

**在启动服务本身后除了HTTPSERVER外，似乎只有启动worker和启动handler会造成延时。**

Handler在上一篇已经找到了执行的源码。这一篇写一点关于worker的资料和思考。

有关worker的文档

<https://github.com/pytorch/serve/blob/master/docs/management_api.md>

managementAPI的可选配置

initial\_workers - the number of initial workers to create.

TorchServe will not run inference until there is at least one work assigned.（注册函数就会启动worker?）

而且默认注册配置是0个worker。Worker是否只能使用一个注册时特定的handler（即一个worker会分配给多个handler吗）？我觉得handler和worker是一对多的关系，而worker和handler是一对一的关系。因为worker是process，结合前一份对handler的分析(base\_handler.py中)，应该是worker启动的时候就调用了handler的initial加载好了模型。

当前的worker状态，minWorker是函数一直保持的最小worker数（必须unregister这个函数才能释放掉资源）。这样注册许多函数会浪费大量资源。但是文章并没有提maxWorker是如何根据请求增加的，以及它是否会自动释放。

