

#define ngx\_str\_set(str, text) \  
 (str)->len = sizeof(text) - 1; (str)->data = (u\_char \*) text

ngx\_str\_set用于设置字符串str为text，由于使用sizeof计算长度，故text必须为常量字符串。

#define ngx\_str\_null(str) (str)->len = 0; (str)->data = NULL

ngx\_str\_null用于设置字符串str为空串，长度为0，data为NULL。

上面这四个函数，使用时一定要小心，ngx\_string与ngx\_null\_string是"{*，*}"格式的，故只能用于赋值时初始化，如：

ngx\_str\_t str = ngx\_string("hello world");  
ngx\_str\_t str1 = ngx\_null\_string;

如果像下面这样使用，就会有问题，这里涉及到c语言中对结构体变量赋值操作的语法规则，在此不做介绍。

ngx\_str\_t str, str1;  
str = ngx\_string("hello world"); // 编译出错  
str1 = ngx\_null\_string; // 编译出错

这种情况，可以调用ngx\_str\_set与ngx\_str\_null这两个函数来做:

ngx\_str\_t str, str1;  
ngx\_str\_set(&str, "hello world");  
ngx\_str\_null(&str1);

按照C99标准，您也可以这么做：

ngx\_str\_t str, str1;  
str = (ngx\_str\_t) ngx\_string("hello world");  
str1 = (ngx\_str\_t) ngx\_null\_string;

另外要注意的是，ngx\_string与ngx\_str\_set在调用时，传进去的字符串一定是常量字符串，否则会得到意想不到的错误(因为ngx\_str\_set内部使用了sizeof()，如果传入的是u\_char\*，那么计算的是这个指针的长度，而不是字符串的长度)。如：

ngx\_str\_t str;  
u\_char \*a = "hello world";  
ngx\_str\_set(&str, a); // 问题产生

此外，值得注意的是，由于ngx\_str\_set与ngx\_str\_null实际上是两行语句，故在if/for/while等语句中单独使用需要用花括号括起来，例如：

ngx\_str\_t str;  
if (cond)  
 ngx\_str\_set(&str, "true"); // 问题产生  
else  
 ngx\_str\_set(&str, "false"); // 问题产生

void ngx\_strlow(u\_char \*dst, u\_char \*src, size\_t n);

将src的前n个字符转换成小写存放在dst字符串当中，调用者需要保证dst指向的空间大于等于n，且指向的空间必须可写。操作不会对原字符串产生变动。如要更改原字符串，可以：

ngx\_strlow(str->data, str->data, str->len);

ngx\_strncmp(s1, s2, n)

区分大小写的字符串比较，只比较前n个字符。

ngx\_strcmp(s1, s2)

区分大小写的不带长度的字符串比较。

ngx\_int\_t ngx\_strcasecmp(u\_char \*s1, u\_char \*s2);

不区分大小写的不带长度的字符串比较。

ngx\_int\_t ngx\_strncasecmp(u\_char \*s1, u\_char \*s2, size\_t n);

不区分大小写的带长度的字符串比较，只比较前n个字符。

u\_char \* ngx\_cdecl ngx\_sprintf(u\_char \*buf, const char \*fmt, ...);  
u\_char \* ngx\_cdecl ngx\_snprintf(u\_char \*buf, size\_t max, const char \*fmt, ...);  
u\_char \* ngx\_cdecl ngx\_slprintf(u\_char \*buf, u\_char \*last, const char \*fmt, ...);

上面这三个函数用于字符串格式化，ngx\_snprintf的第二个参数max指明buf的空间大小，ngx\_slprintf则通过last来指明buf空间的大小。推荐使用第二个或第三个函数来格式化字符串，ngx\_sprintf函数还是比较危险的，容易产生缓冲区溢出漏洞。在这一系列函数中，nginx在兼容glibc中格式化字符串的形式之外，还添加了一些方便格式化nginx类型的一些转义字符，比如%V用于格式化ngx\_str\_t结构。在nginx源文件的ngx\_string.c中有说明：

/\*  
 \* supported formats:  
 \* %[0][width][x][X]O off\_t  
 \* %[0][width]T time\_t  
 \* %[0][width][u][x|X]z ssize\_t/size\_t  
 \* %[0][width][u][x|X]d int/u\_int  
 \* %[0][width][u][x|X]l long  
 \* %[0][width|m][u][x|X]i ngx\_int\_t/ngx\_uint\_t  
 \* %[0][width][u][x|X]D int32\_t/uint32\_t  
 \* %[0][width][u][x|X]L int64\_t/uint64\_t  
 \* %[0][width|m][u][x|X]A ngx\_atomic\_int\_t/ngx\_atomic\_uint\_t  
 \* %[0][width][.width]f double, max valid number fits to %18.15f  
 \* %P ngx\_pid\_t  
 \* %M ngx\_msec\_t  
 \* %r rlim\_t  
 \* %p void \*  
 \* %V ngx\_str\_t \*  
 \* %v ngx\_variable\_value\_t \*  
 \* %s null-terminated string  
 \* %\*s length and string  
 \* %Z '\0'  
 \* %N '\n'  
 \* %c char  
 \* %% %  
 \*  
 \* reserved:  
 \* %t ptrdiff\_t  
 \* %S null-terminated wchar string  
 \* %C wchar  
 \*/

这里特别要提醒的是，我们最常用于格式化ngx\_str\_t结构，其对应的转义符是%V，传给函数的一定要是指针类型，否则程序就会coredump掉。这也是我们最容易犯的错。比如：

ngx\_str\_t str = ngx\_string("hello world");  
u\_char buffer[1024];  
ngx\_snprintf(buffer, 1024, "%V", &str); // 注意，str取地址

void ngx\_encode\_base64(ngx\_str\_t \*dst, ngx\_str\_t \*src);  
ngx\_int\_t ngx\_decode\_base64(ngx\_str\_t \*dst, ngx\_str\_t \*src);

这两个函数用于对str进行base64编码与解码，调用前，需要保证dst中有足够的空间来存放结果，如果不知道具体大小，可先调用ngx\_base64\_encoded\_length与ngx\_base64\_decoded\_length来预估最大占用空间。

uintptr\_t ngx\_escape\_uri(u\_char \*dst, u\_char \*src, size\_t size,  
 ngx\_uint\_t type);

对src进行编码，根据type来按不同的方式进行编码，如果dst为NULL，则返回需要转义的字符的数量，由此可得到需要的空间大小。type的类型可以是：

#define NGX\_ESCAPE\_URI 0  
#define NGX\_ESCAPE\_ARGS 1  
#define NGX\_ESCAPE\_HTML 2  
#define NGX\_ESCAPE\_REFRESH 3  
#define NGX\_ESCAPE\_MEMCACHED 4  
#define NGX\_ESCAPE\_MAIL\_AUTH 5

void ngx\_unescape\_uri(u\_char \*\*dst, u\_char \*\*src, size\_t size, ngx\_uint\_t type);

对src进行反编码，type可以是0、NGX\_UNESCAPE\_URI、NGX\_UNESCAPE\_REDIRECT这三个值。如果是0，则表示src中的所有字符都要进行转码。如果是NGX\_UNESCAPE\_URI与NGX\_UNESCAPE\_REDIRECT，则遇到\'?\'后就结束了，后面的字符就不管了。而NGX\_UNESCAPE\_URI与NGX\_UNESCAPE\_REDIRECT之间的区别是NGX\_UNESCAPE\_URI对于遇到的需要转码的字符，都会转码，而NGX\_UNESCAPE\_REDIRECT则只会对非可见字符进行转码。

uintptr\_t ngx\_escape\_html(u\_char \*dst, u\_char \*src, size\_t size);

对html标签进行编码。

当然，我这里只介绍了一些常用的api的使用，大家可以先熟悉一下，在实际使用过程中，遇到不明白的，最快最直接的方法就是去看源码，看api的实现或看nginx自身调用api的地方是怎么做的，代码就是最好的文档。

### ngx\_pool\_t(100%)

ngx\_pool\_t是一个非常重要的数据结构，在很多重要的场合都有使用，很多重要的数据结构也都在使用它。那么它究竟是一个什么东西呢？简单的说，它提供了一种机制，帮助管理一系列的资源（如内存，文件等），使得对这些资源的使用和释放统一进行，免除了使用过程中考虑到对各种各样资源的什么时候释放，是否遗漏了释放的担心。

例如对于内存的管理，如果我们需要使用内存，那么总是从一个ngx\_pool\_t的对象中获取内存，在最终的某个时刻，我们销毁这个ngx\_pool\_t对象，所有这些内存都被释放了。这样我们就不必要对对这些内存进行malloc和free的操作，不用担心是否某块被malloc出来的内存没有被释放。因为当ngx\_pool\_t对象被销毁的时候，所有从这个对象中分配出来的内存都会被统一释放掉。

再比如我们要使用一系列的文件，但是我们打开以后，最终需要都关闭，那么我们就把这些文件统一登记到一个ngx\_pool\_t对象中，当这个ngx\_pool\_t对象被销毁的时候，所有这些文件都将会被关闭。

从上面举的两个例子中我们可以看出，使用ngx\_pool\_t这个数据结构的时候，所有的资源的释放都在这个对象被销毁的时刻，统一进行了释放，那么就会带来一个问题，就是这些资源的生存周期（或者说被占用的时间）是跟ngx\_pool\_t的生存周期基本一致（ngx\_pool\_t也提供了少量操作可以提前释放资源）。从最高效的角度来说，这并不是最好的。比如，我们需要依次使用A，B，C三个资源，且使用完B的时候，A就不会再被使用了，使用C的时候A和B都不会被使用到。如果不使用ngx\_pool\_t来管理这三个资源，那我们可能从系统里面申请A，使用A，然后在释放A。接着申请B，使用B，再释放B。最后申请C，使用C，然后释放C。但是当我们使用一个ngx\_pool\_t对象来管理这三个资源的时候，A，B和C的释放是在最后一起发生的，也就是在使用完C以后。诚然，这在客观上增加了程序在一段时间的资源使用量。但是这也减轻了程序员分别管理三个资源的生命周期的工作。这也就是有所得，必有所失的道理。实际上是一个取舍的问题，要看在具体的情况下，你更在乎的是哪个。

可以看一下在nginx里面一个典型的使用ngx\_pool\_t的场景，对于nginx处理的每个http request, nginx会生成一个ngx\_pool\_t对象与这个http request关联，所有处理过程中需要申请的资源都从这个ngx\_pool\_t对象中获取，当这个http request处理完成以后，所有在处理过程中申请的资源，都将随着这个关联的ngx\_pool\_t对象的销毁而释放。

ngx\_pool\_t相关结构及操作被定义在文件src/core/ngx\_palloc.h|c中。

typedef struct ngx\_pool\_s ngx\_pool\_t;  
  
struct ngx\_pool\_s {  
 ngx\_pool\_data\_t d;  
 size\_t max;  
 ngx\_pool\_t \*current;  
 ngx\_chain\_t \*chain;  
 ngx\_pool\_large\_t \*large;  
 ngx\_pool\_cleanup\_t \*cleanup;  
 ngx\_log\_t \*log;  
};

从ngx\_pool\_t的一般使用者的角度来说，可不用关注ngx\_pool\_t结构中各字段作用。所以这里也不会进行详细的解释，当然在说明某些操作函数的使用的时候，如有必要，会进行说明。

下面我们来分别解释下ngx\_pool\_t的相关操作。

ngx\_pool\_t \*ngx\_create\_pool(size\_t size, ngx\_log\_t \*log);

创建一个初始节点大小为size的pool，log为后续在该pool上进行操作时输出日志的对象。 需要说明的是size的选择，size的大小必须小于等于NGX\_MAX\_ALLOC\_FROM\_POOL，且必须大于sizeof(ngx\_pool\_t)。

选择大于NGX\_MAX\_ALLOC\_FROM\_POOL的值会造成浪费，因为大于该限制的空间不会被用到（只是说在第一个由ngx\_pool\_t对象管理的内存块上的内存，后续的分配如果第一个内存块上的空闲部分已用完，会再分配的）。

选择小于sizeof(ngx\_pool\_t)的值会造成程序崩溃。由于初始大小的内存块中要用一部分来存储ngx\_pool\_t这个信息本身。

当一个ngx\_pool\_t对象被创建以后，该对象的max字段被赋值为size-sizeof(ngx\_pool\_t)和NGX\_MAX\_ALLOC\_FROM\_POOL这两者中比较小的。后续的从这个pool中分配的内存块，在第一块内存使用完成以后，如果要继续分配的话，就需要继续从操作系统申请内存。当内存的大小小于等于max字段的时候，则分配新的内存块，链接在d这个字段（实际上是d.next字段）管理的一条链表上。当要分配的内存块是比max大的，那么从系统中申请的内存是被挂接在large字段管理的一条链表上。我们暂且把这个称之为大块内存链和小块内存链。

void \*ngx\_palloc(ngx\_pool\_t \*pool, size\_t size);

从这个pool中分配一块为size大小的内存。注意，此函数分配的内存的起始地址按照NGX\_ALIGNMENT进行了对齐。对齐操作会提高系统处理的速度，但会造成少量内存的浪费。

void \*ngx\_pnalloc(ngx\_pool\_t \*pool, size\_t size);

从这个pool中分配一块为size大小的内存。但是此函数分配的内存并没有像上面的函数那样进行过对齐。

void \*ngx\_pcalloc(ngx\_pool\_t \*pool, size\_t size);

该函数也是分配size大小的内存，并且对分配的内存块进行了清零。内部实际上是转调用ngx\_palloc实现的。

void \*ngx\_pmemalign(ngx\_pool\_t \*pool, size\_t size, size\_t alignment);

按照指定对齐大小alignment来申请一块大小为size的内存。此处获取的内存不管大小都将被置于大内存块链中管理。

ngx\_int\_t ngx\_pfree(ngx\_pool\_t \*pool, void \*p);

对于被置于大块内存链，也就是被large字段管理的一列内存中的某块进行释放。该函数的实现是顺序遍历large管理的大块内存链表。所以效率比较低下。如果在这个链表中找到了这块内存，则释放，并返回NGX\_OK。否则返回NGX\_DECLINED。

由于这个操作效率比较低下，除非必要，也就是说这块内存非常大，确应及时释放，否则一般不需要调用。反正内存在这个pool被销毁的时候，总归会都释放掉的嘛！

ngx\_pool\_cleanup\_t \*ngx\_pool\_cleanup\_add(ngx\_pool\_t \*p, size\_t size);

ngx\_pool\_t中的cleanup字段管理着一个特殊的链表，该链表的每一项都记录着一个特殊的需要释放的资源。对于这个链表中每个节点所包含的资源如何去释放，是自说明的。这也就提供了非常大的灵活性。意味着，ngx\_pool\_t不仅仅可以管理内存，通过这个机制，也可以管理任何需要释放的资源，例如，关闭文件，或者删除文件等等。下面我们看一下这个链表每个节点的类型:

typedef struct ngx\_pool\_cleanup\_s ngx\_pool\_cleanup\_t;  
typedef void (\*ngx\_pool\_cleanup\_pt)(void \*data);  
  
struct ngx\_pool\_cleanup\_s {  
 ngx\_pool\_cleanup\_pt handler;  
 void \*data;  
 ngx\_pool\_cleanup\_t \*next;  
};

data

: 指明了该节点所对应的资源。

handler

: 是一个函数指针，指向一个可以释放data所对应资源的函数。该函数只有一个参数，就是data。

next

: 指向该链表中下一个元素。

看到这里，ngx\_pool\_cleanup\_add这个函数的用法，我相信大家都应该有一些明白了。但是这个参数size是起什么作用的呢？这个size就是要存储这个data字段所指向的资源的大小，该函数会为data分配size大小的空间。

比如我们需要最后删除一个文件。那我们在调用这个函数的时候，把size指定为存储文件名的字符串的大小，然后调用这个函数给cleanup链表中增加一项。该函数会返回新添加的这个节点。我们然后把这个节点中的data字段拷贝为文件名。把hander字段赋值为一个删除文件的函数（当然该函数的原型要按照void (\*ngx\_pool\_cleanup\_pt)(void \*data)）。

void ngx\_destroy\_pool(ngx\_pool\_t \*pool);

该函数就是释放pool中持有的所有内存，以及依次调用cleanup字段所管理的链表中每个元素的handler字段所指向的函数，来释放掉所有该pool管理的资源。并且把pool指向的ngx\_pool\_t也释放掉了，完全不可用了。

void ngx\_reset\_pool(ngx\_pool\_t \*pool);

该函数释放pool中所有大块内存链表上的内存，小块内存链上的内存块都修改为可用。但是不会去处理cleanup链表上的项目。

### ngx\_array\_t(100%)

ngx\_array\_t是nginx内部使用的数组结构。nginx的数组结构在存储上与大家认知的C语言内置的数组有相似性，比如实际上存储数据的区域也是一大块连续的内存。但是数组除了存储数据的内存以外还包含一些元信息来描述相关的一些信息。下面我们从数组的定义上来详细的了解一下。ngx\_array\_t的定义位于src/core/ngx\_array.c|h里面。

typedef struct ngx\_array\_s ngx\_array\_t;  
struct ngx\_array\_s {  
 void \*elts;  
 ngx\_uint\_t nelts;  
 size\_t size;  
 ngx\_uint\_t nalloc;  
 ngx\_pool\_t \*pool;  
};

elts

: 指向实际的数据存储区域。

nelts

: 数组实际元素个数。

size

: 数组单个元素的大小，单位是字节。

nalloc

: 数组的容量。表示该数组在不引发扩容的前提下，可以最多存储的元素的个数。当nelts增长到达nalloc时，如果再往此数组中存储元素，则会引发数组的扩容。数组的容量将会扩展到原有容量的2倍大小。实际上是分配新的一块内存，新的一块内存的大小是原有内存大小的2倍。原有的数据会被拷贝到新的一块内存中。

pool

: 该数组用来分配内存的内存池。

下面介绍ngx\_array\_t相关操作函数。

ngx\_array\_t \*ngx\_array\_create(ngx\_pool\_t \*p, ngx\_uint\_t n, size\_t size);

创建一个新的数组对象，并返回这个对象。

p

: 数组分配内存使用的内存池；

n

: 数组的初始容量大小，即在不扩容的情况下最多可以容纳的元素个数。

size

: 单个元素的大小，单位是字节。

void ngx\_array\_destroy(ngx\_array\_t \*a);

销毁该数组对象，并释放其分配的内存回内存池。

void \*ngx\_array\_push(ngx\_array\_t \*a);

在数组a上新追加一个元素，并返回指向新元素的指针。需要把返回的指针使用类型转换，转换为具体的类型，然后再给新元素本身或者是各字段（如果数组的元素是复杂类型）赋值。

void \*ngx\_array\_push\_n(ngx\_array\_t \*a, ngx\_uint\_t n);

在数组a上追加n个元素，并返回指向这些追加元素的首个元素的位置的指针。

static ngx\_inline ngx\_int\_t ngx\_array\_init(ngx\_array\_t \*array, ngx\_pool\_t \*pool, ngx\_uint\_t n, size\_t size);

如果一个数组对象是被分配在堆上的，那么当调用ngx\_array\_destroy销毁以后，如果想再次使用，就可以调用此函数。

如果一个数组对象是被分配在栈上的，那么就需要调用此函数，进行初始化的工作以后，才可以使用。

**注意事项:** 由于使用ngx\_palloc分配内存，数组在扩容时，旧的内存不会被释放，会造成内存的浪费。因此，最好能提前规划好数组的容量，在创建或者初始化的时候一次搞定，避免多次扩容，造成内存浪费。

### ngx\_hash\_t(100%)

ngx\_hash\_t是nginx自己的hash表的实现。定义和实现位于src/core/ngx\_hash.h|c中。ngx\_hash\_t的实现也与数据结构教科书上所描述的hash表的实现是大同小异。对于常用的解决冲突的方法有线性探测，二次探测和开链法等。ngx\_hash\_t使用的是最常用的一种，也就是开链法，这也是STL中的hash表使用的方法。

但是ngx\_hash\_t的实现又有其几个显著的特点:

1. ngx\_hash\_t不像其他的hash表的实现，可以插入删除元素，它只能一次初始化，就构建起整个hash表以后，既不能再删除，也不能在插入元素了。
2. ngx\_hash\_t的开链并不是真的开了一个链表，实际上是开了一段连续的存储空间，几乎可以看做是一个数组。这是因为ngx\_hash\_t在初始化的时候，会经历一次预计算的过程，提前把每个桶里面会有多少元素放进去给计算出来，这样就提前知道每个桶的大小了。那么就不需要使用链表，一段连续的存储空间就足够了。这也从一定程度上节省了内存的使用。

从上面的描述，我们可以看出来，这个值越大，越造成内存的浪费。就两步，首先是初始化，然后就可以在里面进行查找了。下面我们详细来看一下。

ngx\_hash\_t的初始化。

ngx\_int\_t ngx\_hash\_init(ngx\_hash\_init\_t \*hinit, ngx\_hash\_key\_t \*names,

ngx\_uint\_t nelts);

首先我们来看一下初始化函数。该函数的第一个参数hinit是初始化的一些参数的一个集合。 names是初始化一个ngx\_hash\_t所需要的所有key的一个数组。而nelts就是key的个数。下面先看一下ngx\_hash\_init\_t类型，该类型提供了初始化一个hash表所需要的一些基本信息。

typedef struct {  
 ngx\_hash\_t \*hash;  
 ngx\_hash\_key\_pt key;  
  
 ngx\_uint\_t max\_size;  
 ngx\_uint\_t bucket\_size;  
  
 char \*name;  
 ngx\_pool\_t \*pool;  
 ngx\_pool\_t \*temp\_pool;  
} ngx\_hash\_init\_t;

hash

: 该字段如果为NULL，那么调用完初始化函数后，该字段指向新创建出来的hash表。如果该字段不为NULL，那么在初始的时候，所有的数据被插入了这个字段所指的hash表中。

key

: 指向从字符串生成hash值的hash函数。nginx的源代码中提供了默认的实现函数ngx\_hash\_key\_lc。

max\_size

: hash表中的桶的个数。该字段越大，元素存储时冲突的可能性越小，每个桶中存储的元素会更少，则查询起来的速度更快。当然，这个值越大，越造成内存的浪费也越大，(实际上也浪费不了多少)。

bucket\_size

: 每个桶的最大限制大小，单位是字节。如果在初始化一个hash表的时候，发现某个桶里面无法存的下所有属于该桶的元素，则hash表初始化失败。

name

: 该hash表的名字。

pool

: 该hash表分配内存使用的pool。

temp\_pool

: 该hash表使用的临时pool，在初始化完成以后，该pool可以被释放和销毁掉。

下面来看一下存储hash表key的数组的结构。

typedef struct {  
 ngx\_str\_t key;  
 ngx\_uint\_t key\_hash;  
 void \*value;  
} ngx\_hash\_key\_t;

key和value的含义显而易见，就不用解释了。key\_hash是对key使用hash函数计算出来的值。 对这两个结构分析完成以后，我想大家应该都已经明白这个函数应该是如何使用了吧。该函数成功初始化一个hash表以后，返回NGX\_OK，否则返回NGX\_ERROR。

void \*ngx\_hash\_find(ngx\_hash\_t \*hash, ngx\_uint\_t key, u\_char \*name, size\_t len);

在hash里面查找key对应的value。实际上这里的key是对真正的key（也就是name）计算出的hash值。len是name的长度。

如果查找成功，则返回指向value的指针，否则返回NULL。

### ngx\_hash\_wildcard\_t(100%)

nginx为了处理带有通配符的域名的匹配问题，实现了ngx\_hash\_wildcard\_t这样的hash表。他可以支持两种类型的带有通配符的域名。一种是通配符在前的，例如："\*.abc.com"，也可以省略掉星号，直接写成".abc.com"。这样的key，可以匹配www.abc.com，qqq.www.abc.com之类的。另外一种是通配符在末尾的，例如："mail.xxx.\*"，请特别注意通配符在末尾的不像位于开始的通配符可以被省略掉。这样的通配符，可以匹配mail.xxx.com、mail.xxx.com.cn、mail.xxx.net之类的域名。

有一点必须说明，就是一个ngx\_hash\_wildcard\_t类型的hash表只能包含通配符在前的key或者是通配符在后的key。不能同时包含两种类型的通配符的key。ngx\_hash\_wildcard\_t类型变量的构建是通过函数ngx\_hash\_wildcard\_init完成的，而查询是通过函数ngx\_hash\_find\_wc\_head或者ngx\_hash\_find\_wc\_tail来做的。ngx\_hash\_find\_wc\_head是查询包含通配符在前的key的hash表的，而ngx\_hash\_find\_wc\_tail是查询包含通配符在后的key的hash表的。

下面详细说明这几个函数的用法。

ngx\_int\_t ngx\_hash\_wildcard\_init(ngx\_hash\_init\_t \*hinit, ngx\_hash\_key\_t \*names,  
 ngx\_uint\_t nelts);

该函数用来构建一个可以包含通配符key的hash表。

hinit

: 构造一个通配符hash表的一些参数的一个集合。关于该参数对应的类型的说明，请参见ngx\_hash\_t类型中ngx\_hash\_init函数的说明。

names

: 构造此hash表的所有的通配符key的数组。特别要注意的是这里的key已经都是被预处理过的。例如："\*.abc.com"或者".abc.com"被预处理完成以后，变成了"com.abc."。而"mail.xxx.\*"则被预处理为"mail.xxx."。为什么会被处理这样？这里不得不简单地描述一下通配符hash表的实现原理。当构造此类型的hash表的时候，实际上是构造了一个hash表的一个"链表"，是通过hash表中的key"链接"起来的。比如：对于"\*.abc.com"将会构造出2个hash表，第一个hash表中有一个key为com的表项，该表项的value包含有指向第二个hash表的指针，而第二个hash表中有一个表项abc，该表项的value包含有指向\*.abc.com对应的value的指针。那么查询的时候，比如查询www.abc.com的时候，先查com，通过查com可以找到第二级的hash表，在第二级hash表中，再查找abc，依次类推，直到在某一级的hash表中查到的表项对应的value对应一个真正的值而非一个指向下一级hash表的指针的时候，查询过程结束。\**这里有一点需要特别注意的，就是names数组中元素的value值低两位bit必须为0（有特殊用途）。如果不满足这个条件，这个hash表查询不出正确结果。*\*

nelts

: names数组元素的个数。

该函数执行成功返回NGX\_OK，否则NGX\_ERROR。

void \*ngx\_hash\_find\_wc\_head(ngx\_hash\_wildcard\_t \*hwc, u\_char \*name, size\_t len);

该函数查询包含通配符在前的key的hash表的。

hwc

: hash表对象的指针。

name

: 需要查询的域名，例如: www.abc.com。

len

: name的长度。

该函数返回匹配的通配符对应value。如果没有查到，返回NULL。

void \*ngx\_hash\_find\_wc\_tail(ngx\_hash\_wildcard\_t \*hwc, u\_char \*name, size\_t len);

该函数查询包含通配符在末尾的key的hash表的。 参数及返回值请参见上个函数的说明。

### ngx\_hash\_combined\_t(100%)

组合类型hash表，该hash表的定义如下：

typedef struct {  
 ngx\_hash\_t hash;  
 ngx\_hash\_wildcard\_t \*wc\_head;  
 ngx\_hash\_wildcard\_t \*wc\_tail;  
} ngx\_hash\_combined\_t;

从其定义显见，该类型实际上包含了三个hash表，一个普通hash表，一个包含前向通配符的hash表和一个包含后向通配符的hash表。

nginx提供该类型的作用，在于提供一个方便的容器包含三个类型的hash表，当有包含通配符的和不包含通配符的一组key构建hash表以后，以一种方便的方式来查询，你不需要再考虑一个key到底是应该到哪个类型的hash表里去查了。

构造这样一组合hash表的时候，首先定义一个该类型的变量，再分别构造其包含的三个子hash表即可。

对于该类型hash表的查询，nginx提供了一个方便的函数ngx\_hash\_find\_combined。

void \*ngx\_hash\_find\_combined(ngx\_hash\_combined\_t \*hash, ngx\_uint\_t key,  
u\_char \*name, size\_t len);

该函数在此组合hash表中，依次查询其三个子hash表，看是否匹配，一旦找到，立即返回查找结果，也就是说如果有多个可能匹配，则只返回第一个匹配的结果。

hash

: 此组合hash表对象。

key

: 根据name计算出的hash值。

name

: key的具体内容。

len

: name的长度。

返回查询的结果，未查到则返回NULL。

### ngx\_hash\_keys\_arrays\_t(100%)

大家看到在构建一个ngx\_hash\_wildcard\_t的时候，需要对通配符的哪些key进行预处理。这个处理起来比较麻烦。而当有一组key，这些里面既有无通配符的key，也有包含通配符的key的时候。我们就需要构建三个hash表，一个包含普通的key的hash表，一个包含前向通配符的hash表，一个包含后向通配符的hash表（或者也可以把这三个hash表组合成一个ngx\_hash\_combined\_t）。在这种情况下，为了让大家方便的构造这些hash表，nginx提供给了此辅助类型。

该类型以及相关的操作函数也定义在src/core/ngx\_hash.h|c里。我们先来看一下该类型的定义。

typedef struct {  
 ngx\_uint\_t hsize;  
  
 ngx\_pool\_t \*pool;  
 ngx\_pool\_t \*temp\_pool;  
  
 ngx\_array\_t keys;  
 ngx\_array\_t \*keys\_hash;  
  
 ngx\_array\_t dns\_wc\_head;  
 ngx\_array\_t \*dns\_wc\_head\_hash;  
  
 ngx\_array\_t dns\_wc\_tail;  
 ngx\_array\_t \*dns\_wc\_tail\_hash;  
} ngx\_hash\_keys\_arrays\_t;

hsize

: 将要构建的hash表的桶的个数。对于使用这个结构中包含的信息构建的三种类型的hash表都会使用此参数。

pool

: 构建这些hash表使用的pool。

temp\_pool

: 在构建这个类型以及最终的三个hash表过程中可能用到临时pool。该temp\_pool可以在构建完成以后，被销毁掉。这里只是存放临时的一些内存消耗。

keys

: 存放所有非通配符key的数组。

keys\_hash

: 这是个二维数组，第一个维度代表的是bucket的编号，那么keys\_hash[i]中存放的是所有的key算出来的hash值对hsize取模以后的值为i的key。假设有3个key,分别是key1,key2和key3假设hash值算出来以后对hsize取模的值都是i，那么这三个key的值就顺序存放在keys\_hash[i][0],keys\_hash[i][1], keys\_hash[i][2]。该值在调用的过程中用来保存和检测是否有冲突的key值，也就是是否有重复。

dns\_wc\_head

: 放前向通配符key被处理完成以后的值。比如："\*.abc.com" 被处理完成以后，变成 "com.abc." 被存放在此数组中。

dns\_wc\_tail

: 存放后向通配符key被处理完成以后的值。比如："mail.xxx.\*" 被处理完成以后，变成 "mail.xxx." 被存放在此数组中。

dns\_wc\_head\_hash

: 该值在调用的过程中用来保存和检测是否有冲突的前向通配符的key值，也就是是否有重复。

dns\_wc\_tail\_hash

: 该值在调用的过程中用来保存和检测是否有冲突的后向通配符的key值，也就是是否有重复。

在定义一个这个类型的变量，并对字段pool和temp\_pool赋值以后，就可以调用函数ngx\_hash\_add\_key把所有的key加入到这个结构中了，该函数会自动实现普通key，带前向通配符的key和带后向通配符的key的分类和检查，并将这些值存放到对应的字段中去， 然后就可以通过检查这个结构体中的keys、dns\_wc\_head、dns\_wc\_tail三个数组是否为空，来决定是否构建普通hash表，前向通配符hash表和后向通配符hash表了（在构建这三个类型的hash表的时候，可以分别使用keys、dns\_wc\_head、dns\_wc\_tail三个数组）。

构建出这三个hash表以后，可以组合在一个ngx\_hash\_combined\_t对象中，使用ngx\_hash\_find\_combined进行查找。或者是仍然保持三个独立的变量对应这三个hash表，自己决定何时以及在哪个hash表中进行查询。

ngx\_int\_t ngx\_hash\_keys\_array\_init(ngx\_hash\_keys\_arrays\_t \*ha, ngx\_uint\_t type);

初始化这个结构，主要是对这个结构中的ngx\_array\_t类型的字段进行初始化，成功返回NGX\_OK。

ha

: 该结构的对象指针。

type

: 该字段有2个值可选择，即NGX\_HASH\_SMALL和NGX\_HASH\_LARGE。用来指明将要建立的hash表的类型，如果是NGX\_HASH\_SMALL，则有比较小的桶的个数和数组元素大小。NGX\_HASH\_LARGE则相反。

ngx\_int\_t ngx\_hash\_add\_key(ngx\_hash\_keys\_arrays\_t \*ha, ngx\_str\_t \*key,  
void \*value, ngx\_uint\_t flags);

一般是循环调用这个函数，把一组键值对加入到这个结构体中。返回NGX\_OK是加入成功。返回NGX\_BUSY意味着key值重复。

ha

: 该结构的对象指针。

key

: 参数名自解释了。

value

: 参数名自解释了。

flags

: 有两个标志位可以设置，NGX\_HASH\_WILDCARD\_KEY和NGX\_HASH\_READONLY\_KEY。同时要设置的使用逻辑与操作符就可以了。NGX\_HASH\_READONLY\_KEY被设置的时候，在计算hash值的时候，key的值不会被转成小写字符，否则会。NGX\_HASH\_WILDCARD\_KEY被设置的时候，说明key里面可能含有通配符，会进行相应的处理。如果两个标志位都不设置，传0。

有关于这个数据结构的使用，可以参考src/http/ngx\_http.c中的ngx\_http\_server\_names函数。

### ngx\_chain\_t(100%)

nginx的filter模块在处理从别的filter模块或者是handler模块传递过来的数据（实际上就是需要发送给客户端的http response）。这个传递过来的数据是以一个链表的形式(ngx\_chain\_t)。而且数据可能被分多次传递过来。也就是多次调用filter的处理函数，以不同的ngx\_chain\_t。

该结构被定义在src/core/ngx\_buf.h|c。下面我们来看一下ngx\_chain\_t的定义。

typedef struct ngx\_chain\_s ngx\_chain\_t;  
  
struct ngx\_chain\_s {  
 ngx\_buf\_t \*buf;  
 ngx\_chain\_t \*next;  
};

就2个字段，next指向这个链表的下个节点。buf指向实际的数据。所以在这个链表上追加节点也是非常容易，只要把末尾元素的next指针指向新的节点，把新节点的next赋值为NULL即可。

ngx\_chain\_t \*ngx\_alloc\_chain\_link(ngx\_pool\_t \*pool);

该函数创建一个ngx\_chain\_t的对象，并返回指向对象的指针，失败返回NULL。

#define ngx\_free\_chain(pool, cl) \  
 cl->next = pool->chain; \  
pool->chain = cl

该宏释放一个ngx\_chain\_t类型的对象。如果要释放整个chain，则迭代此链表，对每个节点使用此宏即可。

**注意:** 对ngx\_chain\_t类型的释放，并不是真的释放了内存，而仅仅是把这个对象挂在了这个pool对象的一个叫做chain的字段对应的chain上，以供下次从这个pool上分配ngx\_chain\_t类型对象的时候，快速的从这个pool->chain上取下链首元素就返回了，当然，如果这个链是空的，才会真的在这个pool上使用ngx\_palloc函数进行分配。

### ngx\_buf\_t(99%)

这个ngx\_buf\_t就是这个ngx\_chain\_t链表的每个节点的实际数据。该结构实际上是一种抽象的数据结构，它代表某种具体的数据。这个数据可能是指向内存中的某个缓冲区，也可能指向一个文件的某一部分，也可能是一些纯元数据（元数据的作用在于指示这个链表的读取者对读取的数据进行不同的处理）。

该数据结构位于src/core/ngx\_buf.h|c文件中。我们来看一下它的定义。

struct ngx\_buf\_s {  
 u\_char \*pos;  
 u\_char \*last;  
 off\_t file\_pos;  
 off\_t file\_last;  
  
 u\_char \*start; /\* start of buffer \*/  
 u\_char \*end; /\* end of buffer \*/  
 ngx\_buf\_tag\_t tag;  
 ngx\_file\_t \*file;  
 ngx\_buf\_t \*shadow;  
  
  
 /\* the buf's content could be changed \*/  
 unsigned temporary:1;  
  
 /\*  
 \* the buf's content is in a memory cache or in a read only memory  
 \* and must not be changed  
 \*/  
 unsigned memory:1;  
  
 /\* the buf's content is mmap()ed and must not be changed \*/  
 unsigned mmap:1;  
  
 unsigned recycled:1;  
 unsigned in\_file:1;  
 unsigned flush:1;  
 unsigned sync:1;  
 unsigned last\_buf:1;  
 unsigned last\_in\_chain:1;  
  
 unsigned last\_shadow:1;  
 unsigned temp\_file:1;  
  
 /\* STUB \*/ int num;  
};

pos

: 当buf所指向的数据在内存里的时候，pos指向的是这段数据开始的位置。

last

: 当buf所指向的数据在内存里的时候，last指向的是这段数据结束的位置。

file\_pos

: 当buf所指向的数据是在文件里的时候，file\_pos指向的是这段数据的开始位置在文件中的偏移量。

file\_last

: 当buf所指向的数据是在文件里的时候，file\_last指向的是这段数据的结束位置在文件中的偏移量。

start

: 当buf所指向的数据在内存里的时候，这一整块内存包含的内容可能被包含在多个buf中(比如在某段数据中间插入了其他的数据，这一块数据就需要被拆分开)。那么这些buf中的start和end都指向这一块内存的开始地址和结束地址。而pos和last指向本buf所实际包含的数据的开始和结尾。

end

: 解释参见start。

tag

: 实际上是一个void\*类型的指针，使用者可以关联任意的对象上去，只要对使用者有意义。

file

: 当buf所包含的内容在文件中时，file字段指向对应的文件对象。

shadow

: 当这个buf完整copy了另外一个buf的所有字段的时候，那么这两个buf指向的实际上是同一块内存，或者是同一个文件的同一部分，此时这两个buf的shadow字段都是指向对方的。那么对于这样的两个buf，在释放的时候，就需要使用者特别小心，具体是由哪里释放，要提前考虑好，如果造成资源的多次释放，可能会造成程序崩溃！

temporary

: 为1时表示该buf所包含的内容是在一个用户创建的内存块中，并且可以被在filter处理的过程中进行变更，而不会造成问题。

memory

: 为1时表示该buf所包含的内容是在内存中，但是这些内容却不能被进行处理的filter进行变更。

mmap

: 为1时表示该buf所包含的内容是在内存中, 是通过mmap使用内存映射从文件中映射到内存中的，这些内容却不能被进行处理的filter进行变更。

recycled

: 可以回收的。也就是这个buf是可以被释放的。这个字段通常是配合shadow字段一起使用的，对于使用ngx\_create\_temp\_buf 函数创建的buf，并且是另外一个buf的shadow，那么可以使用这个字段来标示这个buf是可以被释放的。

in\_file

: 为1时表示该buf所包含的内容是在文件中。

flush

: 遇到有flush字段被设置为1的的buf的chain，则该chain的数据即便不是最后结束的数据（last\_buf被设置，标志所有要输出的内容都完了），也会进行输出，不会受postpone\_output配置的限制，但是会受到发送速率等其他条件的限制。

sync

: 为1时操作这块缓冲区使用同步方式。

last\_buf

: 数据被以多个chain传递给了过滤器，此字段为1表明这是最后一个buf。

last\_in\_chain

: 在当前的chain里面，此buf是最后一个。特别要注意的是last\_in\_chain的buf不一定是last\_buf，但是last\_buf的buf一定是last\_in\_chain的。这是因为数据会被以多个chain传递给某个filter模块。

last\_shadow

: 在创建一个buf的shadow的时候，通常将新创建的一个buf的last\_shadow置为1。

temp\_file

: 由于受到内存使用的限制，有时候一些buf的内容需要被写到磁盘上的临时文件中去，那么这时，就设置此标志。

对于此对象的创建，可以直接在某个ngx\_pool\_t上分配，然后根据需要，给对应的字段赋值。也可以使用定义好的2个宏：

#define ngx\_alloc\_buf(pool) ngx\_palloc(pool, sizeof(ngx\_buf\_t))  
#define ngx\_calloc\_buf(pool) ngx\_pcalloc(pool, sizeof(ngx\_buf\_t))

这两个宏使用类似函数，也是不说自明的。

对于创建temporary字段为1的buf（就是其内容可以被后续的filter模块进行修改），可以直接使用函数ngx\_create\_temp\_buf进行创建。

ngx\_buf\_t \*ngx\_create\_temp\_buf(ngx\_pool\_t \*pool, size\_t size);

该函数创建一个ngx\_buf\_t类型的对象，并返回指向这个对象的指针，创建失败返回NULL。

对于创建的这个对象，它的start和end指向新分配内存开始和结束的地方。pos和last都指向这块新分配内存的开始处，这样，后续的操作可以在这块新分配的内存上存入数据。

pool

: 分配该buf和buf使用的内存所使用的pool。

size

: 该buf使用的内存的大小。

为了配合对ngx\_buf\_t的使用，nginx定义了以下的宏方便操作。

#define ngx\_buf\_in\_memory(b) (b->temporary || b->memory || b->mmap)

返回这个buf里面的内容是否在内存里。

#define ngx\_buf\_in\_memory\_only(b) (ngx\_buf\_in\_memory(b) && !b->in\_file)

返回这个buf里面的内容是否仅仅在内存里，并且没有在文件里。

#define ngx\_buf\_special(b) \  
 ((b->flush || b->last\_buf || b->sync) \  
 && !ngx\_buf\_in\_memory(b) && !b->in\_file)

返回该buf是否是一个特殊的buf，只含有特殊的标志和没有包含真正的数据。

#define ngx\_buf\_sync\_only(b) \  
 (b->sync \  
 && !ngx\_buf\_in\_memory(b) && !b->in\_file && !b->flush && !b->last\_buf)

返回该buf是否是一个只包含sync标志而不包含真正数据的特殊buf。

#define ngx\_buf\_size(b) \  
 (ngx\_buf\_in\_memory(b) ? (off\_t) (b->last - b->pos): \  
 (b->file\_last - b->file\_pos))

返回该buf所含数据的大小，不管这个数据是在文件里还是在内存里。

### ngx\_list\_t(100%)

ngx\_list\_t顾名思义，看起来好像是一个list的数据结构。这样的说法，算对也不算对。因为它符合list类型数据结构的一些特点，比如可以添加元素，实现自增长，不会像数组类型的数据结构，受到初始设定的数组容量的限制，并且它跟我们常见的list型数据结构也是一样的，内部实现使用了一个链表。

那么它跟我们常见的链表实现的list有什么不同呢？不同点就在于它的节点，它的节点不像我们常见的list的节点，只能存放一个元素，ngx\_list\_t的节点实际上是一个固定大小的数组。

在初始化的时候，我们需要设定元素需要占用的空间大小，每个节点数组的容量大小。在添加元素到这个list里面的时候，会在最尾部的节点里的数组上添加元素，如果这个节点的数组存满了，就再增加一个新的节点到这个list里面去。

好了，看到这里，大家应该基本上明白这个list结构了吧？还不明白也没有关系，下面我们来具体看一下它的定义，这些定义和相关的操作函数定义在src/core/ngx\_list.h|c文件中。

typedef struct {  
 ngx\_list\_part\_t \*last;  
 ngx\_list\_part\_t part;  
 size\_t size;  
 ngx\_uint\_t nalloc;  
 ngx\_pool\_t \*pool;  
} ngx\_list\_t;

last

: 指向该链表的最后一个节点。

part

: 该链表的首个存放具体元素的节点。

size

: 链表中存放的具体元素所需内存大小。

nalloc

: 每个节点所含的固定大小的数组的容量。

pool

: 该list使用的分配内存的pool。

好，我们在看一下每个节点的定义。

typedef struct ngx\_list\_part\_s ngx\_list\_part\_t;  
struct ngx\_list\_part\_s {  
 void \*elts;  
 ngx\_uint\_t nelts;  
 ngx\_list\_part\_t \*next;  
};

elts

: 节点中存放具体元素的内存的开始地址。

nelts

: 节点中已有元素个数。这个值是不能大于链表头节点ngx\_list\_t类型中的nalloc字段的。

next

: 指向下一个节点。

我们来看一下提供的一个操作的函数。

ngx\_list\_t \*ngx\_list\_create(ngx\_pool\_t \*pool, ngx\_uint\_t n, size\_t size);

该函数创建一个ngx\_list\_t类型的对象，并对该list的第一个节点分配存放元素的内存空间。

pool

: 分配内存使用的pool。

n

: 每个节点（ngx\_list\_part\_t）固定长度的数组的长度，即最多可以存放的元素个数。

size

: 每个元素所占用的内存大小。

返回值

: 成功返回指向创建的ngx\_list\_t对象的指针，失败返回NULL。

void \*ngx\_list\_push(ngx\_list\_t \*list);

该函数在给定的list的尾部追加一个元素，并返回指向新元素存放空间的指针。如果追加失败，则返回NULL。

static ngx\_inline ngx\_int\_t  
ngx\_list\_init(ngx\_list\_t \*list, ngx\_pool\_t \*pool, ngx\_uint\_t n, size\_t size);

该函数是用于ngx\_list\_t类型的对象已经存在，但是其第一个节点存放元素的内存空间还未分配的情况下，可以调用此函数来给这个list的首节点来分配存放元素的内存空间。

那么什么时候会出现已经有了ngx\_list\_t类型的对象，而其首节点存放元素的内存尚未分配的情况呢？那就是这个ngx\_list\_t类型的变量并不是通过调用ngx\_list\_create函数创建的。例如：如果某个结构体的一个成员变量是ngx\_list\_t类型的，那么当这个结构体类型的对象被创建出来的时候，这个成员变量也被创建出来了，但是它的首节点的存放元素的内存并未被分配。

总之，如果这个ngx\_list\_t类型的变量，如果不是你通过调用函数ngx\_list\_create创建的，那么就必须调用此函数去初始化，否则，你往这个list里追加元素就可能引发不可预知的行为，亦或程序会崩溃!

### ngx\_queue\_t(100%)

ngx\_queue\_t是nginx中的双向链表，在nginx源码目录src/core下面的ngx\_queue.h|c里面。它的原型如下：

typedef struct ngx\_queue\_s ngx\_queue\_t;  
  
struct ngx\_queue\_s {  
 ngx\_queue\_t \*prev;  
 ngx\_queue\_t \*next;  
};

不同于教科书中将链表节点的数据成员声明在链表节点的结构体中，ngx\_queue\_t只是声明了前向和后向指针。在使用的时候，我们首先需要定义一个哨兵节点(对于后续具体存放数据的节点，我们称之为数据节点)，比如：

ngx\_queue\_t free;

接下来需要进行初始化，通过宏ngx\_queue\_init()来实现：

ngx\_queue\_init(&free);

ngx\_queue\_init()的宏定义如下：

#define ngx\_queue\_init(q) \  
 (q)->prev = q; \  
 (q)->next = q

可见初始的时候哨兵节点的 prev 和 next 都指向自己，因此其实是一个空链表。ngx\_queue\_empty()可以用来判断一个链表是否为空，其实现也很简单，就是：

#define ngx\_queue\_empty(h) \  
 (h == (h)->prev)

那么如何声明一个具有数据元素的链表节点呢？只要在相应的结构体中加上一个 ngx\_queue\_t 的成员就行了。比如ngx\_http\_upstream\_keepalive\_module中的ngx\_http\_upstream\_keepalive\_cache\_t：

typedef struct {  
 ngx\_http\_upstream\_keepalive\_srv\_conf\_t \*conf;  
  
 ngx\_queue\_t queue;  
 ngx\_connection\_t \*connection;  
  
 socklen\_t socklen;  
 u\_char sockaddr[NGX\_SOCKADDRLEN];  
} ngx\_http\_upstream\_keepalive\_cache\_t;

对于每一个这样的数据节点，可以通过ngx\_queue\_insert\_head()来添加到链表中，第一个参数是哨兵节点，第二个参数是数据节点，比如：

ngx\_http\_upstream\_keepalive\_cache\_t cache;  
ngx\_queue\_insert\_head(&free, &cache.queue);

相应的几个宏定义如下：

#define ngx\_queue\_insert\_head(h, x) \  
 (x)->next = (h)->next; \  
 (x)->next->prev = x; \  
 (x)->prev = h; \  
 (h)->next = x  
  
#define ngx\_queue\_insert\_after ngx\_queue\_insert\_head  
  
#define ngx\_queue\_insert\_tail(h, x) \  
 (x)->prev = (h)->prev; \  
 (x)->prev->next = x; \  
 (x)->next = h; \  
 (h)->prev = x

ngx\_queue\_insert\_head()和ngx\_queue\_insert\_after()都是往头部添加节点，ngx\_queue\_insert\_tail()是往尾部添加节点。从代码可以看出哨兵节点的 prev 指向链表的尾数据节点，next 指向链表的头数据节点。另外ngx\_queue\_head()和ngx\_queue\_last()这两个宏分别可以得到头节点和尾节点。

那假如现在有一个ngx\_queue\_t \*q 指向的是链表中的数据节点的queue成员，如何得到ngx\_http\_upstream\_keepalive\_cache\_t的数据呢？ nginx提供了ngx\_queue\_data()宏来得到ngx\_http\_upstream\_keepalive\_cache\_t的指针，例如：

ngx\_http\_upstream\_keepalive\_cache\_t \*cache = ngx\_queue\_data(q,  
 ngx\_http\_upstream\_keepalive\_cache\_t,  
 queue);

也许您已经可以猜到ngx\_queue\_data是通过地址相减来得到的：

#define ngx\_queue\_data(q, type, link) \  
 (type \*) ((u\_char \*) q - offsetof(type, link))

另外nginx也提供了ngx\_queue\_remove()宏来从链表中删除一个数据节点，以及ngx\_queue\_add()用来将一个链表添加到另一个链表。

## nginx的配置系统(100%)

nginx的配置系统由一个主配置文件和其他一些辅助的配置文件构成。这些配置文件均是纯文本文件，全部位于nginx安装目录下的conf目录下。

配置文件中以#开始的行，或者是前面有若干空格或者TAB，然后再跟#的行，都被认为是注释，也就是只对编辑查看文件的用户有意义，程序在读取这些注释行的时候，其实际的内容是被忽略的。

由于除主配置文件nginx.conf以外的文件都是在某些情况下才使用的，而只有主配置文件是在任何情况下都被使用的。所以在这里我们就以主配置文件为例，来解释nginx的配置系统。

在nginx.conf中，包含若干配置项。每个配置项由配置指令和指令参数2个部分构成。指令参数也就是配置指令对应的配置值。

### 指令概述

配置指令是一个字符串，可以用单引号或者双引号括起来，也可以不括。但是如果配置指令包含空格，一定要引起来。

### 指令参数

指令的参数使用一个或者多个空格或者TAB字符与指令分开。指令的参数有一个或者多个TOKEN串组成。TOKEN串之间由空格或者TAB键分隔。

TOKEN串分为简单字符串或者是复合配置块。复合配置块即是由大括号括起来的一堆内容。一个复合配置块中可能包含若干其他的配置指令。

如果一个配置指令的参数全部由简单字符串构成，也就是不包含复合配置块，那么我们就说这个配置指令是一个简单配置项，否则称之为复杂配置项。例如下面这个是一个简单配置项：

error\_page 500 502 503 504 /50x.html;

对于简单配置，配置项的结尾使用分号结束。对于复杂配置项，包含多个TOKEN串的，一般都是简单TOKEN串放在前面，复合配置块一般位于最后，而且其结尾，并不需要再添加分号。例如下面这个复杂配置项：

location / {  
 root /home/jizhao/nginx-book/build/html;  
 index index.html index.htm;  
}

### 指令上下文

nginx.conf中的配置信息，根据其逻辑上的意义，对它们进行了分类，也就是分成了多个作用域，或者称之为配置指令上下文。不同的作用域含有一个或者多个配置项。

当前nginx支持的几个指令上下文：

main

: nginx在运行时与具体业务功能（比如http服务或者email服务代理）无关的一些参数，比如工作进程数，运行的身份等。

http

: 与提供http服务相关的一些配置参数。例如：是否使用keepalive，是否使用gzip进行压缩等。

server

: http服务上支持若干虚拟主机。每个虚拟主机一个对应的server配置项，配置项里面包含该虚拟主机相关的配置。在提供mail服务的代理时，也可以建立若干server.每个server通过监听的地址来区分。

location

: http服务中，某些特定的URL对应的一系列配置项。

mail

: 实现email相关的SMTP/IMAP/POP3代理时，共享的一些配置项（因为可能实现多个代理，工作在多个监听地址上）。

指令上下文，可能有包含的情况出现。例如：通常http上下文和mail上下文一定是出现在main上下文里的。在一个上下文里，可能包含另外一种类型的上下文多次。例如：如果http服务，支持了多个虚拟主机，那么在http上下文里，就会出现多个server上下文。

我们来看一个示例配置：

user nobody;  
worker\_processes 1;  
error\_log logs/error.log info;  
  
events {  
 worker\_connections 1024;  
}  
  
http {  
 server {  
 listen 80;  
 server\_name www.linuxidc.com;  
 access\_log logs/linuxidc.access.log main;  
 location / {  
 index index.html;  
 root /var/www/linuxidc.com/htdocs;  
 }  
 }  
  
 server {  
 listen 80;  
 server\_name www.Androidj.com;  
 access\_log logs/androidj.access.log main;  
 location / {  
 index index.html;  
 root /var/www/androidj.com/htdocs;  
 }  
 }  
}  
  
mail {  
 auth\_http 127.0.0.1:80/auth.php;  
 pop3\_capabilities "TOP" "USER";  
 imap\_capabilities "IMAP4rev1" "UIDPLUS";  
  
 server {  
 listen 110;  
 protocol pop3;  
 proxy on;  
 }  
 server {  
 listen 25;  
 protocol smtp;  
 proxy on;  
 smtp\_auth login plain;  
 xclient off;  
 }  
}

在这个配置中，上面提到个五种配置指令上下文都存在。

存在于main上下文中的配置指令如下:

* user
* worker\_processes
* error\_log
* events
* http
* mail

存在于http上下文中的指令如下:

* server

存在于mail上下文中的指令如下：

* server
* auth\_http
* imap\_capabilities

存在于server上下文中的配置指令如下：

* listen
* server\_name
* access\_log
* location
* protocol
* proxy
* smtp\_auth
* xclient

存在于location上下文中的指令如下：

* index
* root

当然，这里只是一些示例。具体有哪些配置指令，以及这些配置指令可以出现在什么样的上下文中，需要参考nginx的使用文档。

## nginx的模块化体系结构

nginx的内部结构是由核心部分和一系列的功能模块所组成。这样划分是为了使得每个模块的功能相对简单，便于开发，同时也便于对系统进行功能扩展。为了便于描述，下文中我们将使用nginx core来称呼nginx的核心功能部分。

nginx提供了web服务器的基础功能，同时提供了web服务反向代理，email服务反向代理功能。nginx core实现了底层的通讯协议，为其他模块和nginx进程构建了基本的运行时环境，并且构建了其他各模块的协作基础。除此之外，或者说大部分与协议相关的，或者应用相关的功能都是在这些模块中所实现的。

## 模块概述

nginx将各功能模块组织成一条链，当有请求到达的时候，请求依次经过这条链上的部分或者全部模块，进行处理。每个模块实现特定的功能。例如，实现对请求解压缩的模块，实现SSI的模块，实现与上游服务器进行通讯的模块，实现与FastCGI服务进行通讯的模块。

有两个模块比较特殊，他们居于nginx core和各功能模块的中间。这两个模块就是http模块和mail模块。这2个模块在nginx core之上实现了另外一层抽象，处理与HTTP协议和email相关协议（SMTP/POP3/IMAP）有关的事件，并且确保这些事件能被以正确的顺序调用其他的一些功能模块。

目前HTTP协议是被实现在http模块中的，但是有可能将来被剥离到一个单独的模块中，以扩展nginx支持SPDY协议。

### 模块的分类

nginx的模块根据其功能基本上可以分为以下几种类型：

event module

: 搭建了独立于操作系统的事件处理机制的框架，及提供了各具体事件的处理。包括ngx\_events\_module， ngx\_event\_core\_module和ngx\_epoll\_module等。nginx具体使用何种事件处理模块，这依赖于具体的操作系统和编译选项。

phase handler

: 此类型的模块也被直接称为handler模块。主要负责处理客户端请求并产生待响应内容，比如ngx\_http\_static\_module模块，负责客户端的静态页面请求处理并将对应的磁盘文件准备为响应内容输出。

output filter

: 也称为filter模块，主要是负责对输出的内容进行处理，可以对输出进行修改。例如，可以实现对输出的所有html页面增加预定义的footbar一类的工作，或者对输出的图片的URL进行替换之类的工作。

upstream

: upstream模块实现反向代理的功能，将真正的请求转发到后端服务器上，并从后端服务器上读取响应，发回客户端。upstream模块是一种特殊的handler，只不过响应内容不是真正由自己产生的，而是从后端服务器上读取的。

load-balancer

: 负载均衡模块，实现特定的算法，在众多的后端服务器中，选择一个服务器出来作为某个请求的转发服务器。

## nginx的请求处理

nginx使用一个多进程模型来对外提供服务，其中一个master进程，多个worker进程。master进程负责管理nginx本身和其他worker进程。

所有实际上的业务处理逻辑都在worker进程。worker进程中有一个函数，执行无限循环，不断处理收到的来自客户端的请求，并进行处理，直到整个nginx服务被停止。

worker进程中，ngx\_worker\_process\_cycle()函数就是这个无限循环的处理函数。在这个函数中，一个请求的简单处理流程如下：

1) 操作系统提供的机制（例如epoll, kqueue等）产生相关的事件。 2) 接收和处理这些事件，如是接受到数据，则产生更高层的request对象。 3) 处理request的header和body。 4) 产生响应，并发送回客户端。 5) 完成request的处理。 6) 重新初始化定时器及其他事件。

### 请求的处理流程

为了让大家更好的了解nginx中请求处理过程，我们以HTTP Request为例，来做一下详细地说明。

从nginx的内部来看，一个HTTP Request的处理过程涉及到以下几个阶段。

1) 初始化HTTP Request（读取来自客户端的数据，生成HTTP Request对象，该对象含有该请求所有的信息）。 2) 处理请求头。 3) 处理请求体。 4) 如果有的话，调用与此请求（URL或者Location）关联的handler。 5) 依次调用各phase handler进行处理。

在这里，我们需要了解一下phase handler这个概念。phase字面的意思，就是阶段。所以phase handlers也就好理解了，就是包含若干个处理阶段的一些handler。

在每一个阶段，包含有若干个handler，再处理到某个阶段的时候，依次调用该阶段的handler对HTTP Request进行处理。

通常情况下，一个phase handler对这个request进行处理，并产生一些输出。通常phase handler是与定义在配置文件中的某个location相关联的。

一个phase handler通常执行以下几项任务：

1) 获取location配置。 2) 产生适当的响应。 3) 发送response header。 4) 发送response body。

当nginx读取到一个HTTP Request的header的时候，nginx首先查找与这个请求关联的虚拟主机的配置。如果找到了这个虚拟主机的配置，那么通常情况下，这个HTTP Request将会经过以下几个阶段的处理（phase handlers）：

NGX\_HTTP\_POST\_READ\_PHASE

: 读取请求内容阶段

NGX\_HTTP\_SERVER\_REWRITE\_PHASE

: Server请求地址重写阶段

NGX\_HTTP\_FIND\_CONFIG\_PHASE

: 配置查找阶段:

NGX\_HTTP\_REWRITE\_PHASE

: Location请求地址重写阶段

NGX\_HTTP\_POST\_REWRITE\_PHASE

: 请求地址重写提交阶段

NGX\_HTTP\_PREACCESS\_PHASE

: 访问权限检查准备阶段

NGX\_HTTP\_ACCESS\_PHASE

: 访问权限检查阶段

NGX\_HTTP\_POST\_ACCESS\_PHASE

: 访问权限检查提交阶段

NGX\_HTTP\_TRY\_FILES\_PHASE

: 配置项try\_files处理阶段

NGX\_HTTP\_CONTENT\_PHASE

: 内容产生阶段

NGX\_HTTP\_LOG\_PHASE

: 日志模块处理阶段

在内容产生阶段，为了给一个request产生正确的响应，nginx必须把这个request交给一个合适的content handler去处理。如果这个request对应的location在配置文件中被明确指定了一个content handler，那么nginx就可以通过对location的匹配，直接找到这个对应的handler，并把这个request交给这个content handler去处理。这样的配置指令包括像perl，flv，proxy\_pass，mp4等。

如果一个request对应的location并没有直接有配置的content handler，那么nginx依次尝试:

1) 如果一个location里面有配置random\_index on，那么随机选择一个文件，发送给客户端。 2) 如果一个location里面有配置index指令，那么发送index指令指明的文件，给客户端。 3) 如果一个location里面有配置autoindex on，那么就发送请求地址对应的服务端路径下的文件列表给客户端。 4) 如果这个request对应的location上有设置gzip\_static on，那么就查找是否有对应的.gz文件存在，有的话，就发送这个给客户端（客户端支持gzip的情况下）。 5) 请求的URI如果对应一个静态文件，static module就发送静态文件的内容到客户端。

内容产生阶段完成以后，生成的输出会被传递到filter模块去进行处理。filter模块也是与location相关的。所有的fiter模块都被组织成一条链。输出会依次穿越所有的filter，直到有一个filter模块的返回值表明已经处理完成。

这里列举几个常见的filter模块，例如：

1) server-side includes。 2) XSLT filtering。 3) 图像缩放之类的。 4) gzip压缩。

在所有的filter中，有几个filter模块需要关注一下。按照调用的顺序依次说明如下：

write

: 写输出到客户端，实际上是写到连接对应的socket上。

postpone

: 这个filter是负责subrequest的，也就是子请求的。

copy

: 将一些需要复制的buf(文件或者内存)重新复制一份然后交给剩余的body filter处理。