Nginx 开发从入门到精通

发布 0.1

taobao server platform

2013 年 01 月 25 日

前言

目录

第 1 章

缘起

nginx 由于出色的性能，在世界范围内受到了越来越多人的关注，在淘宝内部它更是被广泛的使用，众多的开发以及运维同学都迫切的想要了解 nginx 模块的开发以及它的内部原理，但是国内却没有一本关于这方面的书，源于此我们决定自己来写一本。本书的作者为淘宝核心系统服务器平台组的成员，本书写作的思路是从模块开发逐渐过渡到 nginx 原理剖析。书籍的内容会定期在这里更新，欢迎大家提出宝贵意见，不管是本书的内容问题，还是字词错误，都欢迎大家提交 issue(章节标题的左侧有评注按钮)，我们会及时的跟进。

更新历史

日期	描述
2012/03/01	创建目录大纲
2012/03/28	增加了样章
2012/05/25	更新样章
2012/06/08	增加第 5 章
2012/06/11	增加第 4 章
2012/06/26	增加第 6 章 (event module)
2012/06/27	更新第 5 章部分内容
2012/07/04	更新第 6 章 event module 部分内容
2012/07/12	增加第 12 章（请求头读取，subrequest 解析）
2012/08/14	增加第 2 章 (nginx 基础架构及基础概念)
2012/08/14	增加第 2 章 (ngx_str_t 数据结构介绍)
2012/08/17	增加第 7 章 (模块开发高级篇之变量)
2012/08/25	增加第 11 章 (nginx 的启动阶段部分内容)
2012/09/26	增加第 2 章 (ngx_array_t,ngx_hash_t 及 ngx_pool_t 介绍)
2012/10/08	增加第 11 章 (配置解析综述)
2012/10/12	增加第 2 章 (ngx_hash_wildcard_t,ngx_hash_combined_t 及 ngx_hash_keys_arrays_t 介绍)
2012/10/21	增加第 2 章 (ngx_chain_t,ngx_list_t 及 ngx_buf_t 介绍)
2012/11/09	增加第 12 章 (请求体的读取和丢弃解析)
2012/11/24	更新第 2 章 (ngx_buf_t 的部分字段以及其他一些书写错误和表达)
2012/12/18	更新第 11 章 (解析 http 块)
2012/12/10	增加第 3 章的内容
2012/12/28	补充和完善了第 3 章的内容

第 2 章

版权申明

第 3 章

目录

3.1 上篇：nginx 模块开发篇

3.1.1 背景介绍

nginx 历史

使用简介

nginx 特点介绍

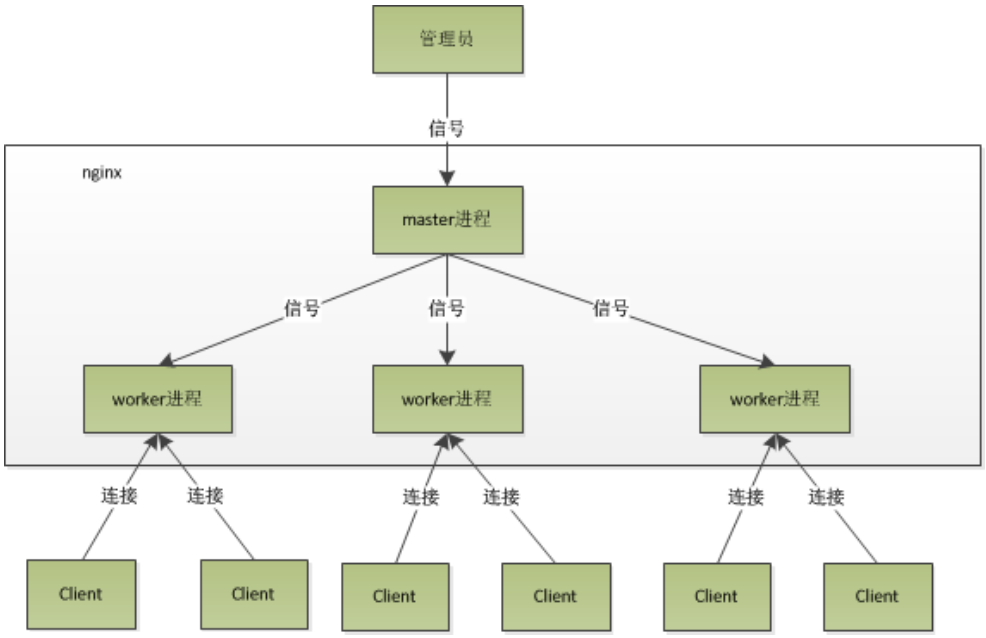
3.1.2 nginx 平台初探 (30%)

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 daemon 模式，让 nginx 在前台运行，这个时候，nginx 就是一个单进程的，很显然，生产环境下我们肯定不会这么做，所以关掉 daemon 的方式，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，然后再启动新的进程，并向所有老的进程发送信号，告诉他们可以光荣退休了。新的进程在启动后，就开始接收新的请求，而老的进程在收到来自 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 之

后，然后再 fork 出多个 worker 进程，这样每个 worker 进程都可以去 accept 这个 socket(当然不是同一个 socket，只是每个进程的这个 socket 会监控在同一个 ip 地址与端口，这个在网络协议里面是允许的)。一般来说，当一个连接进来后，所有在 accept 在这个 socket 上面的进程，都会收到通知，而只有一个进程可以 accept 这个连接，其它的则 accept 失败，这是所谓的惊群现象。当然，nginx 也不会视而不见，所以 nginx 提供了一个 accept_mutex 这个东西，从名字上，我们可以看这是一个加在 accept 上的一把共享锁。有了这把锁之后，同一时刻，就只会有一个进程在 accpet 连接，这样就不会有惊群问题了。accept_mutex 是一个可控选项，我们可以显示地关掉，默认是打开的。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上时，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 进程则很快重新启动新的 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 的是如何处理事件的。

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 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个最小堆里面，每次在进入 epoll_wait 前，先从最小堆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：

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 nginx 进行三次

握手，与 nginx 建立好一个连接后，此时，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。不过，这里我要说的 nginx 对连接数的限制，与 nofile 没有直接关系，可以大于 nofile，也可以小于 nofile。nginx 通过设置 worker_connectons 来设置每个进程可使用的连接最大值。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。

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 是文件协议，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 将会遍历所的有响应头，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)。

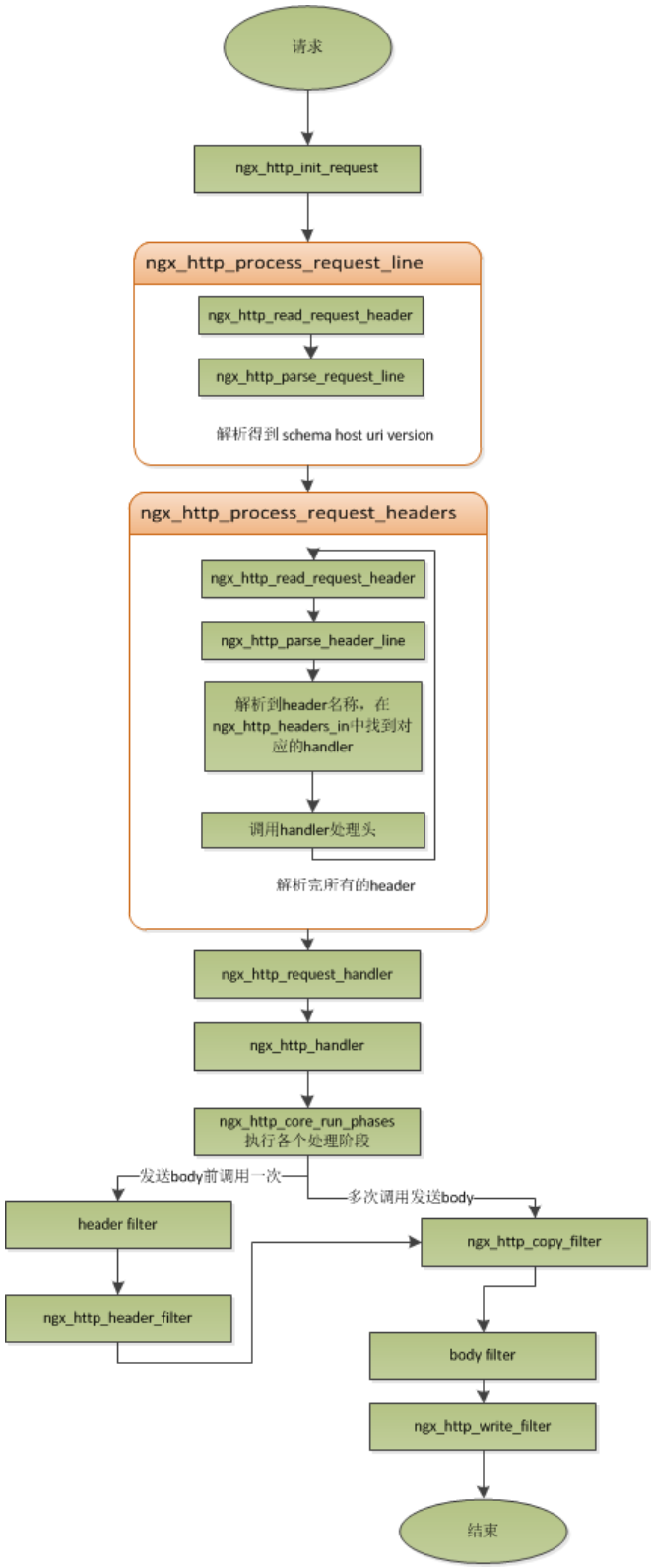
这里要注意的是，nginx 会将整个请求头都放在一个 buffer 里面，这个 buffer 的大小通过配置项 client_header_buffer_size 来设置，如果用户的请求头太大，这个 buffer 装不下，那 nginx 就会重新分配一个新的更大的 buffer 来装请求头，这个大 buffer 可以通过 large_client_header_buffers 来设置，这个 large_buffer 这一组 buffer，比如配置 4 8k，就是表示有四个 8k 大小的 buffer 可以用。注意，为了保存请求行或请求头的完整性，一个完整的请求行或请求头，需要放在一个连续的内存里面，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 请求是某于 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。


请求
ngx_http_init_request
ngx_http_process_request_line
ngx_http_read_request_header
ngx_http_parse_request_line
解析得到 schema host uri version
ngx_http_process_request_headers
ngx_http_read_request_header
ngx_http_parse_header_line
解析到header名称，在ngx_http_headers_in中找到对应的handler
调用handler处理头
解析完所有的header
ngx_http_request_handler
ngx_http_handler
ngx_http_core_run_phases
执行各个处理阶段
发送body前调用一次
多次调用发送body
header filter
ngx_http_header_filter
ngx_http_copy_filter
body filter
ngx_http_write_filter
结束

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。如果的结果为 keepalive，那么，nginx 在输出完响应体后，会设置当前连接的 keepalive 属性，然后再次等待客户端的下一次请求数据。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 直接 keepalive 等待下一次的请求时，会有一个最大等待时间，而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。

如果服务端最后决定的是 keepalive 打开，那么在响应的 http 头里面，也会包含有 connection，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 pipeling 呢？pipeling 其实就是流水线作业，它可以看作为 keepalive 的一种升华，因为 pipeling 也是基于长连接的，目的就是利用一个连接作多次请求。对之前的 keepalive 来说，如果客户端要提交多个请求，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，也就是说，请求是串行进行的，一个请求接一个请求。注意，一个完整的请求，包括发送请求，处理请求，响应请求。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。我们知道，tcp 连接是全双工的，发送与接收可以同时进行，所以，我们可以将多个请求头依次发送出去，在服务端依次处理，客户端再依次接收，这样就多个请求就是同时进行的了。nginx 是直接支持 pipeling 的，但是，nginx 对 pipeling 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，而是再等待一段时间后才真正关掉连接。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。如果客户端正在发送数据，或数据还没有到达服务端，服务端就将连接关掉了。那么，客户端发送的数据会收到 RST 包，此时，客户端对于接收到的服务端的数据，将不会发送 ACK，也就是说，客户端将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 (20%)

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：

从结构体当中，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。正是由于有这样的特性，当你在修改一个字符串时，你就得注意，你修改的字符串是否可以被修改，如果修改后，是否会对其它引用产生影响。在后面介绍 ngx_unescape_uri 函数的时候，就会看到这一点。然后，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。

初始化一个字符串为 str，str 必须为常量字符串，一般只用于声明字符串变量时顺便初始化变量的值。

声明变量时，初始化字符串为空字符串，符串的长度为 0，data 为 NULL。

设置字符串 str 为 text，text 必须为常量字符串。

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 只能用于赋值时初始化，如：

如果这样使用，就会有问题：

这种情况，可以调用 ngx_str_set 与 ngx_str_null 这两个函数来做:

不过要注意的是，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。如：

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。操作不会对原字符串产生变动。如要更改原字符串，可以：

不区分大小写的字符串比较，只比较前 n 个字符。

不区分大小写的不带长度的字符串比较。

区分大小写的不带长度的字符串比较。

区分大小写的带长度的字符串比较，只比较前 n 个字符。

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：

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：

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。

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：

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。

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 对象对销毁的时候，所有从这个对象中分配出来的内存都会被统一释放掉。

在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 的是否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，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 requst 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 requst 处理完成以后，所有在处理过程中申请的资源，都讲随着这个关联的 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。

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。

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) 的值会造成程序奔溃。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，改对象的 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。

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。

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。

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。

对指针 p 指向的一块内存再分配。如果 p 是 NULL，则直接分配一块新的 new_size 大小的内存。

如果 p 不是 NULL, 新分配一块内存，并把旧内存中的内容拷贝至新内存块中，然后释放 p 的旧内存（具体能不能释放旧的，要视具体的情况而定，这里不再详述）。

这个函数实际上也是使用 ngx_palloc 实现的。

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。

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_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，或者删除文件等等的。下面我们看一下这个链表每个节点的类型：

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。该函数的只有一个参数，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。

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)）。

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。

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。

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。

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，即可以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。

销毁该数组对象，并释放其对应的内存给对应的内存池。需要注意的是，调用该函数以后，数组对象上个字段的值并没有被清零。所以即便这个时候对象 a 上各字段还有有意义的值，但是这个对象绝对不应该被再使用了，除非是使用 ngx_array_init 函数。

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。

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。

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: 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课书上所描述的 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，实际上 ngx_hash_t 的使用是非常简单。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。

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。

hash 该字段如果为 NULL，那么调用完初始化韩式有，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，(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 表使用的零时 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。

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。

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。

该函数迎来构建一个可以包含通配符 key 的 hash 表。

hint 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 所对应的值（也就是真正的 value 所在的地址）必须是能被 4 整除的，或者

说是在 4 的倍数的地址上是对齐的。因为这个 value 的值的低两位 bit 是有用的，所以必须为 0。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。

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。

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

ngx_hash_combined_t(100%)

组合类型 hash 表，该 hash 表的定义如下：

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，在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。

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。

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，并将这个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 I 机决定何时以及在哪个 hash 表中进行查询。

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。

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。

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。

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。

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_chaint_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。

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是，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，但是这些内容确不能被进行处理的 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，这些内容确不能被进行处理的 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

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：

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。

该函数创建一个 ngx_but_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。

返回这个 buf 里面的内容是否在内存里。

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。

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。

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。

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。

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。

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。

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 每个节点固定长度的数组的长度。

size 存放的具体元素的个数。

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。

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。

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话，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!

nginx 的配置系统

指令概述

指令参数

指令上下文

nginx 的请求处理

请求的处理流程

nginx 的模块化体系结构

模块概述

模块的分类

3.1.3 handler 模块 (100%)

handler 模块简介

相信大家在看了前一章的模块概述以后，都对 nginx 的模块有了一个基本的认识。基本上作为第三方开发者最可能开发的就是三种类型的模块，即 handler，filter 和 load-balancer。Handler 模块就是接受来自客户端的请求并产生输出的模块。至于有些地方说的 upstream 模块则实际上也是一种 handler。只不过它产生的内容来自于从后端服务器获取的，而非在本机产生的。

当 Nginx 系统启动的时候，每个 handler 都有一次机会把自己关联到一个在配置文件中使用 location 指令配置的一个 location 上。如果有多个 handler 模块都去关联同一个 location，那么实际上只有一个 handler 模块真正会起作用。当然大多数情况下，模块开发人员都会避免出现这种情况。

一个 handler 处理的结果通常有三种情况。处理成功，处理失败（处理的时候发生了错误）或者是拒绝去处理。在拒绝处理的情况下，这个 location 的处理就会由默认的 handler 来进行处理。例如，当在请求一个静态文件的时候，如果你关联一个 handler 到这个 location 上，但是拒绝处理，就会由默认的 ngx_http_static_module 模块进行处理，该模块是一个典型的 handler。

模块的基本结构

在这一节我们将会对通常的模块开发过程中，每个模块所包含的一些常用的部分进行说明。这些部分有些是必须的，有些不是必须的。同时这里所列出的这些东西对于其他类型的模块，例如 filter 模块等也都是相同的。

模块配置结构

基本上每个模块都会提供一些配置指令，以便于用户可以通过配置来控制该模块的行为。那么这些配置信息怎么存储呢？那就需要定义该模块的配置结构来进行存储。

大家都知道 Nginx 的配置信息分成了几个 scope，这就是 main, server, 以及 location。同样的每个模块提供的配置指令也可以出现在这几个 scope 里。那对于这三个 scope 的配置信息，每个模块就需要定义三个不同的数据结构去进行存储。当然，不是每个模块都会在这三个 scope 都提供配置指令的。那么也就不一定每个模块都需要定义三个数据结构去存储这些配置信息了。视模块的实现而言，需要几个就定义几个。

有一点需要特别注意的就是，在模块的开发过程中，我们最好使用 nginx 原有的命名习惯。这样跟原代码的契合度更高，看起来也更舒服。

对于模块配置信息的定义，命名习惯是 ngx_http_<module name>_(main|srv|loc)_conf_t。这里有个例子，就是从我们后面将要展示给大家的 hello module 中截取的。

模块配置指令

一个模块的配置指令是定义在一个静态数组中的。同样地，我们来看一下从 hello module 中截取的模块配置指令的定义。

其实看这个定义，就基本能看出来一些信息。例如，我们是定义了两个配置指令，一个是叫 hello_string，可以接受一个参数，或者是没有参数。另外一个是 hello_counter 的参数。除此之外，似乎看起来有点迷惑。没有关系，我们来详细看一下 ngx_command_t，一旦我们了解这个结构的详细信息，那么我相信上述这个定义所表达的所有信息就不言自明了。

ngx_command_t 的定义，位于 src/core/ngx_conf_file.h 中。

name 配置指令的名称。

type 该配置的类型，其实更准确一点说，是该配置指令属性的集合。nginx 提供了很多预定义的属性值（一些宏定义），通过逻辑或运算符可组合在一起，形成对这个配置指令的详细的说明。下面列出可在这里使用的预定义属性值及说明。

- NGX_CONF_NOARGS：配置指令不接受任何参数。
- NGX_CONF_TAKE1：配置指令接受 1 个参数。
- NGX_CONF_TAKE2：配置指令接受 2 个参数。
- NGX_CONF_TAKE3：配置指令接受 3 个参数。
- NGX_CONF_TAKE4：配置指令接受 4 个参数。
- NGX_CONF_TAKE5：配置指令接受 5 个参数。
- NGX_CONF_TAKE6：配置指令接受 6 个参数。
- NGX_CONF_TAKE7：配置指令接受 7 个参数。

可以组合多个属性，比如一个指令即可以不填参数，也可以接受 1 个或者 2 个参数。那么就是 NGX_CONF_NOARGS|NGX_CONF_TAKE1|NGX_CONF_TAKE2。如果写上面三个属性在一起，你觉得麻烦，那么没有关系，nginx 提供了一些定义，使用起来更简洁。

- NGX_CONF_TAKE12：配置指令接受 1 个或者 2 个参数。
- NGX_CONF_TAKE13：配置指令接受 1 个或者 3 个参数。
- NGX_CONF_TAKE23：配置指令接受 2 个或者 3 个参数。
- NGX_CONF_TAKE123：配置指令接受 1 个或者 2 个或者 3 参数。
- NGX_CONF_TAKE1234：配置指令接受 1 个或者 2 个或者 3 个或者 4 个参数。
- NGX_CONF_1MORE：配置指令接受至少一个参数。
- NGX_CONF_2MORE：配置指令接受至少两个参数。
- NGX_CONF_MULTI: 配置指令可以接受多个参数，即个数不定。
- NGX_CONF_BLOCK：配置指令可以接受的值是一个配置信息块。也就是一对大括号括起来的内容。里面可以再包括很多的配置指令。比如常见的 server 指令就是这个属性的。
- NGX_CONF_FLAG：配置指令可以接受的值是''on'' 或者''off''，最终会被转成 bool 值。
- NGX_CONF_ANY：配置指令可以接受的任意的参数值。一个或者多个，或者''on'' 或者''off''，或者是配置块。

最后要说明的是，无论如何，nginx 的配置指令的参数个数不可以超过 NGX_CONF_MAX_ARGS 个。目前这个值被定义为 8，也就是不能超过 8 个参数值。

下面介绍一组说明配置指令可以出现的位置的属性。

- NGX_DIRECT_CONF：可以出现在配置文件中最外层。例如已经提供的配置指令 daemon，master_process 等。
- NGX_MAIN_CONF: http、mail、events、error_log 等。
- NGX_ANY_CONF: 该配置指令可以出现在任意配置级别上。

对于我们编写的大多数模块而言，都是在处理 http 相关的事情，也就是所谓的都是 NGX_HTTP_MODULE，对于这样类型的模块，其配置可能出现的位置也是分为直接出现在 http 里面，以及其他位置。

- NGX_HTTP_MAIN_CONF: 可以直接出现在 http 配置指令里。
- NGX_HTTP_SRV_CONF: 可以出现在 http 里面的 server 配置指令里。
- NGX_HTTP_LOC_CONF: 可以出现在 http 里面的 location 配置指令里。
- NGX_HTTP_UPS_CONF: 可以出现在 http 里面的 upstream 配置指令里。
- NGX_HTTP_SIF_CONF: 可以出现在 http 里面的 server 配置指令里的 if 语句所在的 block 中。
- NGX_HTTP_LIF_CONF: 可以出现在 http 里面的 limit_except 指令的 block 中。

set 这是一个函数指针，当 nginx 在解析配置的时候，如果遇到这个配置指令，将会把读取到的值传递给这个函数进行分解处理。因为具体每个配置指令的值如何处理，只有定义这个配置指令的人是最清楚的。来看一些这个函数指针要求的函数原型。

先看该函数的返回值，处理成功时，返回 NGX_OK，否则返回 NGX_CONF_ERROR 或者是一个自定义的错误信息的字符串。

在看一下这个函数被调用的时候，传入的三个参数。

- cf: 该参数里面保存里读取到的配置信息的原始字符串以及相关的一些信息。特别注意的是这个参数的 args 字段是一个 ngx_str_t 类型的数组，每个数组元素。该数组的首个元素是这个配置指令本身的字符串，第二个元素是首个参数，第三个元素是第二个参数，依次类推。
- cmd: 这个配置指令对应的 ngx_command_t 结构。
- conf: 就是定义的存储这个配置值的结构体，比如在上面展示的那个 ngx_http_hello_loc_conf_t。当解析这个 hello_string 变量的时候，传入的 conf 就指向一个 ngx_http_hello_loc_conf_t 类型的变量。用户在处理的时候可以使用类型转换，转换成自己知道的类型，再进行字段的赋值。

为了更加方便的实现对配置指令参数的读取，nginx 已经默认提供了对一些标准类型的参数进行读取的函数，可以直接赋值个 set 字段使用。下面来看一下这些已经实现的 set 类型函数。

- ngx_conf_set_flag_slot：读取 NGX_CONF_FLAG 类型的参数。
- ngx_conf_set_str_slot: 读取字符串类型的参数。
- ngx_conf_set_str_array_slot: 读取字符串数组类型的参数。
- ngx_conf_set_keyval_slot：读取键值对类型的参数。
- ngx_conf_set_num_slot: 读取整数类型 (有符号整数 ngx_int_t) 的参数。
- ngx_conf_set_size_slot: 读取 size_t 类型的参数，也就是无符号数。
- ngx_conf_set_off_slot: 读取 off_t 类型的参数。
- ngx_conf_set_msec_slot: 读取毫秒值类型的参数。
- ngx_conf_set_sec_slot: 读取秒值类型的参数。
- ngx_conf_set_bufs_slot：读取的参数值是 2 个，一个是 buf 的个数，一个是 buf 的大小。例如：output_buffers 1 128k;
- ngx_conf_set_enum_slot: 读取枚举类型的参数，将其转换成整数 ngx_uint_t 类型。
- ngx_conf_set_bitmask_slot: 读取参数的值，并将这些参数的值以 bit 位的形式存储。例如：Http-DavModule 模块的 dav_methods 指令。

conf 该字段被 NGX_HTTP_MODULE 类型模块所用 (我们编写的基本上都是 NGX_HTTP_MOUDLE，只有一些 nginx 核心模块是非 NGX_HTTP_MODULE)，该字段指定当前配置项存储的内存位置。实际上是使用哪个内存池的问题。因为 http 模块对所有 http 模块所要保存的配置信息，划分了 main, server 和 location 三个地方进行存储，每个地方都有一个内存池用来分

配存储这些信息的内存。这里可能的值为 NGX_HTTP_MAIN_CONF_OFFSET、NGX_HTTP_SRV_CONF_OFFSET 或 NGX_HTTP_LOC_CONF_OFFSET。当然也可以直接置为 0，就是 NGX_HTTP_MAIN_CONF_OFFSET。

offset 指定该配置项值的精确存放位置，一般指定为某一个结构体变量的字段偏移。因为对于配置信息的存储，一般我们都是定义个结构体来存储的。那么比如我们定义了一个结构体 A，该项配置的值需要存储到该结构体的 b 字段。那么在这里就可以填写为 offsetof(A, b)。对于有些配置项，它的值不需要保存或者是需要保存到更为复杂的结构中时，这里可以设置为 0。

post 该字段存储一个指针。可以指向任何一个在读取配置过程中需要的数据，以便于进行配置读取的处理。大多数时候，都不需要，所以简单地设为 0 即可。

看到这里，应该就比较清楚了。ngx_http_hello_commands 这个数组每 5 个元素为一组，用来描述一个配置项的所有情况。那么如果有多个配置项，只要按照需要再增加 5 个对应的元素对新的配置项进行说明。

需要注意的是，就是在 ngx_http_hello_commands 这个数组定义的最后，都要加一个 ngx_null_command 作为结尾。

模块上下文结构

这是一个 ngx_http_module_t 类型的静态变量。这个变量实际上是提供一组回调函数指针，这些函数有在创建存储配置信息的对象的函数，也有在创建前和创建后会调用的函数。这些函数都将被 nginx 在合适的时间进行调用。

preconfiguration 在创建和读取该模块的配置信息之前被调用。

postconfiguration 在创建和读取该模块的配置信息之后被调用。

create_main_conf 调用该函数创建本模块位于 http block 的配置信息存储结构。该函数成功的时候，返回创建的配置对象。失败的话，返回 NULL。

init_main_conf 调用该函数初始化本模块位于 http block 的配置信息存储结构。该函数成功的时候，返回 NGX_CONF_OK。失败的话，返回 NGX_CONF_ERROR 或错误字符串。

create_srv_conf 调用该函数创建本模块位于 http server block 的配置信息存储结构，每个 server block 会创建一个。该函数成功的时候，返回创建的配置对象。失败的话，返回 NULL。

merge_srv_conf 因为有些配置指令即可以出现在 http block，也可以出现在 http server block 中。那么遇到这种情况，每个 server 都会有自己存储结构来存储该 server 的配置，但是在这种情况下当在 http block 中的配置与 server block 中的配置信息冲突的时候，就需要调用此函数进行合并，该函数并非必须提供，当预计到绝对不会发生需要合并的情况的时候，就无需提供。当然为了安全期间还是建议提供。该函数成功的时候，返回 NGX_CONF_OK。失败的话，返回 NGX_CONF_ERROR 或错误字符串。

create_loc_conf 调用该函数创建本模块位于 location block 的配置信息存储结构。每个在配置中指明的 location 创建一个。该函数成功的时候，返回创建的配置对象。失败的话，返回 NULL。

merge_loc_conf 与 merge_srv_conf 类似，这个也是进行配置值合并的地方。该函数成功的时候，返回 NGX_CONF_OK。失败的话，返回 NGX_CONF_ERROR 或错误字符串。

Nginx 里面的配置信息都是上下一层层的嵌套的，对于具体某个 location 的话，对于同一个配置，如果自己这里没有定义，那么就使用上层的配置，否则是用自己的配置。

这些配置信息一般默认都应该设为一个未初始化的值，针对这个需求，Nginx 定义了一系列的宏定义来代表个中配置所对应数据类型的未初始化值，如下：

又因为对于配置项的合并，逻辑都类似，也就是前面已经说过的，如果在本层次已经配置了，也就是配置项的值已经被读取进来了（那么这些配置项的值就不会等于上面已经定义的那些 UNSET 的值），就使用本层次的值作为定义合并的结果，否则，使用上层的值，如果上层的值也是这些 UNSET 类的值，那就复制为默认值，否则就是用上层的值作为合并的结果。对于这样类似的操作，Nginx 定义了一些宏操作来做这些事情，我们来看其中一个的定义。

显而易见，这个逻辑确实比较简单，所以其它的宏定义也类似，我们就列具其中的一部分吧。

等等。

下面来看一下 hello 模块的模块上下文的定义，加深一下印象。

注意：这里并没有提供 merge_loc_conf 函数，因为我们这个模块的配置指令已经确定只出现在 NGX_HTTP_LOC_CONF 中这一个 level 上，不会发生需要合并的情况。

模块的定义

对于开发一个模块来说，我们都需要定义一个 ngx_module_t 类型的变量来说明这个模块本身的信息，从某种意义上来说，这是这个模块最重要的一个信息，它告诉了 nginx 这个模块的一些信息，上面定义的配置信息，还有模块上下文信息，都是通过这个结构来告诉 nginx 系统的，也就是加载模块的上层代码，都需要通过定义的这个结构，来获取这些信息。

我们来看一下 hello 模块的模块定义。

模块可以提供一些回调函数给 nginx，当 nginx 在创建进程线程或者结束进程线程时进行调用。但大多数模块在这些时刻并不需奥做写什么事情，所以都简单赋值为 NULL。

handler 模块的基本结构

除了上一节介绍的模块的基本结构以外，handler 模块必须提供一个真正的处理函数，这个函数负责对来自客户端请求的真正处理。这个函数的处理，即可以选择自己直接生成内容，也可以选择拒绝处理，由后续的 handler 去进行处理，或者是选择丢给后续的 filter 进行处理。来看一下这个函数的原型申明。

typedef ngx_int_t (*ngx_http_handler_pt)(ngx_http_request_t *r);

r 是 http 请求。里面包含请求所有的信息，这里不相信说明了，可以参考别的章节的介绍。该函数处理成功返回 NGX_OK，处理发生错误返回 NGX_ERROR，拒绝处理（留给后续的 handler 进行处理）返回 NGX_DECLINE。返回 NGX_OK 也就代表给客户端的响应已经生成好了，否则返回 NGX_OK 就发生错误了。

handler 模块的挂载

按处理阶段挂载

为了更精细地控制对于客户端请求的处理过程，nginx 把这个处理过程划分成了 11 个阶段。他们从前到后，依次列举如下：

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

一般情况下，我们自定义的模块，大多数是挂载在 NGX_HTTP_CONTENT_PHASE 阶段的。挂载的动作一般是现在模块上下文调用的 postconfiguration 函数中。

注意：有几个阶段是特例，它不调用挂载地任何的 handler，也就是你就不用挂载到这几个阶段了：

- NGX_HTTP_FIND_CONFIG_PHASE
- NGX_HTTP_POST_ACCESS_PHASE
- NGX_HTTP_POST_REWRITE_PHASE
- NGX_HTTP_TRY_FILES_PHASE

所以其实真正是有 6 个 phase 你可以去挂载 handler。

挂载的代码如下（摘自 hello module）：

使用这种方式挂载的 handler 也被称为**content phase handlers**。

按需挂载

以这种方式挂载的 handler 也被称为**content handler**。

一个请求进来以后，nginx 按照从 NGX_HTTP_POST_READ_PHASE 开始的阶段，去依次执行每个阶段的所有 handler。等到执行到 NGX_HTTP_CONTENT_PHASE 阶段的时候，如果这个 location 对应的有一个

content handler，那么就去执行这个 content handler。否则去依次执行 NGX_HTTP_CONTENT_PHASE 阶段挂载的所有 content phase handlers，直到某个函数处理返回 NGX_OK 或者 NGX_ERROR。

换句话说，如果某个 location 在处理到 NGX_HTTP_CONTENT_PHASE 阶段的时候，如果有 content handler，那么所有的挂载的 content phase handlers 都不会被执行了。

使用这个方法挂载上去的 handler，必须在 NGX_HTTP_CONTENT_PHASE 阶段才能执行到。如果你想自己的 handler 要被更早的执行到的话，那就不要使用这种挂载方式。

另外要提一下，在什么情况会使用这种方式来挂载。一般就是某个模块如果对某个 location 进行了处理以后，发现符合自己处理的逻辑，而且也没有必要再调用 NGX_HTTP_CONTENT_PHASE 阶段的其它 handler 进行处理的时候，就动态挂载上这个 handler。

好了，下面看一下这种挂载方式的具体代码（摘自 Emiller's Guide To Nginx Module Development）。

handler 的编写步骤

好，到了这里，让我们稍微整理一下思路，回顾一下实现一个 handler 的步骤:

1. 编写模块基本结构。
2. 实现 handler 的挂载函数。
3. 编写 handler 处理函数。

看起来不是那么难，对吧？还是那句老话，世上无难事，只怕有心人!

hello handler 模块

我们在前面已经看到了这个 hello handler module 的部分重要的结构。现在我们完整的介绍一下这个示例模块的功能和代码。

该模块提供了 2 个配置指令，仅可以出现在 location 指令的 block 中。这两个指令是 hello_string, 该参数接受一个参数来设置显示的字符串。如果没有跟参数，那么就使用默认的字符串作为响应字符串。

另一个参数是 hello_counter，如果设置为 on，则会在响应的字符串后面追加 Visited Times: 的字样，以统计请求的次数。

这里有两点注意一下：

1. 对于 flag 类型的配置指令，当值为 off 的时候，使用 ngx_conf_set_flag_slot 函数，会转化为 0，为 on，则转化为非 0。
2. 另外一个是，我提供了 merge_loc_conf 函数，但是却没有设置到模块的上下文定义中。这样有一个缺点，就是如果一个指令没有出现在配置文件中的时候，配置信息中的值，将永远会保持在 create_loc_conf 中的初始化的值。那如果，在类似 create_loc_conf 这样的函数中，对创建出来的配置信息的值，没有设置为合理的值的话，后面用户又没有配置，就会出现问题。

下面来完整的给出 ngx_http_hello_module 模块的完整代码。

通过上面一些介绍，我相信大家都能对整个程序有一个比较好的理解。唯一可能感觉有些理解困难的地方在于 ngx_http_hello_handler 函数里面产生和设置输出。但其实大家在本书的前面的相关章节都可以看到对 ngx_buf_t 和 request 等相关数据结构的说明。如果仔细看了这些地方的说明的话，应该对这里代码的实现就比较容易理解了。因此，这里不再赘述解释。

handler 模块的编译和使用

config 文件的编写

对于开发一个模块，我们是需要把这个模块的 C 代码组织到一个目录里，同时需要编写一个 config 文件。这个 config 文件的内容就是告诉 nginx 的编译脚本，该如何进行编译。我们来看一下 hello handler module 的 config 文件的内容，然后再做解释。

其实文件很简单，几乎不需要做什么解释。大家一看都懂了。唯一需要说明的是，如果这个模块的实现有多个源文件，那么都在 NGX_ADDON_SRCS 这个变量里，依次写进去就可以。

编译

对于模块的编译，nginx 并不像 apache 一样，提供了单独的编译工具，可以在没有 nginx 源代码的情况下来单独编译一个模块的代码。nginx 必须去到 nginx 的源代码目录里，通过 configure 指令的参数，来进行编译。下面看一下 hello module 的 configure 指令：

./configure --prefix=/usr/local/nginx-1.3.1 --add-module=/home/jizhao/open_source/book_module

我写的这个示例模块的代码和 config 文件都放在/home/jizhao/open_source/book_module 这个目录下。所以一切都很明了，也没什么好说的了。

使用

使用一个模块需要根据这个模块定义的配置指令来做。比如我们这个简单的 hello handler module 的使用就很简单。在我的测试服务器的配置文件里，就是在 http 里面的默认的 server 里面加入如下的配置：

当我们访问这个地址的时候, lynx http://127.0.0.1/test 的时候，就可以看到返回的结果。

jizhao Visited Times:1

当然你访问多次，这个次数是会增加的。

部分 handler 模块的分析

http access module

该模块的代码位于 src/http/modules/ngx_http_access_module.c 中。该模块的作用是提供对于特定 host 的客户端的访问控制。可以限定特定 host 的客户端对于服务端全部，或者某个 server，或者是某个 location 的访问。该模块的实现非常简单，总共也就只有几个函数。

对于与配置相关的几个函数都不需要做解释了，需要提一下的是函数 ngx_http_access_init，该函数在实现上把本模块挂载到了 NGX_HTTP_ACCESS_PHASE 阶段的 handler 上，从而使自己的被调用时机发生在了 NGX_HTTP_CONTENT_PHASE 等阶段前。因为进行客户端地址的限制检查，根本不需要等到这么后面。

另外看一下这个模块的主处理函数 ngx_http_access_handler。这个函数的逻辑也非常简单，主要是根据客户端地址的类型，来分别选则 ipv4 类型的处理函数 ngx_http_access_inet 还是 ipv6 类型的处理函数 ngx_http_access_inet6。

而这个两个处理函数内部也非常简单，就是循环检查每个规则，检查是否有匹配的规则，如果有就返回匹配的结果，如果都没有匹配，就默认拒绝。

http static module

从某种程度上来说，此模块可以算的上是“最正宗的”，“最古老”的 content handler。因为本模块的作用就是读取磁盘上的静态文件，并把文件内容作为产生的输出。在 Web 技术发展的早期，只有静态页面，没有服务端脚本来动态生成 HTML 的时候。恐怕开发个 Web 服务器的时候，第一个要开发就是这样一个 content handler。

http static module 的代码位于 srchttpmodulesngx_http_static_module.c 中，总共只有两百多行近三百行。可以说是非常短小。

我们首先来看一下该模块的模块上下文的定义。

是非常的简洁吧，连任何与配置相关的函数都没有。对了，因为该模块没有提供任何配置指令。大家想想也就知道了，这个模块做的事情实在是太简单了，也确实没什么好配置的。唯一需要调用的函数是一个 ngx_http_static_init 函数。好了，来看一下这个函数都干了写什么。

仅仅是挂载这个 handler 到 NGX_HTTP_CONTENT_PHASE 处理阶段。简单吧？

下面我们就看一下这个模块最核心的处理逻辑所在的 ngx_http_static_handler 函数。该函数大概占了这个模块代码量的百分之八九十。

首先是检查 NGX_HTTP_GET|NGX_HTTP_HEAD|NGX_HTTP_POST，对就是客户端的请求类型，就这三种，其他一律 NGX_HTTP_NOT_ALLOWED。

其次是检查请求的 url 的结尾字符是不是斜杠‘/’，如果是说明请求的不是一个文件，给后续的 handler 去处理，比如后续的 ngx_http_autoindex_handler（如果是请求的是一个目录下面，可以列出这个目录的文件），或者是 ngx_http_index_handler（如果请求的路径下面有个默认的 index 文件，直接返回 index 文件的内容）。

然后接下来调用了一个 ngx_http_map_uri_to_path 函数，该函数的作用是把请求的 http 协议的路径转化成一个文件系统的路径。

然后根据转化出来的具体路径，去打开文件，打开文件的时候做了 2 中检查，一种是，如果请求的文件是个 symbol link，根据配置，是否允许符号链接，不允许返回错误。还有一个检查是，如果请求的是一个名称，是一个目录的名字，也返回错误。如果都没有文件，就读取文件，返回内容。其实说返回内容可能不是特别准确，比较准确的说法是，把产生的内容传递给后续的 filter 去处理。

http log module

该模块提供了对于每一个 http 请求进行记录的功能，也就是我们见到的 access.log。当然这个模块对于 log 提供了一些配置指令，使得可以比较方便的定制 access.log。

这个模块的代码位于 src/http/modules/ngx_http_log_module.c，虽然这个模块的代码有接近 1400 行，但是主要的逻辑在于对日志本身格式啊，等细节的处理。我们在这里进行分析主要是关注，如何编写一个 log handler 的问题。

由于 log handler 的时候，拿到的参数也是 requset 这个东西，那么也就意味着我们如果需要，可以好好研究下这个结构，把我们需要的所有信息都记录下来。

对于 log handler，有一点特别需要注意的就是，log handler 是无论如何都会被调用的，就是只要服务端接受到了一个客户端的请求，也就是产生了一个 requset 对象，那么这些个 log handler 的处理函数都会被调用的，就是在释放 requset 的时候被调用的（ngx_http_free_request 函数）。

那么当然绝对不能忘记的就是 log handler 最好，也是建议被挂载在 NGX_HTTP_LOG_PHASE 阶段。因为挂载在其他阶段，有可能在某些情况下被跳过，而没有执行到，导致你的 log 模块记录的信息不全。

还有一点要说明的是，由于 nginx 是允许在某个阶段有多个 handler 模块存在的，根据其处理结果，确定是否要调用下一个 handler。但是对于挂载在 NGX_HTTP_LOG_PHASE 阶段的 handler，则根本不关注这里 handler 的具体处理函数的返回值，所有的都被调用。如下，位于 src/http/ngx_http_request.c 中的 ngx_http_log_request 函数。

3.1.4 过滤模块 (90%)

过滤模块简介 (90%)

执行时间和内容 (90%)

过滤（filter）模块是过滤响应头和内容的模块，可以对回复的头和内容进行处理。它的处理时间在获取回复内容之后，向用户发送请求之前。它的处理过程分为两个阶段，过滤 HTTP 回复的头部和主体，在这两个阶段可以分别对头部和主体进行修改。

在代码中有类似的函数：

就是分别对头部和主体进行过滤的函数。所有模块的响应内容要返回给客户端，都必须调用这两个接口。

执行顺序 (90%)

过滤模块的调用是有顺序的，它的顺序在编译的时候就决定了。控制编译的脚本位于 auto/modules 中，当你编译完 Nginx 以后，可以在 objs 目录下面看到一个 ngx_modules.c 的文件。打开这个文件，有类似的代码：

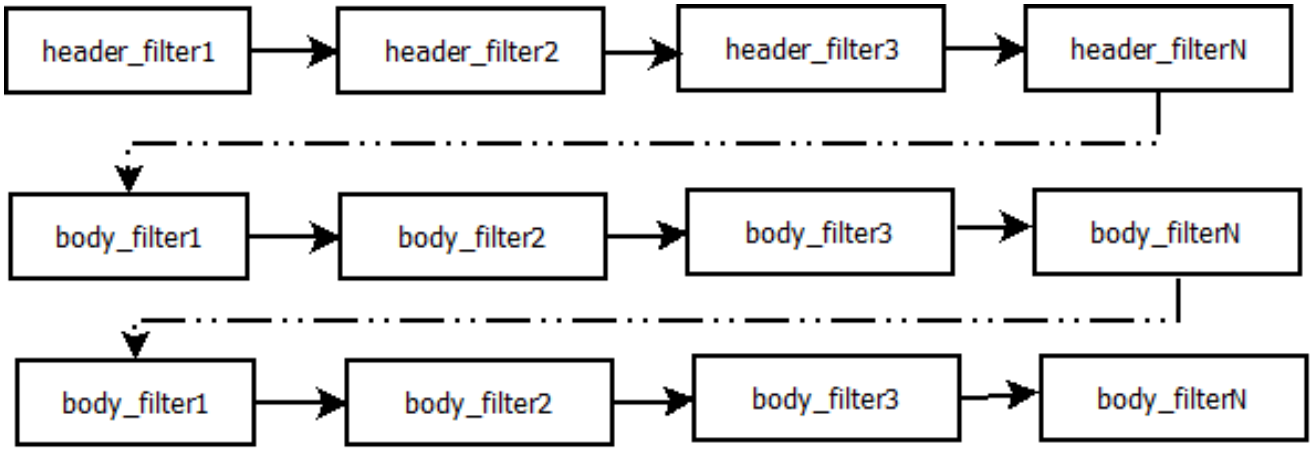
从 write_filter 到 not_modified_filter，模块的执行顺序是反向的。也就是说最早执行的是 not_modified_filter，然后各个模块依次执行。所有第三方的模块只能加入到 copy_filter 和 headers_filter 模块之间执行。

Nginx 执行的时候是怎么按照次序依次来执行各个过滤模块呢？它采用了一种很隐晦的方法，即通过局部的全局变量。比如，在每个 filter 模块，很可能看到如下代码：

ngx_http_top_header_filter 是一个全局变量。当编译进一个 filter 模块的时候，就被赋值为当前 filter 模块的处理函数。而 ngx_http_next_header_filter 是一个局部全局变量，它保存了编译前上一个 filter 模块的处理函数。所以整体看来，就像用全局变量组成的一条单向链表。

每个模块想执行下一个过滤函数，只要调用一下 ngx_http_next_header_filter 这个局部变量。而整个过滤模块链的入口，需要调用 ngx_http_top_header_filter 这个全局变量。ngx_http_top_body_filter 的行为与 header fitler 类似。

响应头和响应体过滤函数的执行顺序如下所示：



这图只表示了 head_filter 和 body_filter 之间的执行顺序，在 header_filter 和 body_filter 处理函数之间，在 body_filter 处理函数之间，可能还有其他执行代码。

模块编译 (90%)

Nginx 可以方便的加入第三方的过滤模块。在过滤模块的目录里，首先需要加入 config 文件，文件的内容如下：

说明把这个名为 ngx_http_example_filter_module 的过滤模块加入，ngx_http_example_filter_module.c 是该模块的源代码。

注意 HTTP_AUX_FILTER_MODULES 这个变量与一般的内容处理模块不同。

过滤模块的分析 (90%)

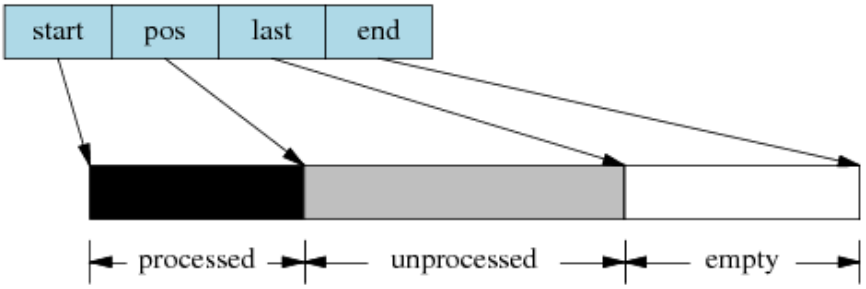
相关结构体 (90%)

ngx_chain_t 结构非常简单，是一个单向链表：

在过滤模块中，所有输出的内容都是通过一条单向链表所组成。这种单向链表的设计，正好应和了 Nginx 流式的输出模式。每次 Nginx 都是读到一部分的内容，就放到链表，然后输出出去。这种设计的好处是简单，非阻塞，但是相应的问题就是跨链表的内容操作非常麻烦，如果需要跨链表，很多时候都只能缓存链表的内容。

单链表负载的就是 ngx_buf_t，这个结构体使用非常广泛，先让我们看下该结构体的代码：

一般 buffer 结构体可以表示一块内存，内存的起始和结束地址分别用 start 和 end 表示，pos 和 last 表示实际的内容。如果内容已经处理过了，pos 的位置就可以往后移动。如果读取到新的内容，last 的位置就会往后移动。所以 buffer 可以在多次调用过程中使用。如果 last 等于 end，就说明这块内存已经用完了。如果 pos 等于 last，说明内存已经处理完了。下面是一个简单的示意图，说明 buffer 中指针的用法：



响应头过滤函数 (90%)

响应头过滤函数主要的用处就是处理 HTTP 响应的头，可以根据实际情况对于响应头进行修改或者添加删除。响应头过滤函数先于响应体过滤函数，而且只调用一次，所以一般可作过滤模块的初始化工作。

响应头过滤函数的入口只有一个：

该函数向客户端发送回复的时候调用，然后按前一节所述的执行顺序。该函数的返回值一般是 NGX_OK，NGX_ERROR 和 NGX_AGAIN，分别表示处理成功，失败和未完成。

你可以把 HTTP 响应头的存储方式想象成一个 hash 表，在 Nginx 内部可以很方便地查找和修改各个响应头部，ngx_http_header_filter_module 过滤模块把所有的 HTTP 头组合成一个完整的 buffer，最终 ngx_http_write_filter_module 过滤模块把 buffer 输出。

按照前一节过滤模块的顺序，依次讲解如下：

filter module	description
ngx_http_not_modified_filter_module	默认打开，如果请求的 if-modified-since 等于回复的 last-modified 间值，说明回复没有变化，清空所有回复的内容，返回 304。
ngx_http_range_body_filter_module	默认打开，只是响应体过滤函数，支持 range 功能，如果请求包含 range 请求，那就只发送 range 请求的一段内容。
ngx_http_copy_filter_module	始终打开，只是响应体过滤函数，主要工作是把文件中内容读到内存中，以便进行处理。
ngx_http_headers_filter_module	始终打开，可以设置 expire 和 Cache-control 头，可以添加任意名称的头
ngx_http_userid_filter_module	默认关闭，可以添加统计用的识别用户的 cookie。
ngx_http_charset_filter_module	默认关闭，可以添加 charset，也可以将内容从一种字符集转换到另外一种字符集，不支持多字节字符集。
ngx_http_ssi_filter_module	默认关闭，过滤 SSI 请求，可以发起子请求，去获取 include 进来的文件
ngx_http_postpone_filter_module	始终打开，用来将子请求和主请求的输出链合并
ngx_http_gzip_filter_module	默认关闭，支持流式的压缩内容
ngx_http_range_header_filter_module	默认打开，此是响应头过滤函数，用来解析 range 头，并产生 range 响应的头。
ngx_http_chunked_filter_module	默认打开，对于 HTTP/1.1 和缺少 content-length 的回复自动打开。
ngx_http_header_filter_module	始终打开，用来将所有 header 组成一个完整的 HTTP 头。
ngx_http_write_filter_module	始终打开，将输出链拷贝到 r->out 中，然后输出内容。

响应体过滤函数 (90%)

响应体过滤函数是过滤响应主体的函数。ngx_http_top_body_filter 这个函数每个请求可能会被执行多次，它的入口函数是 ngx_http_output_filter，比如：

ngx_http_output_filter 可以被一般的静态处理模块调用，也有可能是在 upstream 模块里面被调用，对于整个请求的处理阶段来说，他们处于的用处都是一样的，就是把响应内容过滤，然后发给客户端。

具体模块的响应体过滤函数的格式类似这样：

该函数的返回值一般是 NGX_OK，NGX_ERROR 和 NGX_AGAIN，分别表示处理成功，失败和未完成。

主要功能介绍 (90%)　响应的主体内容就存于单链表 in，链表一般不会太长，有时 in 参数可能为 NULL。in 中存有 buf 结构体中，对于静态文件，这个 buf 大小默认是 32K；对于反向代理的应用，这个 buf 可能是 4k 或者 8k。为了保持内存的低消耗，Nginx 一般不会分配过大的内存，处理的原则是收到一定的数据，就发送出去。一个简单的例子，可以看看 Nginx 的 chunked_filter 模块，在没有 content-length 的情况下，chunk 模块可以流式（stream）的加上长度，方便浏览器接收和显示内容。

在响应体过滤模块中，尤其要注意的是 buf 的标志位，完整描述可以在“相关结构体”这个节中看到。如果 buf 中包含 last 标志，说明是最后一块 buf，可以直接输出并结束请求了。如果有 flush 标志，说明这块 buf 需要马上输出，不能缓存。如果整块 buffer 经过处理完以后，没有数据了，你可以把 buffer 的 sync 标志置上，表示只是同步的用处。

当所有的过滤模块都处理完毕时，在最后的 write_fitler 模块中，Nginx 会将 in 输出链拷贝到 r->out 输出链的末尾，然后调用 sendfile 或者 writev 接口输出。由于 Nginx 是非阻塞的 socket 接口，写操作并不一定会成功，可能会有部分数据还残存在 r->out。在下次的调用中，Nginx 会继续尝试发送，直至成功。

发出子请求 (90%)　Nginx 过滤模块一大特色就是可以发出子请求，也就是在过滤响应内容的时候，你可以发送新的请求，Nginx 会根据你调用的先后顺序，将多个回复的内容拼接成正常的响应主体。一个简单的例子可以参考 addtion 模块。

Nginx 是如何保证父请求和子请求的顺序呢？当 Nginx 发出子请求时，就会调用 ngx_http_subrequest 函数，将子请求插入父请求的 r->postponed 链表中。子请求会在主请求执行完毕时获得依次调用。子请求同样会有一个请求所有的生存期和处理过程，也会进入过滤模块流程。

关键点是在 postpone_filter 模块中，它会拼接主请求和子请求的响应内容。r->postponed 按次序保存有父请求和子请求，它是一个链表，如果前面一个请求未完成，那后一个请求内容就不会输出。当前一个请求完成时并输出时，后一个请求才可输出，当所有的子请求都完成时，所有的响应内容也就输出完毕了。

一些优化措施 (90%)　Nginx 过滤模块涉及到的结构体，主要就是 chain 和 buf，非常简单。在日常的过滤模块中，这两类结构使用非常频繁，Nginx 采用类似 freelist 重复利用的原则，将使用完毕的 chain 或者 buf 结构体，放置到一个固定的空闲链表里，以待下次使用。

比如，在通用内存池结构体中，pool->chain 变量里面就保存着释放的 chain。而一般的 buf 结构体，没有模块间公用的空闲链表池，都是保存在各模块的缓存空闲链表池里面。对于 buf 结构体，还有一种 busy 链表，表示该链表中的 buf 都处于输出状态，如果 buf 输出完毕，这些 buf 就可以释放并重复利用了。

功能	函数名
chain 分配	ngx_alloc_chain_link
chain 释放	ngx_free_chain
buf 分配	ngx_chain_get_free_buf
buf 释放	ngx_chain_update_chains

过滤内容的缓存 (90%)　由于 Nginx 设计流式的输出结构，当我们需要对响应内容作全文过滤的时候，必须缓存部分的 buf 内容。该类过滤模块往往比较复杂，比如 sub，ssi，gzip 等模块。这类模块的设计非常灵活，我简单讲一下设计原则：

1. 输入链 in 需要拷贝操作，经过缓存的过滤模块，输入输出链往往已经完全不一样了，所以需要拷贝，通过 ngx_chain_add_copy 函数完成。
2. 一般有自己的 free 和 busy 缓存链表池，可以提高 buf 分配效率。
3. 如果需要分配大块内容，一般分配固定大小的内存卡，并设置 recycled 标志，表示可以重复利用。
4. 原有的输入 buf 被替换缓存时，必须将其 buf->pos 设为 buf->last，表明原有的 buf 已经被输出完毕。或者在新建立的 buf，将 buf->shadow 指向旧的 buf，以便输出完毕时及时释放旧的 buf。

3.1.5 upstream 模块

upstream 模块 (100%)

nginx 模块一般被分成三大类：handler、filter 和 upstream。前面的章节中，读者已经了解了 handler、filter。利用这两类模块，可以使 nginx 轻松完成任何单机工作。而本章介绍的 upstream，将使 nginx 将跨越单机的限制，完成网络数据的接收、处理和转发。

数据转发功能，为 nginx 提供了跨越单机的横向处理能力，使 nginx 摆脱只能为终端节点提供单一功能的限制，而使它具备了网路应用级别的拆分、封装和整合的战略功能。在云模型大行其道的今天，数据转发使 nginx 有能力构建一个网络应用的关键组件。当然，一个网络应用的关键组件往往一开始都会考虑通过高级开发语言编写，因为开发比较方便，但系统到达一定规模，需要更重视性能的时候，这些高级语言为了达成目标所做的结构化修改所付出的代价会使 nginx 的 upstream 模块就呈现出极大的吸引力，因为他天生就快。作为附带，nginx 的配置提供的层次化和松耦合使得系统的扩展性也可能达到比较高的程度。

言归正传，下面介绍 upstream 的写法。

upstream 模块接口

从本质上说，upstream 属于 handler，只是他不产生自己的内容，而是通过请求后端服务器得到内容，所以才称为 upstream（上游）。请求并取得响应内容的整个过程已经被封装到 nginx 内部，所以 upstream 模块只需要开发若干回调函数，完成构造请求和解析响应等具体的工作。

这些回调函数如下表所示：

create_request	生成发送到后端服务器的请求缓冲（缓冲链）。
reinit_request	在某台后端服务器出错的情况，nginx 会尝试另一台后端服务器。nginx 选定新的服务器以后，会先调用此函数，然后再次调用 create_request，以重新初始化 upstream 模块的工作状态。
process_header	处理后端服务器返回的信息头部。所谓头部是与 upstream server 通信的协议规定的，比如 HTTP 协议的 header 部分，或者 memcached 协议的响应状态部分。
abort_request	在客户端放弃请求时被调用。不需要在函数中实现关闭后端服务器连接的功能，系统会自动完成关闭连接的步骤，所以一般此函数不会进行任何具体工作。
finalize_request	正常完成与后端服务器的请求后调用该函数，与 abort_request 相同，一般也不会进行任何具体工作。
input_filter	处理后端服务器返回的响应正文。nginx 默认的 input_filter 会将收到的内容封装成为缓冲区链 ngx_chain。该链由 upstream 的 out_bufs 指针域定位，所以开发人员可以在模块以外通过该指针得到后端服务器返回的正文数据。memcached 模块实现了自己的 input_filter，在后面会具体分析这个模块。
input_filter_init	初始化 input filter 的上下文。nginx 默认的 input_filter_init 直接返回。

memcached 模块分析

memcache 是一款高性能的分布式 cache 系统，得到了非常广泛的应用。memcache 定义了一套私有通信协议，使得不能通过 HTTP 请求来访问 memcache。但协议本身简单高效，而且 memcache 使用广泛，所以大部分现代开发语言和平台都提供了 memcache 支持，方便开发者使用 memcache。

nginx 提供了 ngx_http_memcached 模块，提供从 memcache 读取数据的功能，而不提供向 memcache 写数据的功能。作为 web 服务器，这种设计是可以接受的。

下面，我们开始分析 ngx_http_memcached 模块，一窥 upstream 的奥秘。

Handler 模块？ 初看 memcached 模块，大家可能觉得并无特别之处。如果稍微细看，甚至觉得有点像 handler 模块，当大家看到这段代码以后，必定疑惑为什么会跟 handler 模块一模一样。

因为 upstream 模块使用的就是 handler 模块的接入方式。同时，upstream 模块的指令系统的设计也是遵循 handler 模块的基本规则：配置该模块才会执行该模块。

所以大家觉得眼熟是好事，说明大家对 Handler 的写法已经很熟悉了。

Upstream 模块！ 那么，upstream 模块的特别之处究竟在哪里呢？答案是就在模块处理函数的实现中。upstream 模块的处理函数进行的操作都包含一个固定的流程。在 memcached 的例子中，可以观察 ngx_http_memcached_handler 的代码，可以发现，这个固定的操作流程是：

1. 创建 upstream 数据结构。
2. 设置模块的 tag 和 schema。schema 现在只会用于日志，tag 会用于 buf_chain 管理。
3. 设置 upstream 的后端服务器列表数据结构。
4. 设置 upstream 回调函数。在这里列出的代码稍稍调整了代码顺序。
5. 创建并设置 upstream 环境数据结构。
6. 完成 upstream 初始化并进行收尾工作。

任何 upstream 模块，简单如 memcached，复杂如 proxy、fastcgi 都是如此。不同的 upstream 模块在这 6 步中的最大差别会出现在第 2、3、4、5 上。其中第 2、4 两步很容易理解，不同的模块设置的标志和使用的回调函数肯定不同。第 5 步也不难理解，只有第 3 步是最为晦涩的，不同的模块在取得后端服务器列表时，策略的差异非常大，有如 memcached 这样简单明了的，也有如 proxy 那样逻辑复杂的。这个问题先记下来，等把 memcached 剖析清楚了，再单独讨论。

第 6 步是一个常态。将 count 加 1，然后返回 NGX_DONE。nginx 遇到这种情况，虽然会认为当前请求的处理已经结束，但是不会释放请求使用的内存资源，也不会关闭与客户端的连接。之所以需要这样，是因为 nginx 建立了 upstream 请求和客户端请求之间一对一的关系，在后续使用 ngx_event_pipe 将 upstream 响应发送回客户端时，还要使用到这些保存着客户端信息的数据结构。这部分会在后面的原理篇做具体介绍，这里不再展开。

将 upstream 请求和客户端请求进行一对一绑定，这个设计有优势也有缺陷。优势就是简化模块开发，可以将精力集中在模块逻辑上，而缺陷同样明显，一对一的设计很多时候都不能满足复杂逻辑的需要。对于这一点，将

会在后面的原理篇来阐述。

回调函数　前面剖析了 memcached 模块的骨架，现在开始逐个解决每个回调函数。

1. ngx_http_memcached_create_request：很简单的按照设置的内容生成一个 key，接着生成一个“get $key”的请求，放在 r->upstream->request_bufs 里面。

2. ngx_http_memcached_reinit_request：无需初始化。

3. ngx_http_memcached_abort_request：无需额外操作。

4. ngx_http_memcached_finalize_request：无需额外操作。

5. ngx_http_memcached_process_header：模块的业务重点函数。memcache 协议将头部信息被定义为第一行文本，可以找到这段代码证明：

如果在已读入缓冲的数据中没有发现 LF(`n') 字符，函数返回 NGX_AGAIN，表示头部未完全读入，需要继续读取数据。nginx 在收到新的数据以后会再次调用该函数。

nginx 处理后端服务器的响应头时只会使用一块缓存，所有数据都在这块缓存中，所以解析头部信息时不需要考虑头部信息跨越多块缓存的情况。而如果头部过大，不能保存在这块缓存中，nginx 会返回错误信息给客户端，并记录 error log，提示缓存不够大。

process_header 的重要职责是将后端服务器返回的状态翻译成返回给客户端的状态。例如，在 ngx_http_memcached_process_header 中，有这样几段代码：

u->state 用于计算 upstream 相关的变量。比如 u->status->status 将被用于计算变量“upstream_status”的值。u->headers_in 将被作为返回给客户端的响应返回状态码。而第一行则是设置返回给客户端的响应的长度。

在这个函数中不能忘记的一件事情是处理完头部信息以后需要将读指针 pos 后移，否则这段数据也将被复制到返回给客户端的响应的正文中，进而导致正文内容不正确。

process_header 函数完成响应头的正确处理，应该返回 NGX_OK。如果返回 NGX_AGAIN，表示未读取完整数据，需要从后端服务器继续读取数据。返回 NGX_DECLINED 无意义，其他任何返回值都被认为是出错状态，nginx 将结束 upstream 请求并返回错误信息。

6. ngx_http_memcached_filter_init：修正从后端服务器收到的内容长度。因为在处理 header 时没有加上这部分长度。

7. ngx_http_memcached_filter：memcached 模块是少有的带有处理正文的回调函数的模块。因为 memcached 模块需要过滤正文末尾 CRLF ``END'' CRLF，所以实现了自己的 filter 回调函数。处理正文的实际意义是将从后端服务器收到的正文有效内容封装成 ngx_chain_t，并加在 u->out_bufs 末尾。nginx 并不进行数据拷贝，而是建立 ngx_buf_t 数据结构指向这些数据内存区，然后由 ngx_chain_t 组织这些 buf。这种实现避免了内存大量搬迁，也是 nginx 高效的奥秘之一。

本节回顾

这一节介绍了 upstream 模块的基本组成。upstream 模块是从 handler 模块发展而来，指令系统和模块生效方式与 handler 模块无异。不同之处在于，upstream 模块在 handler 函数中设置众多回调函数。实际工作都是由这些回调函数完成的。每个回调函数都是在 upstream 的某个固定阶段执行，各司其职，大部分回调函数一般不会真正用到。upstream 最重要的回调函数是 create_request、process_header 和 input_filter，他们共同实现了与后端服务器的协议的解析部分。

负载均衡模块 (100%)

负载均衡模块用于从''upstream'' 指令定义的后端主机列表中选取一台主机。nginx 先使用负载均衡模块找到一台主机，再使用 upstream 模块实现与这台主机的交互。为了方便介绍负载均衡模块，做到言之有物，以下选取 nginx 内置的 ip hash 模块作为实际例子进行分析。

配置

要了解负载均衡模块的开发方法，首先需要了解负载均衡模块的使用方法。因为负载均衡模块与之前书中提到的模块差别比较大，所以我们从配置入手比较容易理解。

在配置文件中，我们如果需要使用 ip hash 的负载均衡算法。我们需要写一个类似下面的配置：

从配置我们可以看出负载均衡模块的使用场景：1. 核心指令''ip_hash'' 只能在 upstream {} 中使用。这条指令用于通知 nginx 使用 ip hash 负载均衡算法。如果没加这条指令，nginx 会使用默认的 round robin 负载均衡模块。请各位读者对比 handler 模块的配置，是不是有共同点？2. upstream {} 中的指令可能出现在''server'' 指令前，可能出现在''server'' 指令后，也可能出现在两条''server'' 指令之间。各位读者可能会有疑问，有什么差别么？那么请各位读者尝试下面这个配置：

神奇的事情出现了：

可见 ip_hash 指令的确能影响到配置的解析。

指令

配置决定指令系统，现在就来看 ip_hash 的指令定义：

没有特别的东西，除了指令属性是 NGX_HTTP_UPS_CONF。这个属性表示该指令的适用范围是 upstream{}。

钩子

以从前面的章节得到的经验，大家应该知道这里就是模块的切入点了。负载均衡模块的钩子代码都是有规律的，这里通过 ip_hash 模块来分析这个规律。

这段代码中有两点值得我们注意。一个是 uscf->flags 的设置，另一个是设置 init_upstream 回调。

设置 uscf->flags

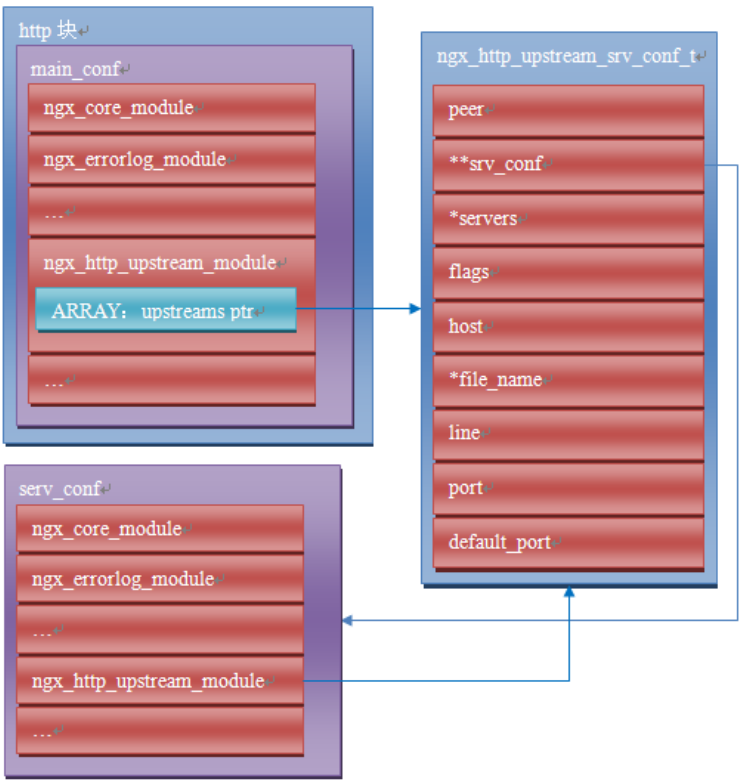
1. NGX_HTTP_UPSTREAM_CREATE：创建标志，如果含有创建标志的话，nginx 会检查重复创建，以及必要参数是否填写；
2. NGX_HTTP_UPSTREAM_MAX_FAILS：可以在 server 中使用 max_fails 属性；
3. NGX_HTTP_UPSTREAM_FAIL_TIMEOUT：可以在 server 中使用 fail_timeout 属性；
4. NGX_HTTP_UPSTREAM_DOWN：可以在 server 中使用 down 属性；

此外还有下面属性：

5. NGX_HTTP_UPSTREAM_WEIGHT：可以在 server 中使用 weight 属性；
6. NGX_HTTP_UPSTREAM_BACKUP：可以在 server 中使用 backup 属性。

聪明的读者如果联想到刚刚遇到的那个神奇的配置错误，可以得出一个结论：在负载均衡模块的指令处理函数中可以设置并修改 upstream{} 中''server'' 指令支持的属性。这是一个很重要的性质，因为不同的负载均衡模块对各种属性的支持情况都是不一样的，那么就需要在解析配置文件的时候检测出是否使用了不支持的负载均衡属性并给出错误提示，这对于提升系统维护性是很有意义的。但是，这种机制也存在缺陷，正如前面的例子所示，没有机制能够追加检查在更新支持属性之前已经配置了不支持属性的''server'' 指令。

设置 init_upstream 回调　nginx 初始化 upstream 时，会在 ngx_http_upstream_init_main_conf 函数中调用设置的回调函数初始化负载均衡模块。这里不太好理解的是 uscf 的具体位置。通过下面的示意图，说明 upstream 负载均衡模块的配置的内存布局。



从图上可以看出，MAIN_CONF 中 ngx_upstream_module 模块的配置项中有一个指针数组 upstreams，数组中的每个元素对应就是配置文件中每一个 upstream{} 的信息。更具体的将会在后面的原理篇讨论。

初始化配置

init_upstream 回调函数执行时需要初始化负载均衡模块的配置，还要设置一个新钩子，这个钩子函数会在 nginx 处理每个请求时作为初始化函数调用，关于这个新钩子函数的功能，后面会有详细的描述。这里，我们先分析 IP hash 模块初始化配置的代码：

这段代码非常简单：IP hash 模块首先调用另一个负载均衡模块 Round Robin 的初始化函数，然后再设置自己的处理请求阶段初始化钩子。实际上几个负载均衡模块可以组成一条链表，每次都是从链首的模块开始进行处理。如果模块决定不处理，可以将处理权交给链表中的下一个模块。这里，IP hash 模块指定 Round Robin 模块作为自己的后继负载均衡模块，所以在自己的初始化配置函数中也对 Round Robin 模块进行初始化。

初始化请求

nginx 收到一个请求以后，如果发现需要访问 upstream，就会执行对应的 peer.init 函数。这是在初始化配置时设置的回调函数。这个函数最重要的作用是构造一张表，当前请求可以使用的 upstream 服务器被依次添加到这张表中。之所以需要这张表，最重要的原因是如果 upstream 服务器出现异常，不能提供服务时，可以从这张表中取得其他服务器进行重试操作。此外，这张表也可以用于负载均衡的计算。之所以构造这张表的行为放在这里而不是在前面初始化配置的阶段，是因为 upstream 需要为每一个请求提供独立隔离的环境。

为了讨论 peer.init 的核心，我们还是看 IP hash 模块的实现：

第一行是设置数据指针，这个指针就是指向前面提到的那张表；

第二行是调用 Round Robin 模块的回调函数对该模块进行请求初始化。面前已经提到，一个负载均衡模块可以调用其他负载均衡模块以提供功能的补充。

第三行是设置一个新的回调函数 get。该函数负责从表中取出某个服务器。除了 get 回调函数，还有另一个 r->upstream->peer.free 的回调函数。该函数在 upstream 请求完成后调用，负责做一些善后工作。比如我们需要维护一个 upstream 服务器访问计数器，那么可以在 get 函数中对其加 1，在 free 中对其减 1。如果是 SSL 的话，nginx 还提供两个回调函数 peer.set_session 和 peer.save_session。一般来说，有两个切入点实现负载均衡算法，其一是在这里，其二是在 get 回调函数中。

peer.get 和 peer.free 回调函数

这两个函数是负载均衡模块最底层的函数，负责实际获取一个连接和回收一个连接的预备操作。之所以说是预备操作，是因为在这两个函数中，并不实际进行建立连接或者释放连接的动作，而只是执行获取连接的地址或维护连接状态的操作。需要理解的清楚一点，在 peer.get 函数中获取连接的地址信息，并不代表这时连接一定没有被建立，相反的，通过 get 函数的返回值，nginx 可以了解是否存在可用连接，连接是否已经建立。这些返回值总结如下：

返回值	说明	nginx 后续动作
NGX_DONE	得到了连接地址信息，并且连接已经建立。	直接使用连接，发送数据。
NGX_OK	得到了连接地址信息，但连接并未建立。	建立连接，如连接不能立即建立，设置事件，暂停执行本请求，执行别的请求。
NGX_BUSY	所有连接均不可用。	返回 502 错误至客户端。

各位读者看到上面这张表，可能会有几个问题浮现出来：

Q 什么时候连接是已经建立的？

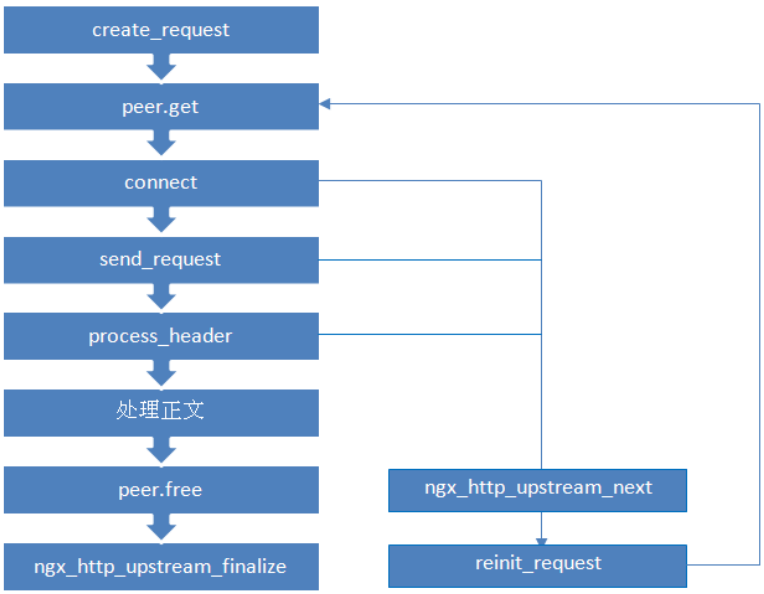
A 使用后端 keepalive 连接的时候，连接在使用完以后并不关闭，而是存放在一个队列中，新的请求只需要从队列中取出连接，这些连接都是已经准备好的。

Q 什么叫所有连接均不可用？

A 初始化请求的过程中，建立了一张表，get 函数负责每次从这张表中不重复的取出一个连接，当无法从表中取得一个新的连接时，即所有连接均不可用。

Q 对于一个请求，peer.get 函数可能被调用多次么？

A 正式如此。当某次 peer.get 函数得到的连接地址连接不上，或者请求对应的服务器得到异常响应，nginx 会执行 ngx_http_upstream_next，然后可能再次调用 peer.get 函数尝试别的连接。upstream 整体流程如下：



本节回顾

这一节介绍了负载均衡模块的基本组成。负载均衡模块的配置区集中在 upstream{} 块中。负载均衡模块的回调函数体系是以 init_upstream 为起点，经历 init_peer，最终到达 peer.get 和 peer.free。其中 init_peer 负责建立每个请求使用的 server 列表，peer.get 负责从 server 列表中选择某个 server（一般是不重复选

择），而 peer.free 负责 server 释放前的资源释放工作。最后，这一节通过一张图将 upstream 模块和负载均衡模块在请求处理过程中的相互关系展现出来。

3.1.6 其他模块 (40%)

Nginx 的模块种类挺多的，除了 HTTP 模块，还有一些核心模块和 mail 系列模块。核心模块主要是做一些基础功能，比如 Nginx 的启动初始化，event 处理机制，错误日志的初始化，ssl 的初始化，正则处理初始化。

mail 模块可以对 imap，pop3，smtp 等协议进行反向代理，这些模块本身不对邮件内容进行处理。

core 模块 (40%)

Nginx 的启动模块 (40%)

启动模块从启动 Nginx 进程开始，做了一系列的初始化工作，源代码位于 src/core/nginx.c，从 main 函数开始:

1. 时间、正则、错误日志、ssl 等初始化
2. 读入命令行参数
3. OS 相关初始化
4. 读入并解析配置
5. 核心模块初始化
6. 创建各种暂时文件和目录
7. 创建共享内存
8. 打开 listen 的端口
9. 所有模块初始化
10. 启动 worker 进程

event 模块 (40%)

event 的类型和功能 (40%)

Nginx 是以 event（事件）处理模型为基础的模块。它为了支持跨平台，抽象出了 event 模块。它支持的 event 处理类型有：AIO（异步 IO），/dev/pool（Solaris 和 Unix 特有），epoll（Linux 特有），eventport（Solaris 10 特有），kqueue（BSD 特有），pool，rtsig（实时信号），select 等。

event 模块的主要功能就是，监听 accept 后建立的连接，对读写事件进行添加删除。事件处理模型和 Nginx 的非阻塞 IO 模型结合在一起使用。当 IO 可读可写的时候，相应的读写时间就会被唤醒，此时就会去处理事件的回调函数。

特别对于 Linux，Nginx 采用的是 epoll 的 EPOLLET（边沿触发）的方法来触发事件，而不是 EPOLLLT（水平触发），所以如果出现了可读事件，进行处理时，必须读取所有的可读数据，否则可能会出现读事件不再触发，连接饿死的情况。

上述是 event 处理抽象出来的关键结构体，可以看到，每个 event 处理模型，都需要实现部分功能。最关键的是 add 和 del 功能，就是最基本的添加和删除事件的函数。

accept 锁 (40%)

Nginx 是多进程程序，80 端口是各进程所共享的，多进程同时 listen 80 端口，势必会产生竞争，也产生了所谓的“惊群”效应。当内核 accept 一个连接时，会唤醒所有等待中的进程，但实际上只有一个进程能获取连接，其他的进程都是被无效唤醒的。所以 Nginx 采用了自有的一套 accept 加锁机制，避免多个进程同时调用 accept。Nginx 多进程的锁在底层默认是通过 CPU 自旋锁来实现。如果操作系统不支持自旋锁，就采用文件锁。

Nginx 事件处理的入口函数是 ngx_process_events_and_timers()，下面是部分代码，可以看到其加锁的过程：

在 ngx_trylock_accept_mutex() 函数里面，如果拿到了锁，Nginx 会把 listen 的端口读事件加入 event 处理，该进程在有新连接进来时就可以进行 accept 了。注意 accept 操作是一个普通的读事件。下面的代码说明了这点：

ngx_process_events() 函数是所有事件处理的入口，它会遍历所有的事件。抢到了 accept 锁的进程跟一般进程稍微不同的是，它的被加上了 NGX_POST_EVENTS 标志，也就是说在 ngx_process_events() 函数里面只接受而不处理事件，并加入 post_events 的队列里面。直到 ngx_accept_mutex 锁去掉以后才去处理具体的事件。为什么这样？因为 ngx_accept_mutex 是全局锁，这样做可以尽量减少该进程抢到锁以后，从 accept 开始到结束的时间，以便其他进程继续接收新的连接，提高吞吐量。

ngx_posted_accept_events 和 ngx_posted_events 就分别是 accept 延迟事件队列和普通延迟事件队列。可以看到 ngx_posted_accept_events 还是放到 ngx_accept_mutex 锁里面处理的。该队列里面处理的都是 accept 事件，它会一口气把内核 backlog 里等待的连接都 accept 进来，注册到读写事件里。

而 ngx_posted_events 是普通的延迟事件队列。一般情况下，什么样的事件会放到这个普通延迟队列里面呢？我的理解是，那些 CPU 耗时比较多的都可以放进去。因为 Nginx 事件处理都是根据触发顺序在一个大循环里依次处理的，因为 Nginx 一个进程同时只能处理一个事件，所以有些耗时多的事件会把后面所有的事件处理都被耽搁了。

除了加锁，Nginx 也对各进程的请求处理的均衡性作了优化，也就是说，如果在负载高的时候，进程抢到的锁过多，会导致这个进程被禁止接受请求一段时间。

比如，在 ngx_event_accept 函数中，有类似代码：

ngx_cycle->connection_n 是进程可以分配的连接总数，ngx_cycle->free_connection_n 是空闲的进程数。上述等式说明了，当前进程的空闲进程数小于 1/8 的话，就会被禁止 accept 一段时间。

定时器 (40%)

Nginx 在需要用到超时的时候，都会用到定时器机制。比如，建立连接以后的那些读写超时。Nginx 使用红黑树来构造定期器，红黑树是一种有序的二叉平衡树，其查找插入和删除的复杂度都为 O(logn)，所以是一种比较理想的二叉树。

定时器的机制就是，二叉树的值是其超时时间，每次查找二叉树的最小值，如果最小值已经过期，就删除该节点，然后继续查找，直到所有超时节点都被删除。

mail 模块

mail 模块的实现

mail 模块的功能

3.1.7 模块开发高级篇 (30%)

变量 (80%)

综述

在 Nginx 中同一个请求需要在模块之间数据的传递或者说在配置文件里面使用模块动态的数据一般来说都是使用变量，比如在 HTTP 模块中导出了 host/remote_addr 等变量，这样我们就可以在配置文件中以及在其他的模块使用这个变量。在 Nginx 中，有两种定义变量的方式，一种是在配置文件中, 使用 set 指令，一种就是上面我们提到的在模块中定义变量，然后导出.

在 Nginx 中所有的变量都是与 HTTP 相关的 (也就是说赋值都是在请求阶段)，并且基本上是同时保存在两个数据结构中，一个就是 hash 表 (可选)，另一个是数组. 比如一些特殊的变量，比如 arg_xxx/cookie_xxx 等，这些变量的名字是不确定的 (因此不能内置)，而且他们还是只读的 (不能交由用户修改)，如果每个都要放到 hash 表中的话 (不知道用户会取多少个), 会很占空间的，因此这些变量就没有 hash, 只有索引. 这里要注意，用户不能定义这样的变量，这样的变量只存在于 Nginx 内部.

对应的变量结构体是这样子 (每一个变量都是一个 ngx_http_variable_s 结构体) 的：

其中 name 表示对应的变量名字，set/get_handler 表示对应的设置以及读取回调，而 data 是传递给回调的参数，flags 表示变量的属性，index 提供了一个索引 (数组的脚标)，从而可以迅速定位到对应的变量。set/get_handler 只有在真正读取设置变量的时候才会被调用.

这里要注意 flag 属性,flag 属性就是由下面的几个属性组合而成:

1. NGX_HTTP_VAR_CHANGEABLE 表示这个变量是可变的, 比如 arg_xxx 这类变量，如果你使用 set 指令来修改，那么 Nginx 就会报错.
2. NGX_HTTP_VAR_NOCACHEABLE 表示这个变量每次都要去取值，而不是直接返回上次 cache 的值 (配合对应的接口).
3. NGX_HTTP_VAR_INDEXED 表示这个变量是用索引读取的.

4. NGX_HTTP_VAR_NOHASH 表示这个变量不需要被 hash.

而变量在 Nginx 中的初始化流程是这样的:

1. 首先当解析 HTTP 之前会调用 ngx_http_variables_add_core_vars(pre_config) 来将 HTTP core 模块导出的变量 (http_host/remote_addr...) 添加进全局的 hash key 链中.
2. 解析完 HTTP 模块之后，会调用 ngx_http_variables_init_vars 来初始化所有的变量 (不仅包括 HTTP core 模块的变量，也包括其他的 HTTP 模块导出的变量, 以及配置文件中使用 set 命令设置的变量), 这里的初始化包括初始化 hash 表, 以及初始化数组索引.
3. 当每次请求到来时会给每个请求创建一个变量数组 (数组的个数就是上面第二步所保存的变量个数)。然后只有取变量值的时候，才会将变量保存在对应的变量数组位置。

创建变量

在 Nginx 中，创建变量有两种方式，分别是在配置文件中使用 set 指令，和在模块中调用对应的接口，在配置文件中创建变量比较简单，因此我们主要来看如何在模块中创建自己的变量。

在 Nginx 中提供了下面的接口，可以供模块调用来创建变量。

这个函数所做的工作就是将变量 ``name'' 添加进全局的 hash key 表中, 然后初始化一些域，不过这里要注意，对应的变量的 get/set 回调，需要当这个函数返回之后，显示的设置, 比如在 split_clients 模块中的例子:

而对应的回调函数原型是这样的:

回调函数比较简单，第一个参数是当前请求，第二个是需要设置或者获取的变量值，第三个是初始化时的回调指针，这里我们着重来看一下 ngx_http_variable_value_t, 下面就是这个结构体的原型:

这里主要是 data 域，当我们在 get_handle 中设置变量值的时候，只需要将对应的值放入到 data 中就可以了，这里 data 需要在 get_handle 中分配内存, 比如下面的例子 (ngx_http_fastcgi_script_name_variable), 就是 fastcgi_script_name 变量的 get_handler 代码片段:

使用变量

Nginx 的内部变量指的就是 Nginx 的官方模块中所导出的变量，在 Nginx 中，大部分常用的变量都是 CORE HTTP 模块导出的。而在 Nginx 中，不仅可以在模块代码中使用变量，而且还可以在配置文件中使用。

假设我们需要在配置文件中使用 http 模块的 host 变量，那么只需要这样在变量名前加一个 $ 符号就可以了 ($host). 而如果需要在模块中使用 host 变量，那么就比较麻烦，Nginx 提供了下面几个接口来取得变量:

他们的区别是这样子的，ngx_http_get_indexed_variable 和 ngx_http_get_flushed_variable 都是用来取得有索引的变量，不过他们的区别是后一个会处理 NGX_HTTP_VAR_NOCACHEABLE 这个标记，也就是说如果你想要 cache 你的变量值，那么你的变量属性就不能设置 NGX_HTTP_VAR_NOCACHEABLE, 并且通过 ngx_http_get_flushed_variable 来获取变量值. 而 ngx_http_get_variable 和上面的区别就是它能够得到没有索引的变量值.

通过上面我们知道可以通过索引来得到变量值，可是这个索引改如何取得呢，Nginx 也提供了对应的接口：

通过这个接口，就可以取得对应变量名的索引值。

接下来来看对应的例子，比如在 http_log 模块中，如果在 log_format 中配置了对应的变量，那么它会调用 ngx_http_get_variable_index 来保存索引：

然后 http_log 模块会使用 ngx_http_get_indexed_variable 来得到对应的变量值, 这里要注意，就是使用这个接口的时候，判断返回值，不仅要判断是否为空，也需要判断 value->not_found, 这是因为只有第一次调用才会返回空，后续返回就不是空，因此需要判断 value->not_found:

upstream

使用 subrequest 访问 upstream

超越 upstream

event 机制

例讲（主动健康检查模块）

使用 lua 模块

3.2 下篇：nginx 原理解析篇

3.2.1 高性能服务器设计

c10k 问题

高性能服务器编写的关键原则

事件驱动的核心引擎

定时器管理

3.2.2 nginx 架构详解

nginx 的源码目录结构

nginx 的 configure 原理

auto 脚本

模块编译顺序

nginx 的事件机制

event 框架及非阻塞模型

定时器实现

信号处理

惊群问题

nginx 的进程机制

master 进程

worker 进程

进程间通讯

3.2.3 nginx 基础设施

内存池

共享内存

在不考虑 nginx 单进城工作的情况下，这个过程包含三种方式：

1. 启动新的 nginx
2. reload 配置
3. 热替换 nginx 代码

三种方式有共同的流程，下面这幅图向我们展现了这个流程：

图 11-1

流程的开端是解析 nginx 配置、初始化模块，接着是初始化文件句柄，初始化共享内存，然后是监听端口，再后来创建 worker 子进程和其他辅助子进程，最后是 worker 初始化事件机制。以上步骤结束以后，nginx 各个子进程开始各司其职，比如 worker 进程开始 accept 请求并按最新配置处理请求，cache-manager 进程开始管理 cache 文件目录等等。

除了这些共同流程，这三种方式的差异也非常明显。第一种方式包含命令行解析的过程，同时输出有一段时间是输出到控制台。reload 配置有两种形式，一种是使用 nginx 命令行，一种是向 master 进程发送 HUP 信号，前者表面上与第一种方式无异，但实际上差别很大，后者则完全不支持控制台输出，无法直接查看 nginx 的启动情况。而且 reload 配置时，nginx 需要自动停止以往生成的子进程，所以还包含复杂的进程管理操作，这一点在启动新的 nginx 的方式中是不存在的。热替换 nginx 代码虽然使用上与 reload 配置的后一种形式相似，但在解析 nginx 配置方面，与 reload 配置的方式差距非常大。另外，热替换 nginx 代码时，对以往创建的子进程管理也不像 reload 配置那样，需要手工触发进行。所以，我们想弄懂 nginx 的启动阶段，就必须理解所有这三种方式下 nginx 都是如何工作的。

共有流程 (100%)

从概述中我们了解到，nginx 启动分为三种方式，虽然各有不同，但也有一段相同的流程。在这一节中，我们对 nginx 启动阶段的共用流程进行讨论。

共有流程的代码主要集中在 ngx_cycle.c、ngx_process.c、ngx_process_cycle.c 和 ngx_event.c 这四个文件中。我们这一节只讨论 nginx 的框架代码，而与 http 相关的模块代码，我们会在后面进行分析。

共有流程开始于解析 nginx 配置，这个过程集中在 ngx_init_cycle 函数中。ngx_init_cycle 是 nginx 的一个核心函数，共有流程中与配置相关的几个过程都在这个函数中实现，其中包括解析 nginx 配置、初始化 CORE 模块，接着是初始化文件句柄，初始化错误日志，初始化共享内存，然后是监听端口。可以说共有流程 80% 都是现在 ngx_init_cycle 函数中。

在具体介绍以前，我们先解决一个概念问题——什么叫 cycle？

cycle 就是周期的意思，对应着一次启动过程。也就是说，不论发生了上节介绍的三种启动方式的哪一种，nginx 都会创建一个新的 cycle 与这次启动对应。

配置解析接口 (100%)

ngx_init_cycle 提供的是配置解析接口。接口是一个切入点，通过少量代码提供一个完整功能的调用。配置解析接口分为两个阶段，一个是准备阶段，另一个就是真正开始调用配置解析。准备阶段指什么呢？主要是准备三

点：

1. 准备内存

nginx 根据以往的经验（old_cycle）预测这一次的配置需要分配多少内存。比如，我们可以看这段：

这段代码的意思是遍历 old_cycle，统计上一次系统中分配了多少块共享内存，接着就按这个数据初始化当前 cycle 中共享内存的规模。

2. 准备错误日志

nginx 启动可能出错，出错就要记录到错误日志中。而错误日志本身也是配置的一部分，所以不解析完配置，nginx 就不能了解错误日志的信息。nginx 通过使用上一个周期的错误日志来记录解析配置时发生的错误，而在配置解析完成以后，nginx 就用新的错误日志替换旧的错误日志。具体代码摘抄如下，以说明 nginx 解析配置时使用 old_cycle 的错误日志：

3. 准备数据结构

主要是两个数据结果，一个是 ngx_cycle_t 结构，一个是 ngx_conf_t 结构。前者用于存放所有 CORE 模块的配置，后者则是用于存放解析配置的上下文信息。具体代码如下：

准备好了这些内容，nginx 开始调用配置解析模块，其代码如下：

第一个 if 解析 nginx 命令行参数'-g' 加入的配置。第二个 if 解析 nginx 配置文件。好的设计就体现在接口极度简化，模块之间的耦合非常低。这里只使用区区 10 行完成了配置的解析。在这里，我们先浅尝辄止，具体 nginx 如何解析配置，我们将在后面的小节做细致的介绍。

配置解析

通用过程 (100%)

配置解析模块在 ngx_conf_file.c 中实现。模块提供的接口函数主要是 ngx_conf_parse，另外，模块提供一个单独的接口 ngx_conf_param，用来解析命令行传递的配置，当然，这个接口也是对 ngx_conf_parse 的包装。

ngx_conf_parse 函数支持三种不同的解析环境：

1. parse_file：解析配置文件；
2. parse_block：解析块配置。块配置一定是由“{”和“}”包裹起来的；
3. parse_param：解析命令行配置。命令行配置中不支持块指令。

我们先来鸟瞰 nginx 解析配置的流程，整个过程可参见下面示意图：

图 11-2

这是一个递归的过程。nginx 首先解析 core 模块的配置。core 模块提供一些块指令，这些指令引入其他类型的模块，nginx 遇到这些指令，就重新迭代解析过程，解析其他模块的配置。这些模块配置中又有一些块指令引入新的模块类型或者指令类型，nginx 就会再次迭代，解析这些新的配置类型。比如上图，nginx 遇到“events”指令，就重新调用 ngx_conf_parse() 解析 event 模块配置，解析完以后 ngx_conf_parse() 返回，

nginx 继续解析 core 模块指令，直到遇到“http”指令。nginx 再次调用 ngx_conf_parse() 解析 http 模块配置的 http 级指令，当遇到“server”指令时，nginx 又一次调用 ngx_conf_parse() 解析 http 模块配置的 server 级指令。

了解了 nginx 解析配置的流程，我们来看其中的关键函数 ngx_conf_parse()。

ngx_conf_parse() 解析配置分成两个主要阶段，一个是词法分析，一个是指令解析。

词法分析通过 ngx_conf_read_token() 函数完成。指令解析有两种方式，其一是使用 nginx 内建的指令解析机制，其二是使用第三方自定义指令解析机制。自定义指令解析可以参见下面的代码：

这里注意 cf->handler 和 cf->handler_conf 两个属性，其中 handler 是自定义解析函数指针，handler_conf 是 conf 指针。

下面着重介绍 nginx 内建的指令解析机制。本机制分为 4 个步骤：

1. 只有处理的模块的类型是 NGX_CONF_MODULE 或者是当前正在处理的模块类型，才可能被执行。nginx 中有一种模块类型是 NGX_CONF_MODULE，当前只有 ngx_conf_module 一种，只支持一条指令“include”。“include”指令的实现我们后面再进行介绍。
2. 匹配指令名，判断指令用法是否正确。

1. 指令的 Context 必须当前解析 Context 相符；
2. 非块指令必须以“;”结尾；
3. 块指令必须后接“{”；
4. 指令参数个数必须正确。注意指令参数有最大值 NGX_CONF_MAX_ARGS，目前值为 8。

3. 取得指令工作的 conf 指针。

1. NGX_DIRECT_CONF 常量单纯用来指定配置存储区的寻址方法，只用于 core 模块。
2. NGX_MAIN_CONF 常量有两重含义，其一是指定指令的使用上下文是 main（其实还是指 core 模块），其二是指定配置存储区的寻址方法。所以，在代码中常常可以见到使用上下文是 main 的指令的 cmd->type 属性定义如下：

表示指令使用上下文是 main，conf 寻址方式是直接寻址。

使用 NGX_MAIN_CONF 还表示指定配置存储区的寻址方法的指令有 4 个：“events”、“http”、“mail”、“imap”。这四个指令也有共同之处——都是使用上下文是 main 的块指令，并且块中的指令都使用其他类型的模块（分别是 event 模块、http 模块、mail 模块和 mail 模块）来处理。

后面分析 ngx_http_block() 函数时，再具体分析为什么需要 NGX_MAIN_CONF 这种配置寻址方式。

3. 除开 core 模块，其他类型的模块都会使用第三种配置寻址方式，也就是根据 cmd->conf 的值从 cf->ctx 中取出对应的配置。举 http 模块为例，cf->conf 的可选值是 NGX_HTTP_MAIN_CONF_OFFSET、NGX_HTTP_SRV_CONF_OFFSET、NGX_HTTP_LOC_CONF_OFFSET，分别对应“http{}”、“server{}”、“location{}”这三个 http 配置级别。
4. 执行指令解析回调函数

cmd 是词法分析得到的结果，conf 是上一步得到的配置存贮区地址。

http 的解析

http 是作为一个 core 模块被 nginx 通用解析过程解析的，其核心就是“http”块指令回调，它完成了 http 解析的整个功能，从初始化到计算配置结果。

因为这是本书第一次提到块指令，所以在这里对其做基本介绍。

块指令的流程是：

1. 创建并初始化上下文环境；
2. 调用通用解析流程解析；
3. 根据解析结果进行后续合并处理；
4. 善后工作。

下面我们以“http”指令为例来介绍这个流程：

创建并初始化上下文环境　http 模块的上下文环境 ctx（注意我们在通用解析流程中提到的 ctx 是同一个东西）非常复杂，它是由三个指针数组组成的：main_conf、srv_conf、loc_conf。根据上面的代码可以看到，这三个数组的元素个数等于系统中 http 模块的个数。想想我们平时三四十个 http 模块的规模，大家也应该可以理解这一块结构的庞大。nginx 还为每个模块分别执行对应的 create 函数分配空间。我们需要注意后面的这一句“cf->ctx = ctx;”，正是这一句将解析配置的上下文切换成刚刚建立的 ctx。最后一段代码通过调用各个 http 模块的 preconfiguration 回调函数完成了对应模块的预处理操作，其主要工作是创建模块用到的变量。

调用通用解析流程解析　基本上所有的块指令都类似上面的三行语句（例外是 map，它用的是 cf->handler），改变通用解析流程的工作状态，然后调用通用解析流程。

根据解析结果进行后续合并处理　以上是 http 配置处理最重要的步骤。首先，在这里调用了各个模块的 postconfiguration 回调函数完成了模块配置过程。更重要的是，它为 nginx 建立了一棵完整的配置树（叶子节点为 location，包含 location 的完整配置）、完整的 location 搜索树、一张变量表、一张完成的阶段处理回调表 (phase handler)、一张 server 对照表和一张端口对照表。下面我们将分别介绍这些配置表的生成过程。

location 配置树　介绍这部分以前，先说明一个 nginx 的公理

公理 11-1：所有存放参数为 NGX_HTTP_SRV_CONF_OFFSET 的配置，配置仅在请求匹配的虚拟主机 (server) 上下文中生效，而所有存放参数为 NGX_HTTP_LOC_CONF_OFFSET 的配置，配置仅在请求匹配的路径 (location) 上下文中生效。

正因为有公理 11-1，所以 nginx 需要调用 merge_XXX 回调函数合并配置。具体的原因是很多配置指令可以放在不同配置层级，比如 access_log 既可以在 http 块中配置，又可以在 server 块中配置，还可以在 location 块中配置。但是因为公理 11-1，access_log 指令配置只有在路径 (location) 上下文中生效，所以需要将在

http 块中配置的 access_log 指令的配置向路径上下文做两次传递，第一次从 HTTP(http) 上下文到虚拟主机(server) 上下文，第二次从虚拟主机上下文到路径上下文。

可能有人会疑惑，为什么需要传递和合并呢？难道它们不在一张表里么？对，在创建并初始化上下文环境的过程中，大家已经看到，nginx 为 HTTP 上下文创建了 main_conf，为虚拟主机上下文创建了 srv_conf，为路径上下文创建了 loc_conf。但是，这张表只是用于解析在 http 块但不包含 server 块中定义的指令。而后面我们会看到，在 server 块指令中，同样建立了 srv_conf 和 loc_conf，用于解析在 server 块但不含 location 块中定义的指令。所以 nginx 其实维护了很多张配置表，因此 nginx 必须将配置在这些表中从顶至下不断传递。

前面列出的

就是初始化 HTTP 上下文，并且完成两步配置合并操作：从 HTTP 上下文合并到虚拟主机上下文，以及从虚拟主机上下文合并到路径上下文。其中，合并到路径上下问的操作是在 ngx_http_merge_servers 函数中进行的，见

大家注意观察 ngx_http_merge_servers 函数中的这段，先将 HTTP 上下文中的 location 配置合并到虚拟主机上下文，再将虚拟主机上下文中的 location 配置合并到路径上下文。

location 搜索树　公理 11-2：nginx 搜索路径时，正则匹配路径和其他的路径分开搜。

公理 11-3：nginx 路径可以嵌套。

所以，nginx 存放 location 的有两个指针，分别是

通过这段代码，大家还可以发现一点——nginx 的正则表达式需要 PCRE 支持。

正则表达式的路径是个指针数组，指针类型就是 ngx_http_core_loc_conf_t，所以数据结构决定算法，正则表达式路径的添加非常简单，就是在表中插入一项，这里不做介绍。

而其他路径，保存在 ngx_http_location_tree_node_t 指针指向的搜索树 static_locations，则是变态复杂，可以看得各位大汗淋漓。

为了说明这棵树的构建，我们先了解其他路径包含哪些：

1. 普通前端匹配的路径，例如 location / {}
2. 抢占式前缀匹配的路径，例如 location ^~ / {}
3. 精确匹配的路径，例如 location = / {}
4. 命名路径，比如 location @a {}
5. 无名路径，比如 if {} 或者 limit_except {} 生成的路径

我们再来看 ngx_http_core_loc_conf_t 中如何体现这些路径：

普通前端匹配的路径	无
抢占式前缀匹配的路径	noregex = 1
精确匹配的路径	exact_match = 1
命名路径	named = 1
无名路径	noname = 1
正则路径	regex != NULL

有了这些基础知识，可以看代码了。首先是 ngx_http_init_locations 函数

大家可以看到，这个函数正是根据不同的路径类型将 locations 分成多段，并以不同的指针引用。首先注意开始的排序，根据 ngx_http_cmp_locations 比较各个 location，排序以后的顺序依次是

1. 精确匹配的路径和两类前缀匹配的路径 (字母序，如果某个精确匹配的路径的名字和前缀匹配的路径相同，精确匹配的路径排在前面)
2. 正则路径 (出现序)
3. 命名路径 (字母序)
4. 无名路径 (出现序)

这样 nginx 可以简单的截断列表得到不同类型的路径，nginx 也正是这样处理的。

另外还要注意一点，就是 ngx_http_init_locations 的迭代调用，这里的 clcf 引用了两个我们没有介绍过的字段 exact 和 inclusive。这两个字段最初是在 ngx_http_add_location 函数（添加 location 配置时必然调用）中设置的：

当然这部分的具体逻辑我们在介绍 location 解析是再具体说明。

接着我们看 ngx_http_init_static_location_trees 函数。通过刚才的 ngx_http_init_locations 函数，留在 locations 数组里面的还有哪些类型的路径呢？

还有普通前端匹配的路径、抢占式前缀匹配的路径和精确匹配的路径这三类。

请注意除开这段核心代码，这个函数也有一个自迭代过程。

ngx_http_join_exact_locations 函数是将名字相同的精确匹配的路径和两类前缀匹配的路径合并，合并方法

简言之，就是将前缀匹配的路径放入精确匹配的路径的 inclusive 指针中，然后从列表删除前缀匹配的路径。

ngx_http_create_locations_list 函数将和某个路径名拥有相同名称前缀的路径添加到此路径节点的 list 指针域下，并将这些路径从 locations 中摘除。其核心代码是

ngx_http_create_locations_tree 函数则将刚才划分的各个 list 继续细分，形成一个二分搜索树，每个中间节点代表一个 location，每个 location 有如下字段：

1. exact：两类前缀匹配路径的 inclusive 指针域指向这两类路径的配置上下文；
2. inclusive：精确匹配路径的 exact 指针域指向这些路径的配置上下文；
3. auto_redirect：为各种 upstream 模块，比如 proxy、fastcgi 等等开启自动 URI 填充的功能；
4. len：路径前缀的长度。任何相同前缀的路径的 len 等于该路径名长度减去公共前缀的长度。比如路径/a 和/ab，前者的 len 为 2，后者的 len 也为 1；
5. name：路径前缀，任何相同前缀的路径的 name 是其已于公共前缀的部分。仍举路径/a 和/ab 为例，前者的 name 为/a，后者的 name 为 b；
6. left：左子树，当然是长度短或者字母序小的不同前缀的路径；
7. right：右子树，当然是长度长或者字母序大的不同前缀的路径。

通过上面三个步骤，nginx 就将 locations 列表中各种类型的路径分类处理并由不同的指针引用。对于前缀路径和精确匹配的路径，形成一棵独特的二分前缀树。

变量表　变量表的处理相对简单，即对照变量名表，为变量表中的每一个元素设置对应的 get_handler 和 data 字段。在前面的章节大家已经知道，变量表 variables 用以处理索引变量，而变量名表 variables_keys 用于处理可按变量名查找的变量。对于通过 ngx_http_get_variable_index 函数创建的索引变量，在变量表 variables 中的 get_handler 初始为空，如果没有认为设置的话，将会在这里进行初始化。

特殊变量的 get_handler 初始化也在这里进行：

变量前缀	get_handler	标志
http	ngx_http_variable_unknown_header_in	
sent_http	ngx_http_variable_unknown_header_out	
upstream_http	ngx_http_upstream_header_variable	NGX_HTTP_VAR_NOCACHEABLE
cookie	ngx_http_variable_cookie	
arg	ngx_http_variable_argument	NGX_HTTP_VAR_NOCACHEABLE

阶段处理回调表　按照下表顺序将各个模块设置的 phase handler 依次加入 cmcf->phase_engine.handlers 列表，各个 phase 的 phase handler 的 checker 不同。checker 主要用于限定某个 phase 的框架逻辑，包括处理返回值。

处理阶段 PHASE	checker	可自定义 handler
NGX_HTTP_POST_READ_PHASE	是	ngx_http_core_generic_phase
NGX_HTTP_SERVER_REWRITE_PHASE	是	ngx_http_core_rewrite_phase
NGX_HTTP_FIND_CONFIG_PHASE	否	ngx_http_core_find_config_phase
NGX_HTTP_REWRITE_PHASE	是	ngx_http_core_rewrite_phase
NGX_HTTP_POST_REWRITE_PHASE	否	ngx_http_core_post_rewrite_phase
NGX_HTTP_PREACCESS_PHASE	是	ngx_http_core_generic_phase
NGX_HTTP_ACCESS_PHASE	是	ngx_http_core_access_phase
NGX_HTTP_POST_ACCESS_PHASE	否	ngx_http_core_post_access_phase
NGX_HTTP_TRY_FILES_PHASE	否	ngx_http_core_try_files_phase
NGX_HTTP_CONTENT_PHASE	是	ngx_http_core_content_phase

注意相同 PHASE 的 phase handler 是按模块顺序的反序加入回调表的。另外在 NGX_HTTP_POST_REWRITE_PHASE 中，ph->next 指向 NGX_HTTP_FIND_CONFIG_PHASE 第一个 phase handler，以实现 rewrite last 逻辑。

server 对照表　大家如果读过 nginx 的“Server names”这篇官方文档，会了解 nginx 对于 server name 的处理分为 4 中情况：精确匹配、前缀通配符匹配、后缀通配符匹配和正则匹配。那么，下面是又一个公理，

公理 11-4：nginx 对于不同类型的 server name 分别处理。

所以，所谓 server 对照表，其实是四张表，分别对应四种类型的 server。数据结构决定算法，四张表决定

了 nginx 必须建立这四张表的行为。鉴于前三种类型和正则匹配可以分成两大类，nginx 使用两套策略生成 server 对照表——对正则匹配的虚拟主机名，nginx 为其建立一个数组，按照主机名在配置文件的出现顺序依次写入数组；而对于其他虚拟主机名，nginx 根据它们的类型为它们分别存放在三张 hash 表中。三张 hash 表的结构完全相同，但对于前缀通配或者后缀通配这两种类型的主机名，nginx 对通配符进行的预处理不同。其中“.taobao.com”这种特殊的前缀通配与普通的前缀通配处理又有不同。我们现在来介绍这些不同。

处理前缀通配是将字符串按节翻转，然后去掉通配符。举个例子，“*.example.com”会被转换成“com.example.\0”，而特殊的前缀通配“.example.com”会被转换成“com.example\0”。

处理后缀通配更简单，直接去掉通配符。也举个例子，“www.example.*”会被转换成“www.example\0”。

端口对照表　对于所有写在 server 配置中的 listen 指令，nginx 都会以 ip 和端口两个维度

善后工作　善后工作基本的就是一件事，还原解析上下文。“http”指令是这个干的

server 的管理

这里我们略去了 http 解析的过程，因为 http 是作为一个 core 模块被 nginx 标准解析

location 的管理

模块初始化

热代码部署

reload 过程解析

upgrade 过程解析

3.2.5 nginx 的请求处理阶段 (30%)

接收请求流程 (99%)

http 请求格式简介 (99%)

首先介绍一下 rfc2616 中定义的 http 请求基本格式：

第一行是请求行（request line），用来说明请求方法，要访问的资源以及所使用的 HTTP 版本：

请求方法（Method）的定义如下，其中最常用的是 GET，POST 方法：

要访问的资源由统一资源地位符 URI(Uniform Resource Identifier) 确定，它的一个比较通用的组成格式（rfc2396）如下：

一般来说根据请求方法（Method）的不同，请求 URI 的格式会有所不同，通常只需写出 path 和 query 部分。

http 版本 (version) 定义如下，现在用的一般为 1.0 和 1.1 版本：

请求行的下一行则是请求头，rfc2616 中定义了 3 种不同类型的请求头，分别为 general-header，request-header 和 entity-header，每种类型 rfc 中都定义了一些通用的头，其中 entity-header 类型可以包含自定义的头。

请求头读取 (99%)

这一节介绍 nginx 中请求头的解析，nginx 的请求处理流程中，会涉及到 2 个非常重要的数据结构，ngx_connection_t 和 ngx_http_request_t，分别用来表示连接和请求，这 2 个数据结构在本书的前篇中已经做了比较详细的介绍，没有印象的读者可以翻回去复习一下，整个请求处理流程从头到尾，对应着这 2 个数据结构的分配，初始化，使用，重用和销毁。

nginx 在初始化阶段，具体是在 init process 阶段的 ngx_event_process_init 函数中会为每一个监听套接字分配一个连接结构（ngx_connection_t），并将该连接结构的读事件成员（read）的事件处理函数设置为 ngx_event_accept，并且如果没有使用 accept 互斥锁的话，在这个函数中会将该读事件挂载到 nginx 的事件处理模型上（poll 或者 epoll 等），反之则会等到 init process 阶段结束，在工作进程的事件处理循环中，某个进程抢到了 accept 锁才能挂载该读事件。

当一个工作进程在某个时刻将监听事件挂载上事件处理模型之后，nginx 就可以正式的接收并处理客户端过来的请求了。这时如果有一个用户在浏览器的地址栏内输入一个域名，并且域名解析服务器将该域名解析到一台由 nginx 监听的服务器上，nginx 的事件处理模型接收到这个读事件之后，会速度交由之前注册好的事件处理函数 ngx_event_accept 来处理。

在 ngx_event_accept 函数中，nginx 调用 accept 函数，从已连接队列得到一个连接以及对应的套接字，接着分配一个连接结构（ngx_connection_t），并将新得到的套接字保存在该连接结构中，这里还会做一些基本的连接初始化工作：

1, 首先给该连接分配一个内存池，初始大小默认为 256 字节，可通过 connection_pool_size 指令设置；

2, 分配日志结构，并保存在其中，以便后续的日志系统使用；

3, 初始化连接相应的 io 收发函数，具体的 io 收发函数和使用的事件模型及操作系统相关；

4, 分配一个套接口地址（sockaddr），并将 accept 得到的对端地址拷贝在其中，保存在 sockaddr 字段；

5, 将本地套接口地址保存在 local_sockaddr 字段，因为这个值是从监听结构 ngx_listening_t 中可得，而监听结构中保存的只是配置文件中设置的监听地址，但是配置的监听地址可能是通配符 *，即监听在所有的地址上，所以连接中保存的这个值最终可能还会变动，会被确定为真正的接收地址；

6, 将连接的写事件设置为已就绪，即设置 ready 为 1，nginx 默认连接第一次为可写；

7, 如果监听套接字设置了 TCP_DEFER_ACCEPT 属性，则表示该连接上已经有数据包过来，于是设置读事件为就绪；

8, 将 sockaddr 字段保存的对端地址格式化为可读字符串，并保存在 addr_text 字段；

最后调用 ngx_http_init_connection 函数初始化该连接结构的其他部分。

ngx_http_init_connection 函数最重要的工作是初始化读写事件的处理函数：将该连接结构的写事件的处理函数设置为 ngx_http_empty_handler，这个事件处理函数不会做任何操作，实际上 nginx 默认连接第一次可写，不会挂载写事件，如果有数据需要发送，nginx 会直接写到这个连接，只有在发生一次写不完的情况下，才会挂载写事件到事件模型上，并设置真正的写事件处理函数，这里后面的章节还会做详细介绍；读事件的处理函数设置为 ngx_http_init_request，此时如果该连接上已经有数据过来（设置了 deferred accept)，则会直接调用 ngx_http_init_request 函数来处理该请求，反之则设置一个定时器并在事件处理模型上挂载一个读事件，等待数据到来或者超时。当然这里不管是已经有数据到来，或者需要等待数据到来，又或者等待超时，最终都会进入读事件的处理函数 -ngx_http_init_request。

ngx_http_init_request 函数主要工作即是初始化请求，由于它是一个事件处理函数，它只有唯一一个 ngx_event_t * 类型的参数，ngx_event_t 结构在 nginx 中表示一个事件，事件处理的上下文类似于一个中断处理的上下文，为了在这个上下文得到相关的信息，nginx 中一般会将连接结构的引用保存在事件结构的 data 字段，请求结构的引用则保存在连接结构的 data 字段，这样在事件处理函数中可以方便的得到对应的连接结构和请求结构。进入函数内部看一下，首先判断该事件是否是超时事件，如果是的话直接关闭连接并返回；反之则是指之前 accept 的连接上有请求过来需要处理。

ngx_http_init_request 函数首先在连接的内存池中为该请求分配一个 ngx_http_request_t 结构，这个结构将用来保存该请求所有的信息。分配完之后，这个结构的引用会被包存在连接的 hc 成员的 request 字段，以便于在长连接或 pipelined 请求中复用该请求结构。在这个函数中，nginx 根据该请求的接收端口和地址找到一个默认虚拟服务器配置（listen 指令的 default_server 属性用来标识一个默认虚拟服务器，否则监听在相同端口和地址的多个虚拟服务器，其中第一个定义的则为默认）。

nginx 配置文件中可以设置多个监听在不同端口和地址的虚拟服务器（每个 server 块对应一个虚拟服务器），另外还根据域名（server_name 指令可以配置该虚拟服务器对应的域名）来区分监听在相同端口和地址的虚拟服务器，每个虚拟服务器可以拥有不同的配置内容，而这些配置内容决定了 nginx 在接收到一个请求之后如何处理该请求。找到之后，相应的配置被保存在该请求对应的 ngx_http_request_t 结构中。注意这里根据端口和地址找到的默认配置只是临时使用一下，最终 nginx 会根据域名找到真正的虚拟服务器配置，随后的初始化工作还包括：

1, 将连接的读事件的处理函数设置为 ngx_http_process_request_line 函数，这个函数用来解析请求行，将请求的 read_event_handler 设置为 ngx_http_block_reading 函数，这个函数实际上什么都不做（当然在事件模型设置为水平触发时，唯一做的事情就是将事件从事件模型监听列表中删除，防止该事件一直被触发），后面会说到这里为什么会将 read_event_handler 设置为此函数；

2, 为这个请求分配一个缓冲区用来保存它的请求头，地址保存在 header_in 字段，默认大小为 1024 个字节，可以使用 client_header_buffer_size 指令修改，这里需要注意一下，nginx 用来保存请求头的缓冲区是在该请求所在连接的内存池中分配，而且会将地址保存一份在连接的 buffer 字段中，这样做的目的也是为了给该连接的下一次请求重用这个缓冲区，另外如果客户端发过来的请求头大于 1024 个字节，nginx 会重新分配更大的缓存区，默认用于大请求的头的缓冲区最大为 8K，最多 4 个，这 2 个值可以用 large_client_header_buffers 指令设置，后面还会说到请求行和一个请求头都不能超过一个最大缓冲区的大小；

3, 为这个请求分配一个内存池，后续所有与该请求相关的内存分配一般都会使用该内存池，默认大小为 4096 个字节，可以使用 request_pool_size 指令修改；

4, 为这个请求分配响应头链表，初始大小为 20；

5, 创建所有模块的上下文 ctx 指针数组，变量数据；

6, 将该请求的 main 字段设置为它本身，表示这是一个主请求，nginx 中对应的还有子请求概念，后面的章节会做详细的介绍；

7, 将该请求的 count 字段设置为 1，count 字段表示请求的引用计数；

8, 将当前时间保存在 start_sec 和 start_msec 字段，这个时间是该请求的起始时刻，将被用来计算一个请求的处理时间（request time），nginx 使用的这个起始点和 apache 略有差别，nginx 中请求的起始点是接收到客户端的第一个数据包开始，而 apache 则是接收到客户端的整个 request line 后开始算起；

9, 初始化请求的其他字段，比如将 uri_changes 设置为 11，表示最多可以将该请求的 uri 改写 10 次，subrequests 被设置为 201，表示一个请求最多可以发起 200 个子请求；

做完所有这些初始化工作之后，ngx_http_init_request 函数会调用读事件的处理函数来真正的解析客户端发过来的数据，也就是会进入 ngx_http_process_request_line 函数中处理。

解析请求行 (99%) ngx_http_process_request_line 函数的主要作用即是解析请求行，同样由于涉及到网络 IO 操作，即使是很短的一行请求行可能也不能被一次读完，所以在之前的 ngx_http_init_request 函数中，ngx_http_process_request_line 函数被设置为读事件的处理函数，它也只拥有一个唯一的 ngx_event_t * 类型参数，并且在函数的开头，同样需要判断是否是超时事件，如果是的话，则关闭这个请求和连接；否则开始正常的解析流程。先调用 ngx_http_read_request_header 函数读取数据。

由于可能多次进入 ngx_http_process_request_line 函数，ngx_http_read_request_header 函数首先检查请求的 header_in 指向的缓冲区内是否有数据，有的话直接返回；否则从连接读取数据并保存在请求的 header_in 指向的缓存区，而且只要缓冲区有空间的话，会一次尽可能多的读数据，读到多少返回多少；如果客户端暂时没有发任何数据过来，并返回 NGX_AGAIN，返回之前会做 2 件事情：

1，设置一个定时器，时长默认为 60s，可以通过指令 client_header_timeout 设置，如果定时事件到达之前没有任何可读事件，nginx 将会关闭此请求；

2，调用 ngx_handle_read_event 函数处理一下读事件 -如果该连接尚未在事件处理模型上挂载读事件，则将其挂载上；

如果客户端提前关闭了连接或者读取数据发生了其他错误，则给客户端返回一个 400 错误（当然这里并不保证客户端能够接收到响应数据，因为客户端可能都已经关闭了连接），最后函数返回 NGX_ERROR；

如果 ngx_http_read_request_header 函数正常的读取到了数据，ngx_http_process_request_line 函数将调用 ngx_http_parse_request_line 函数来解析，这个函数根据 http 协议规范中对请求行的定义实现了一个有限状态机，经过这个状态机，nginx 会记录请求行中的请求方法（Method），请求 uri 以及 http 协议版本在缓冲区中的起始位置，在解析过程中还会记录一些其他有用的信息，以便后面的处理过程中使用。如果解析请求行的过程中没有产生任何问题，该函数会返回 NGX_OK；如果请求行不满足协议规范，该函数会立即终止解析过程，并返回相应错误号；如果缓冲区数据不够，该函数返回 NGX_AGAIN。

在整个解析 http 请求的状态机中始终遵循着两条重要的原则：减少内存拷贝和回溯。

内存拷贝是一个相对比较昂贵的操作，大量的内存拷贝会带来较低的运行时效率。nginx 在需要做内存拷贝的地方尽量只拷贝内存的起始和结束地址而不是内存本身，这样做的话仅仅只需要两个赋值操作而已，大大降低了开

销，当然这样带来的影响是后续的操作不能修改内存本身，如果修改的话，会影响到所有引用到该内存区间的地方，所以必须很小心的管理，必要的时候需要拷贝一份。

这里不得不提到 nginx 中最能体现这一思想的数据结构，ngx_buf_t，它用来表示 nginx 中的缓存，在很多情况下，只需要将一块内存的起始地址和结束地址分别保存在它的 pos 和 last 成员中，再将它的 memory 标志置 1，即可表示一块不能修改的内存区间，在另外的需要一块能够修改的缓存的情形中，则必须分配一块所需大小的内存并保存其起始地址，再将 ngx_bug_t 的 temprary 标志置 1，表示这是一块能够被修改的内存区域。

再回到 ngx_http_process_request_line 函数中，如果 ngx_http_parse_request_line 函数返回了错误，则直接给客户端返回 400 错误；如果返回 NGX_AGAIN，则需要判断一下是否是由于缓冲区空间不够，还是已读数据不够。如果是缓冲区大小不够了，nginx 会调用 ngx_http_alloc_large_header_buffer 函数来分配另一块大缓冲区，如果大缓冲区还不够装下整个请求行，nginx 则会返回 414 错误给客户端，否则分配了更大的缓冲区并拷贝之前的数据之后，继续调用 ngx_http_read_request_header 函数读取数据来进入请求行自动机处理，直到请求行解析结束；

如果返回了 NGX_OK，则表示请求行被正确的解析出来了，这时先记录好请求行的起始地址以及长度，并将请求 uri 的 path 和参数部分保存在请求结构的 uri 字段，请求方法起始位置和长度保存在 method_name 字段，http 版本起始位置和长度记录在 http_protocol 字段。还要从 uri 中解析出参数以及请求资源的拓展名，分别保存在 args 和 exten 字段。接下来将要解析请求头，将在下一小节中接着介绍。

解析请求头 (99%)　在 ngx_http_process_request_line 函数中，解析完请求行之后，如果请求行的 uri 里面包含了域名部分，则将其保存在请求结构的 headers_in 成员的 server 字段，headers_in 用来保存所有请求头，它的类型为 ngx_http_headers_in_t：

接着，该函数会检查进来的请求是否使用的是 http0.9，如果是的话则使用从请求行里得到的域名，调用 ngx_http_find_virtual_server（）函数来查找用来处理该请求的虚拟服务器配置，之前通过端口和地址找到的默认配置不再使用，找到相应的配置之后，则直接调用 ngx_http_process_request（）函数处理该请求，因为 http0.9 是最原始的 http 协议，它里面没有定义任何请求头，显然就不需要读取请求头的操作。

当然，如果是 1.0 或者更新的 http 协议，接下来要做的就是读取请求头了，首先 nginx 会为请求头分配空间，ngx_http_headers_in_t 结构的 headers 字段为一个链表结构，它被用来保存所有请求头，初始为它分配了 20 个节点，每个节点的类型为 ngx_table_elt_t，保存请求头的 name/value 值对，还可以看到 ngx_http_headers_in_t 结构有很多类型为 ngx_table_elt_t* 的指针成员，而且从它们的命名可以看出是一些常见的请求头名字，nginx 对这些常用的请求头在 ngx_http_headers_in_t 结构里面保存了一份引用，后续需要使用的话，可以直接通过这些成员得到，另外也事先为 cookie 头分配了 2 个元素的数组空间，做完这些内存准备工作之后，该请求对应的读事件结构的处理函数被设置为 ngx_http_process_request_headers，并随后马上调用了该函数。

ngx_http_process_request_headers 函数循环的读取所有的请求头，并保存和初始化和请求头相关的结构，下面详细分析一下该函数：

因为 nginx 对读取请求头有超时限制，ngx_http_process_request_headers 函数作为读事件处理函数，一并处理了超时事件，如果读超时了，nginx 直接给该请求返回 408 错误：

读取和解析请求头的逻辑和处理请求行差不多，总的流程也是循环的调用 ngx_http_read_request_header（）函数读取数据，然后再调用一个解析函数来从读取的数据中解析请求头，直到解析完所有请求头，或者发生

解析错误为主。当然由于涉及到网络 io，这个流程可能发生在多个 io 事件的上下文中。

接着来细看该函数，先调用了 ngx_http_read_request_header（）函数读取数据，如果当前连接并没有数据过来，再直接返回，等待下一次读事件到来，如果读到了一些数据则调用 ngx_http_parse_header_line（）函数来解析，同样的该解析函数实现为一个有限状态机，逻辑很简单，只是根据 http 协议来解析请求头，每次调用该函数最多解析出一个请求头，该函数返回 4 种不同返回值，表示不同解析结果：

1，返回 NGX_OK，表示解析出了一行请求头，这时还要判断解析出的请求头名字里面是否有非法字符，名字里面合法的字符包括字母，数字和连字符（-），另外如果设置了 underscores_in_headers 指令为 on，则下划线也是合法字符，但是 nginx 默认下划线不合法，当请求头里面包含了非法的字符，nginx 默认只是忽略这一行请求头；如果一切都正常，nginx 会将该请求头及请求头名字的 hash 值保存在请求结构体的 headers_in 成员的 headers 链表, 而且对于一些常见的请求头，如 Host，Connection，nginx 采用了类似于配置指令的方式，事先给这些请求头分配了一个处理函数，当解析出一个请求头时，会检查该请求头是否有设置处理函数，有的话则调用之，nginx 所有有处理函数的请求头都记录在 ngx_http_headers_in 全局数组中：

ngx_http_headers_in 数组当前包含了 25 个常用的请求头，每个请求头都设置了一个处理函数，其中一部分请求头设置的是公共处理函数，这里有 2 个公共处理函数，ngx_http_process_header_line 和 ngx_http_process_unique_header_line。先来看一下处理函数的函数指针定义：

它有 3 个参数，r 为对应的请求结构，h 为指向该请求头在 headers_in.headers 链表中对应节点的指针，offset 为该请求头对应字段在 ngx_http_headers_in_t 结构中的偏移。

再来看 ngx_http_process_header_line 函数：

这个函数只是简单将该请求头在 ngx_http_headers_in_t 结构中保存一份引用。ngx_http_process_unique_header_line 功能类似，不同点在于该函数会检查这个请求头是否是重复的，如果是的话，则给该请求返回 400 错误。

ngx_http_headers_in 数组中剩下的请求头都有自己特殊的处理函数，这些特殊的函数根据对应的请求头有一些特殊的处理，下面拿 Host 头的处理函数 ngx_http_process_host 做一下介绍：

此函数的目的也是保存 Host 头的快速引用，它会对 Host 头的值做一些合法性检查，并从中解析出域名，保存在 headers_in.server 字段，实际上前面在解析请求行时，headers_in.server 可能已经被赋值为从请求行中解析出来的域名，根据 http 协议的规范，如果请求行中的 uri 带有域名的话，则域名以它为准，所以这里需检查一下 headers_in.server 是否为空，如果不为空则不需要再赋值。

其他请求头的特殊处理函数，不再做介绍，大致都是根据该请求头在 http 协议中规定的意义及其值设置请求的一些属性，必备后续使用。

对一个合法的请求头的处理大致为如上所述；

2，返回 NGX_AGAIN，表示当前接收到的数据不够，一行请求头还未结束，需要继续下一轮循环。在下一轮循环中，nginx 首先检查请求头缓冲区 header_in 是否已满，如够满了，则调用 ngx_http_alloc_large_header_buffer（）函数分配更多缓冲区，下面分析一下 ngx_http_alloc_large_header_buffer 函数：

当 ngx_http_alloc_large_header_buffer 函数返回 NGX_DECLINED 时，表示客户端发送了一行过大的请求头，或者是整个请求头部超过了限制，nginx 会返回 494 错误，注意到 nginx 再返回 494 错误之前将请求的 lingering_close 标识置为了 1，这样做的目的是在返回响应之丢弃掉客户端发过来的其他数据；

3，返回 NGX_HTTP_PARSE_INVALID_HEADER，表示请求头解析过程中遇到错误，一般为客户端发送了不符合协议规范的头部，此时 nginx 返回 400 错误；

4，返回 NGX_HTTP_PARSE_HEADER_DONE，表示所有请求头已经成功的解析，这时请求的状态被设置为 NGX_HTTP_PROCESS_REQUEST_STATE，意味着结束了请求读取阶段，正式进入了请求处理阶段，但是实际上请求可能含有请求体，nginx 在请求读取阶段并不会去读取请求体，这个工作交给了后续的请求处理阶段的模块，这样做的目的是 nginx 本身并不知道这些请求体是否有用，如果后续模块并不需要的话，一方面请求体一般较大，如果全部读取进内存，则白白耗费大量的内存空间，另一方面即使 nginx 将请求体写进磁盘，但是涉及到磁盘 io，会耗费比较多时间。所以交由后续模块来决定读取还是丢弃请求体是最明智的办法。

读取完请求头之后，nginx 调用了 ngx_http_process_request_header（）函数，这个函数主要做了两个方面的事情，一是调用 ngx_http_find_virtual_server（）函数查找虚拟服务器配置；二是对一些请求头做一些协议的检查。比如对那些使用 http1.1 协议但是却没有发送 Host 头的请求，nginx 给这些请求返回 400 错误。还有 nginx 现在的版本并不支持 chunked 格式的输入，如果某些请求申明自己使用了 chunked 格式的输入（请求带有值为 chunked 的 transfer_encoding 头部)，nginx 给这些请求返回 411 错误。等等。

最后调用 ngx_http_process_request（）函数处理请求, 至此，nginx 请求头接收流程就介绍完毕。

请求体读取 (100%)

上节说到 nginx 核心本身不会主动读取请求体，这个工作是交给请求处理阶段的模块来做，但是 nginx 核心提供了 ngx_http_read_client_request_body() 接口来读取请求体，另外还提供了一个丢弃请求体的接口 -ngx_http_discard_request_body()，在请求执行的各个阶段中，任何一个阶段的模块如果对请求体感兴趣或者希望丢掉客户端发过来的请求体，可以分别调用这两个接口来完成。这两个接口是 nginx 核心提供的处理请求体的标准接口，如果希望配置文件中一些请求体相关的指令（比如 client_body_in_file_only，client_body_buffer_size 等）能够预期工作，以及能够正常使用 nginx 内置的一些和请求体相关的变量（比如 $request_body 和 $request_body_file），一般来说所有模块都必须调用这些接口来完成相应操作，如果需要自定义接口来处理请求体，也应尽量兼容 nginx 默认的行为。

读取请求体　请求体的读取一般发生在 nginx 的 content handler 中，一些 nginx 内置的模块，比如 proxy 模块，fastcgi 模块，uwsgi 模块等，这些模块的行为必须将客户端过来的请求体（如果有的话）以相应协议完整的转发到后端服务进程，所有的这些模块都是调用了 ngx_http_read_client_request_body() 接口来完成请求体读取。值得注意的是这些模块会把客户端的请求体完整的读取后才开始往后端转发数据。

由于内存的限制，ngx_http_read_client_request_body() 接口读取的请求体会部分或者全部写入一个临时文件中，根据请求体的大小以及相关的指令配置，请求体可能完整放置在一块连续内存中，也可能分别放置在两块不同内存中，还可能全部存在一个临时文件中，最后还可能一部分在内存，剩余部分在临时文件中。下面先介绍一下和这些不同存储行为相关的指令：

client_body_buffer_size 设置缓存请求体的 buffer 大小，默认为系统页大小的 2 倍，当请求体的大小超过此大小时，nginx 会把请求体写入到临时文件中。可以根据业务需求设置合适的大小，尽量避免磁盘 io 操作;

client_body_in_single_buffer 指示是否将请求体完整的存储在一块连续的内存中，默认为 off，如果此指令被设置为 on，则 nginx 会保证请求体在不大于 client_body_buffer_size 设置的

值时，被存放在一块连续的内存中，但超过大小时会被整个写入一个临时文件；

client_body_in_file_only 设置是否总是将请求体保存在临时文件中，默认为 off，当此指定被设置为 on 时，即使客户端显示指示了请求体长度为 0 时，nginx 还是会为请求创建一个临时文件。

接着介绍 ngx_http_read_client_request_body() 接口的实现，它的定义如下：

该接口有 2 个参数，第 1 个为指向请求结构的指针，第 2 个为一个函数指针，当请求体读完时，它会被调用。之前也说到根据 nginx 现有行为，模块逻辑会在请求体读完后执行，这个回调函数一般就是模块的逻辑处理函数。ngx_http_read_client_request_body() 函数首先将参数 r 对应的主请求的引用加 1，这样做的目的和该接口被调用的上下文有关，一般而言，模块是在 content handler 中调用此接口，一个典型的调用如下：

上面的代码是在 porxy 模块的 content handler，ngx_http_proxy_handler() 中调用了 ngx_http_read_client_request_body() 函数，其中 ngx_http_upstream_init() 被作为回调函数传入进接口中，另外 nginx 中模块的 content handler 调用的上下文如下：

上面的代码中，content handler 调用之后，它的返回值作为参数调用了 ngx_http_finalize_request() 函数，在请求体没有被接收完全时，ngx_http_read_client_request_body() 函数返回值为 NGX_AGAIN，此时 content handler，比如 ngx_http_proxy_handler() 会返回 NGX_DONE，而 NGX_DONE 作为参数传给 ngx_http_finalize_request() 函数会导致主请求的引用计数减 1，所以正好抵消了 ngx_http_read_client_request_body() 函数开头对主请求计数的加 1。

接下来回到 ngx_http_read_client_request_body() 函数，它会检查该请求的请求体是否已经被读取或者被丢弃了，如果是的话，则直接调用回调函数并返回 NGX_OK，这里实际上是为子请求检查，子请求是 nginx 中的一个概念，nginx 中可以在当前请求中发起另外一个或多个全新的子请求来访问其他的 location，关于子请求的具体介绍会在后面的章节作详细分析，一般而言子请求不需要自己去读取请求体。

函数接着调用 ngx_http_test_expect() 检查客户端是否发送了 Expect: 100-continue 头，是的话则给客户端回复''HTTP/1.1 100 Continue''，根据 http 1.1 协议，客户端可以发送一个 Expect 头来向服务器表明期望发送请求体，服务器如果允许客户端发送请求体，则会回复''HTTP/1.1 100 Continue''，客户端收到时，才会开始发送请求体。

接着继续为接收请求体做准备工作，分配一个 ngx_http_request_body_t 结构，并保存在 r->request_body，这个结构用来保存请求体读取过程用到的缓存引用，临时文件引用，剩余请求体大小等信息，它的定义如下:

temp_file 指向储存请求体的临时文件的指针；

bufs 指向保存请求体的链表头；

buf 指向当前用于保存请求体的内存缓存；

rest 当前剩余的请求体大小；

post_handler 保存传给 ngx_http_read_client_request_body() 函数的回调函数。

做好准备工作之后，函数开始检查请求是否带有 content_length 头，如果没有该头或者客户端发送了一个值为 0 的 content_length 头，表明没有请求体，这时直接调用回调函数并返回 NGX_OK 即可。当然如果 client_body_in_file_only 指令被设置为 on，且 content_length 为 0 时，该函数在调用回调函数之前，会创建一个空的临时文件。

进入到函数下半部分，表明客户端请求确实表明了要发送请求体，该函数会先检查是否在读取请求头时预读了请求体，这里的检查是通过判断保存请求头的缓存 (r->header_in) 中是否还有未处理的数据。如果有预读数据，则分配一个 ngx_buf_t 结构，并将 r->header_in 中的预读数据保存在其中，并且如果 r->header_in 中还有剩余空间，并且能够容下剩余未读取的请求体，这些空间将被继续使用，而不用分配新的缓存，当然甚至如果请求体已经被整个预读了，则不需要继续处理了，此时调用回调函数后返回。

如果没有预读数据或者预读不完整，该函数会分配一块新的内存（除非 r->header_in 还有足够的剩余空间），另外如果 request_body_in_single_buf 指令被设置为 no，则预读的数据会被拷贝进新开辟的内存块中，真正读取请求体的操作是在 ngx_http_do_read_client_request_body() 函数，该函数循环的读取请求体并保存在缓存中，如果缓存被写满了，其中的数据会被清空并写回到临时文件中。当然这里有可能不能一次将数据读到，该函数会挂载读事件并设置读事件 handler 为 ngx_http_read_client_request_body_handler，另外 nginx 核心对两次请求体的读事件之间也做了超时设置，client_body_timeout 指令可以设置这个超时时间，默认为 60 秒，如果下次读事件超时了，nginx 会返回 408 给客户端。

最终读完请求体后，ngx_http_do_read_client_request_body() 会根据配置，将请求体调整到预期的位置 (内存或者文件)，所有情况下请求体都可以从 r->request_body 的 bufs 链表得到，该链表最多可能有 2 个节点，每个节点为一个 buffer，但是这个 buffer 的内容可能是保存在内存中，也可能是保存在磁盘文件中。另外 $request_body 变量只在当请求体已经被读取并且是全部保存在内存中，才能取得相应的数据。

丢弃请求体 一个模块想要主动的丢弃客户端发过的请求体，可以调用 nginx 核心提供的 ngx_http_discard_request_body() 接口，主动丢弃的原因可能有很多种，如模块的业务逻辑压根不需要请求体，客户端发送了过大的请求体，另外为了兼容 http1.1 协议的 pipeline 请求，模块有义务主动丢弃不需要的请求体。总之为了保持良好的客户端兼容性，nginx 必须主动丢弃无用的请求体。下面开始分析 ngx_http_discard_request_body() 函数：

由于函数不长，这里把它完整的列出来了，函数的开始同样先判断了不需要再做处理的情况：子请求不需要处理，已经调用过此函数的也不需要再处理。接着调用 ngx_http_test_expect() 处理 http1.1 expect 的情况，根据 http1.1 的 expect 机制，如果客户端发送了 expect 头，而服务端不希望接收请求体时，必须返回 417(Expectation Failed) 错误。nginx 并没有这样做，它只是简单的让客户端把请求体发送过来，然后丢弃掉。接下来，函数删掉了读事件上的定时器，因为这时本身就不需要请求体，所以也无所谓客户端发送的快还是慢了，当然后面还会讲到，当 nginx 已经处理完该请求但客户端还没有发送完无用的请求体时，nginx 会在读事件上再挂上定时器。

客户端如果打算发送请求体，就必须发送 content-length 头，所以函数会检查请求头中的 content-length 头，同时还会查看其他地方是不是已经读取了请求体。如果确实有待处理的请求体，函数接着检查请求头 buffer 中预读的数据，预读的数据会直接被丢掉，当然如果请求体已经被全部预读，函数就直接返回了。

接下来，如果还有剩余的请求体未处理，该函数调用 ngx_handle_read_event() 在事件处理机制中挂载好读事件，并把读事件的处理函数设置为 ngx_http_discarded_request_body_handler。做好这些准备之后，该函数最后调用 ngx_http_read_discarded_request_body() 接口读取客户端过来的请求体并丢弃。如果客户端并没有一次将请求体发过来，函数会返回，剩余的数据等到下一次读事件过来时，交给 ngx_http_discarded_request_body_handler() 来处理，这时，请求的 discard_body 将被设置为 1 用来标识这种情况。另外请求的引用数 (count) 也被加 1，这样做的目的是客户端可能在 nginx 处理完请求之后仍未完整发送待发送的请求体，增加引用是防止 nginx 核心在处理完请求后直接释放了请求的相关资源。

ngx_http_read_discarded_request_body() 函数非常简单，它循环的从链接中读取数据并丢弃，直到读完接收缓冲区的所有数据，如果请求体已经被读完了，该函数会设置读事件的处理函数为 ngx_http_block_reading，这个函数仅仅删除水平触发的读事件，防止同一事件不断被触发。

最后看一下读事件的处理函数 ngx_http_discarded_request_body_handler，这个函数每次读事件来时会被调用，先看一下它的源码：

函数一开始就处理了读事件超时的情况，之前说到在 ngx_http_discard_request_body() 函数中已经删除了读事件的定时器，那么什么时候会设置定时器呢？答案就是在 nginx 已经处理完该请求，但是又没有完全将该请求的请求体丢弃的时候（客户端可能还没有发送过来），在 ngx_http_finalize_connection() 函数中，如果检查到还有未丢弃的请求体时，nginx 会添加一个读事件定时器，它的时长为 lingering_timeout 指令所指定，默认为 5 秒，不过这个时间仅仅两次读事件之间的超时时间，等待请求体的总时长为 lingering_time 指令所指定，默认为 30 秒。这种情况中，该函数如果检测到超时事件则直接返回并断开连接。同样，还需要控制整个丢弃请求体的时长不能超过 lingering_time 设置的时间，如果超过了最大时长，也会直接返回并断开连接。

如果读事件发生在请求处理完之前，则不用处理超时事件，也不用设置定时器，函数只是简单的调用 ngx_http_read_discarded_request_body() 来读取并丢弃数据。

子请求并不是 http 标准里面的概念，它是在当前请求中发起的一个新的请求，它拥有自己的 ngx_http_request_t 结构，uri 和 args。一般来说使用 subrequest 的效率可能会有些影响，因为它需要重新从 server rewrite 开始走一遍 request 处理的 PHASE，但是它在某些情况下使用能带来方便，比较常用的是用 subrequest 来访问一个 upstream 的后端，并给它一个 ngx_http_post_subrequest_t 的回调 handler，这样有点类似于一个异步的函数调用。对于从 upstream 返回的数据，subrequest 允许根据创建时指定的 flag，来决定由用户自己处理 (回调 handler 中) 还是由 upstream 模块直接发送到 out put filter。简单的说一下 subrequest 的行为，nginx 使用 subrequest 访问某个 location，产生相应的数据，并插入到 nginx 输出链的相应位置（创建 subrequest 时的位置），下面用 nginx 代码内的 addition 模块 (默认未编译进 nginx 核心，请使用 --with-http_addition_module 选项包含此模块) 来举例说明一下：

访问/main.htm，将得到如下响应：

上面的 add_before_body 指令发起一个 subrequest 来访问/hello.htm，并将产生的内容 (hello) 插入主请求响应体的开头，add_after_body 指令发起一个 subrequest 访问/world.htm，并将产生的内容 (world) 附加在主请求响应体的结尾。addition 模块是一个 filter 模块，但是 subrequest 既可以在 phase 模块中使用，也可以在 filter 模块中使用。

在进行源码解析之前，先来想想如果是我们自己要实现 subrequest 的上述行为，该如何来做？subrequest 还可能有自己的 subrequest，而且每个 subrequest 都不一定按照其创建的顺序来输出数据，所以简单的采用链表不好实现，于是进一步联想到可以采用树的结构来做，主请求即为根节点，每个节点可以有自己的子节点，遍历某节点表示处理某请求，自然的可以想到这里可能是用后根 (序) 遍历的方法，没错，实际上 Igor 采用树和链表结合的方式实现了 subrequest 的功能，但是由于节点（请求）产生数据的顺序不是固定按节点创建顺序 (左 -> 右)，而且可能分多次产生数据，不能简单的用后根 (序) 遍历。Igor 使用了 2 个链表的结构来实现，第一个是每个请求都有的 postponed 链表，一般情况下每个链表节点保存了该请求的一个子请求，该链表节点定义如下：

可以看到它有一个 request 字段，可以用来保存子请求，另外还有一个 ngx_chain_t 类型的 out 字段，实际上一个请求的 postponed 链表里面除了保存子请求的节点，还有保存该请求自己产生的数据的节点，数据保存在 out 字段；第二个是 posted_requests 链表，它挂载了当前需要遍历的请求（节点），该链表保存在主请求（根节点）的 posted_requests 字段，链表节点定义如下：

在 ngx_http_run_posted_requests 函数中会顺序的遍历主请求的 posted_requests 链表：

ngx_http_run_posted_requests 函数的调用点后面会做说明。

了解了一些实现的原理，来看代码就简单多了，现在正式进行 subrequest 的源码解析，首先来看一下创建 subrequest 的函数定义：

参数 r 为当前的请求，uri 和 args 为新的要发起的 uri 和 args，当然 args 可以为 NULL，psr 为指向一个 ngx_http_request_t 指针的指针，它的作用就是获得创建的子请求，ps 的类型为 ngx_http_post_subrequest_t，它的定义如下：

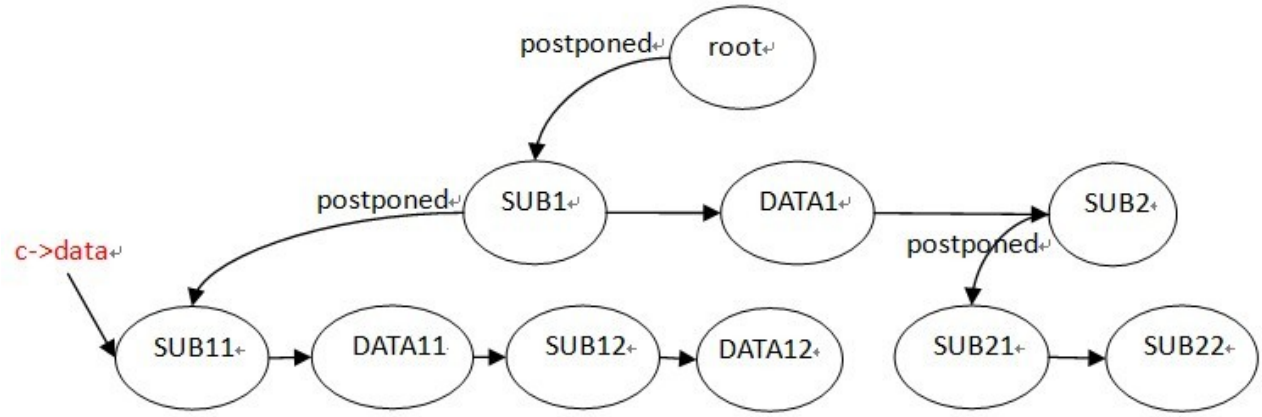
它就是之前说到的回调 handler，结构里面的 handler 类型为 ngx_http_post_subrequest_pt，它是函数指针，data 为传递给 handler 的额外参数。再来看一下 ngx_http_subrequest 函数的最后一个是 flags，现在的源码中实际上只有 2 种类型的 flag，分别为 NGX_HTTP_SUBREQUEST_IN_MEMORY 和 NGX_HTTP_SUBREQUEST_WAITED，第一个就是指定文章开头说到的子请求的 upstream 处理数据的方式，第二个参数表示如果该子请求提前完成 (按后续遍历的顺序)，是否设置将它的状态设为 done，当设置该参数时，提前完成就会设置 done，不设时，会让该子请求等待它之前的子请求处理完毕才会将状态设置为 done。

进入 ngx_http_subrequest 函数内部看看：

到这时，子请求创建完毕，一般来说子请求的创建都发生在某个请求的 content handler 或者某个 filter 内，从上面的函数可以看到子请求并没有马上被执行，只是被挂载在了主请求的 posted_requests 链表中，那它什么时候可以执行呢？之前说到 posted_requests 链表是在 ngx_http_run_posted_requests 函数中遍历，那么 ngx_http_run_posted_requests 函数又是在什么时候调用？它实际上是在某个请求的读（写）事件的 handler 中，执行完该请求相关的处理后被调用，比如主请求在走完一遍 PHASE 的时候会调用 ngx_http_run_posted_requests，这时子请求得以运行。

这时实际还有 1 个问题需要解决，由于 nginx 是多进程，是不能够随意阻塞的（如果一个请求阻塞了当前进程，就相当于阻塞了这个进程 accept 到的所有其他请求，同时该进程也不能 accept 新请求），一个请求可能由于某些原因需要阻塞（比如访问 io），nginx 的做法是设置该请求的一些状态并在 epoll 中添加相应的事件，然后转去处理其他请求，等到该事件到来时再继续处理该请求，这样的行为就意味着一个请求可能需要多次执行机会才能完成，对于一个请求的多个子请求来说，意味着它们完成的先后顺序可能和它们创建的顺序是不一样的，所以必须有一种机制让提前完成的子请求保存它产生的数据，而不是直接输出到 out chain，同时也能够让当前能够往 out chain 输出数据的请求及时的输出产生的数据。作者 Igor 采用 ngx_connection_t 中的 data 字段，以及一个 body filter，即 ngx_http_postpone_filter，还有 ngx_http_finalize_request 函数中的一些逻辑来解决这个问题。

下面用一个图来做说明，下图是某时刻某个主请求和它的所有子孙请求的树结构：



图中的 root 节点即为主请求，它的 postponed 链表从左至右挂载了 3 个节点，SUB1 是它的第一个子请求，DATA1 是它产生的一段数据，SUB2 是它的第 2 个子请求，而且这 2 个子请求分别有它们自己的子请求及数据。ngx_connection_t 中的 data 字段保存的是当前可以往 out chain 发送数据的请求，文章开头说到发到客户端的数据必须按照子请求创建的顺序发送，这里即是按后续遍历的方法（SUB11->DATA11->SUB12->DATA12->(SUB1)->DATA1->SUB21->SUB22->(SUB2)->(ROOT)），上图中当前能够往客户端（out chain）发送数据的请求显然就是 SUB11，如果 SUB12 提前执行完成，并产生数据 DATA121，只要前面它还有节点未发送完毕，DATA121 只能先挂载在 SUB12 的 postponed 链表下。这里还要注意一下的是 c->data 的设置，当 SUB11 执行完并且发送完数据之后，下一个将要发送的节点应该是 DATA11，但是该节点实际上保存的是数据，而不是子请求，所以 c->data 这时应该指向的是拥有改数据节点的 SUB1 请求。

下面看下源码具体是怎样实现的，首先是 ngx_http_postpone_filter 函数：

再来看 ngx_http_finalzie_request 函数：

3.2.6 nginx 的 upstream 原理解析

无缓冲的处理流程

带缓冲的处理流程

subrequest 访问 upstream 的处理流程

负载均衡算法

round-robin 算法

基于 ip 的 hash 算法

upstream 缓存机制

常用 upstream 模块分析

proxy 模块

fastcgi 模块

3.2.7 nginx 的 script 处理

正则表达式

变量

复杂变量

if 的处理

3.3 附录 A 编码风格 (100%)

3.3.1 Nginx 代码风格图示 (100%)

一、基本原则

K&R 编码风格（偏 BSD 子类）。

每行不能超过 80 列。

不用 TAB 对齐，用空格。

默认对齐单元是 4 个空格。

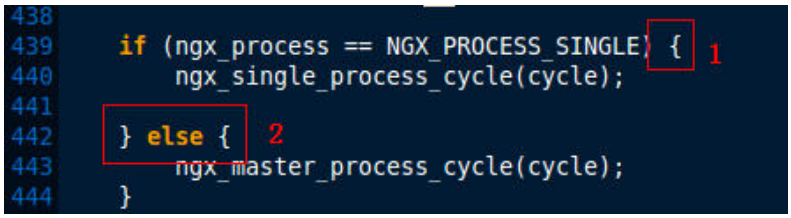
除宏定义外，字母均为小写，单词间用下划线 _ 间隔。

使用 C 方式的注释，不得使用//形式注释。

中缀运算符的前后须空一格，如 3 + 2 以及 a > 3。

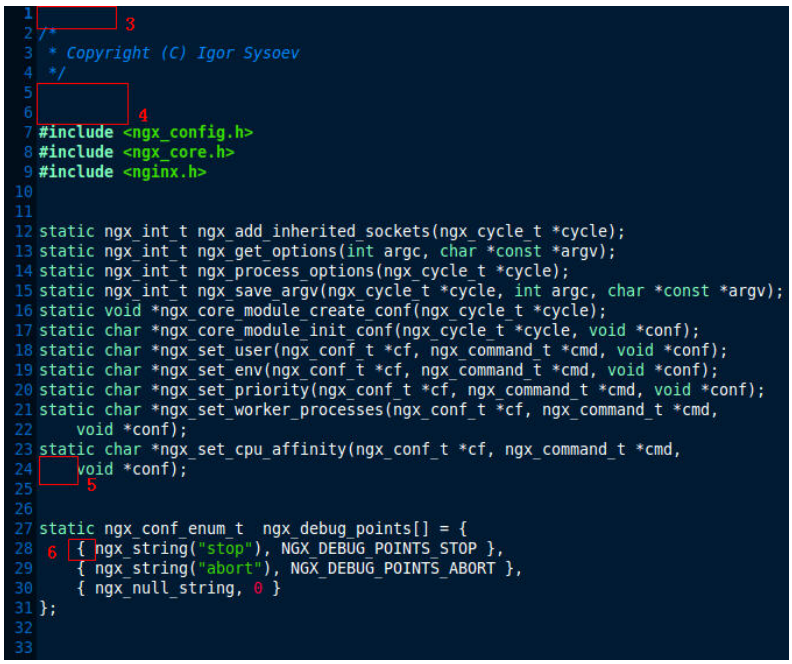
逗号后须空一格，如 foo(a, b, c);

二、风格图示



1、if/while/for/switch 语句的左花括号和关键字在同一行上，和括号之间空一个空格。

2、else 关键字和两个花括号在同一行上。



3、文件开始的注释空一行。

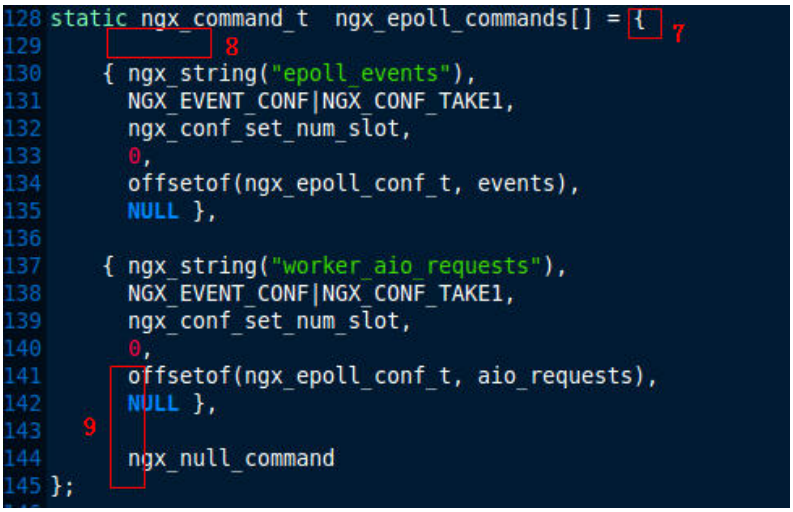
4、较为完整的代码块间的距离为空两行。如函数声明、函数定义之间等。

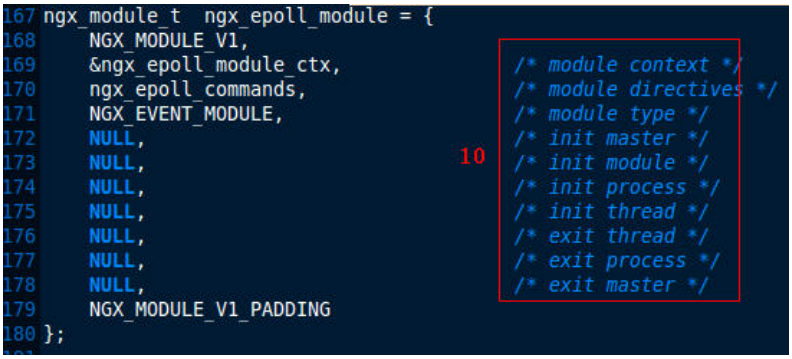
5、函数声明或定义若一行显示不小，则函数原型空 4 个空格。

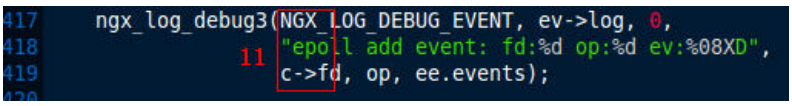
6、结构体数组的花括号和内容之间空一个空格。7、结构体数组的左花括号放在同一行上。

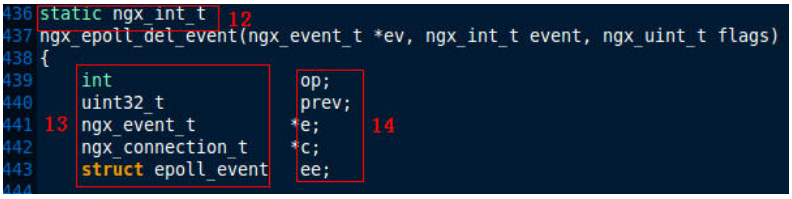
8、较大的结构体数组元素最开始空一行。

9、元素内容上下对齐。10、注释上下对齐。11、函数调用折行时，参数上下对齐。12、函数定义时，类型单独一行。


static ngx_command_t ngx_epoll_commands[] = {
{ ngx_string("epoll_events"),
NGX_EVENT_CONF|NGX_CONF_TAKE1,
ngx_conf_set_num_slot,
0,
offsetof(ngx_epoll_conf_t, events),
NULL },
{ ngx_string("worker_aio_requests"),
NGX_EVENT_CONF|NGX_CONF_TAKE1,
ngx_conf_set_num_slot,
0,
offsetof(ngx_epoll_conf_t, aio_requests),
NULL },
ngx_null_command
};
7
8
9


ngx_module_t ngx_epoll_module = {
NGX_MODULE_V1,
&ngx_epoll_module_ctx,
ngx_epoll_commands,
NGX_EVENT_MODULE,
NULL,
NULL,
NULL,
NULL,
NULL,
NULL,
NULL,
NGX_MODULE_V1_PADDING
};
/* module context */
/* module directives */
/* module type */
/* init master */
/* init module */
/* init process */
/* init thread */
/* exit thread */
/* exit process */
/* exit master */
10


ngx_log_debug3(NGX_LOG_DEBUG_EVENT, ev->log, 0,
"epoll add event: fd:%d op:%d ev:%08XD",
c->fd, op, ee.events);
11


static ngx_int_t
ngx_epoll_del_event(ngx_event_t *ev, ngx_int_t event, ngx_uint_t flags)
{
int op;
uint32_t prev;
ngx_event_t *e;
ngx_connection_t *c;
struct epoll_event ee;
12
13
14

13、变量声明的类型上下排列按照从短到长的顺序。注意，最下面的变量的类型和名称间的空格为 2-3 个。一般情况下为 2 个，这是 Nginx 中最小的变量声明中类型和名称的距离。

14、变量名称上下对齐——字母对齐，不包括指针的 * 号。15、结构体内变量上下对齐（字母，不包括指针的

```
typedef struct {
    ngx_msec_t                       bl_time;
    ngx_uint_t                       bl_state;

    ngx_uint_t                       status;
    time_t                           response_sec;
    ngx_uint_t                       response_msec;
    off_t                            response_length;

    ngx_str_t                       *peer;
} ngx_http_upstream_state_t;
```

的 * 号）。16、单行注释格式为/* something */ 17、多行注释的格式为：18、函数定义的左花括号独占一行。

```
    default:

    /* cached NGX_HTTP_BAD_GATEWAY, NGX_HTTP_GATEWAY_TIME_OUT, etc. */

        u->cache_status = NGX_HTTP_CACHE_HIT;

        return rc;
    }
```

```
    {
        /*
         * the r->request_body->buf can be reused for one request only,
         * the subrequests should allocate their own temporay bufs
         */

        u->output.free = ngx_alloc_chain_link(r->pool);
```

```
size_t
ngx_inet_ntop(int family, void *addr, u_char *text, size_t len)
{
    u_char  *p;

    switch (family) {

    case AF_INET:

        p = addr;

        return ngx_snprintf(text, len, "%ud.%ud.%ud.%ud",
                            p[0], p[1], p[2], p[3])
               - text;
```

19、switch 语句中，switch 和 case 关键字上下对齐。20、当条件表达式过长需要折行时，关系运算符须位

```
    if (lock_file.len) {
        lock_file.len--;

        if (ccf->lock_file.len != lock_file.len
            || ngx_strncmp(ccf->lock_file.data, lock_file.data, lock_file.len)
               != 0)
        {
            ngx_log_error(NGX_LOG_EMERG, cycle->log, 0,
                          "\"lock_file\" could not be changed, ignored");
        }
```

于下一行的行首，并与上一行的条件表达式的第一个字符对齐，同时右花括号须位于单独的一行，并与 if/while 等关键字对齐。21、else 语句之前须空出一行。22、在函数中，相同类型的变量声明放在一行上。

```
if (shift) {
    cidr->u.in.mask = htonl((ngx_uint_t) (0 - (1 << (32 - shift))));
} else {
    /* x86 compilers use a shl instruction that shifts by modulo 32 */
    cidr->u.in.mask = 0;
}
```

21

```
613 static ngx_int_t
614 ngx_parse_inet_url(ngx_pool_t *pool, ngx_url_t *u)
615 {
616     u_char              *p, *host, *port, *last, *uri, *args;
617     size_t               len;
618     ngx_int_t            n;
619     struct hostent      *h;
620     struct sockaddr_in  *sin;
```

22

3.4 附录 B 常用 API

3.4.1 B.1 字符串操作

3.4.2 B.2 读取请求体

3.4.3 B.3 操作请求头

3.4.4 B.4 取参数

3.4.5 B.5 URI 处理

3.4.6 B.6 buffer 操作

3.4.7 B.7 变量操作

3.4.8 B.8 日志输出

3.4.9 B.9 配置解析接口

3.4.10 B.10 文件操作

3.4.11 B.11 锁操作

3.5 附录 C 模块编译，调试与测试

3.5.1 C.1 添加第 3 方库

3.5.2 C.2 添加额外编译，链接选项

3.5.3 C.3 debug point

3.5.4 C.4 使用 gdb

3.5.5 C.5 调试日志

3.5.6 C.6 单元测试

3.5.7 C.7 功能测试

3.5.8 C.8 性能测试

第 4 章

团队成员

叔度 (http://blog.zhuzhaoyuan.com)

雕梁 (http://www.pagefault.info)

文景 (http://yaoweibin.cn)

李子 (http://blog.lifeibo.com)

卫越 (http://blog.sina.com.cn/u/1929617884)

袁茁 (http://yzprofile.me)

小熊 (http://dinic.iteye.com)

吉兆 (http://jizhao.blog.chinaunix.net)

静龙 (http://blog.csdn.net/fengmo_q)

竹权 (http://weibo.com/u/2199139545)

公远 (http://100continue.iteye.com/)

布可 (http://weibo.com/sifeierss)