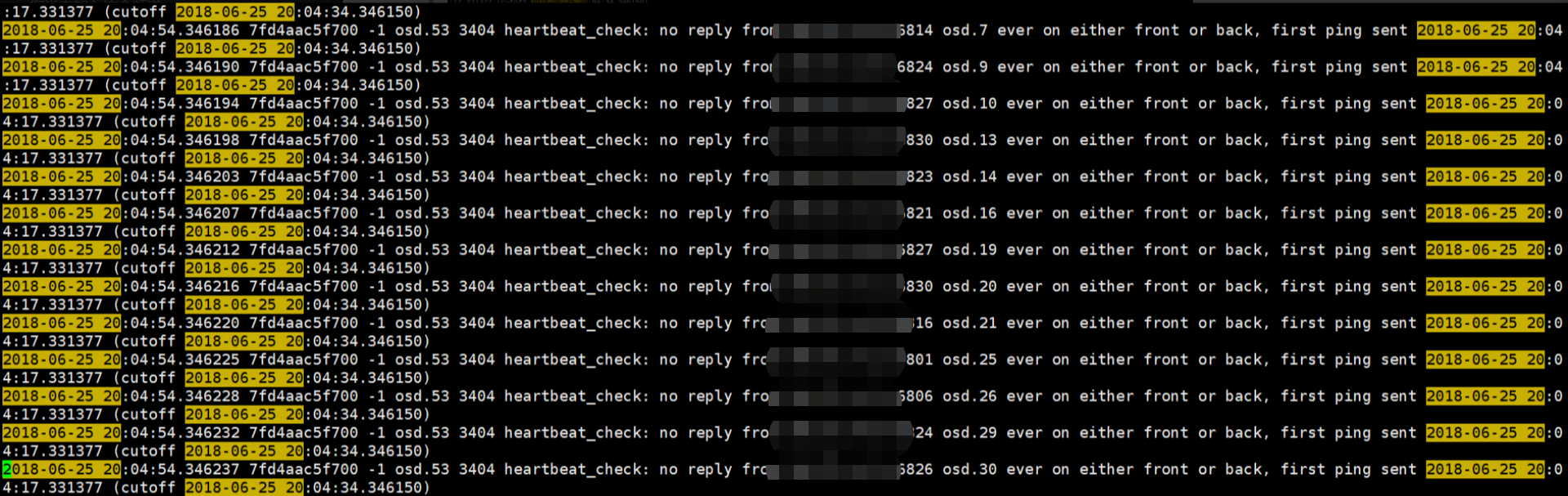
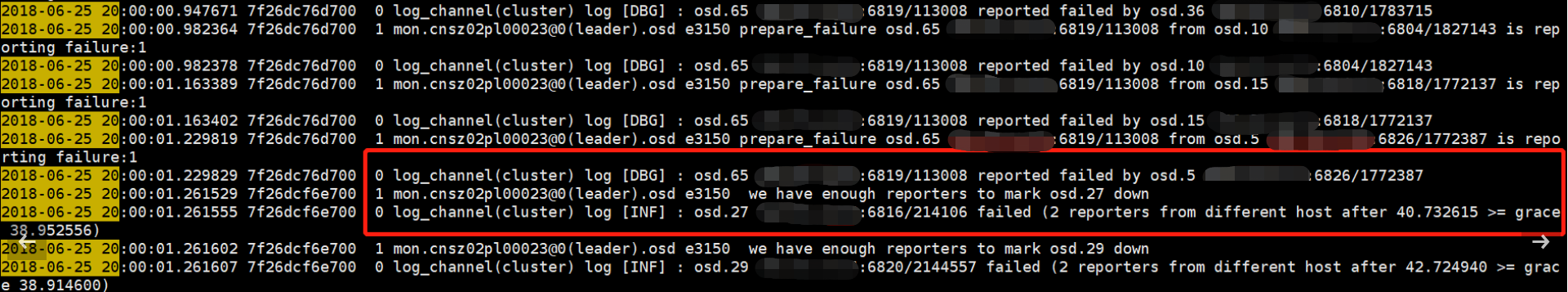
最近在维护ceph集群时发现有时osd时而UP时而Down, 发现是由于网络不稳定导致OSD之间不能及时的心跳汇报，导致误认为对方down掉了，故而想monitor报告OSD挂掉了。





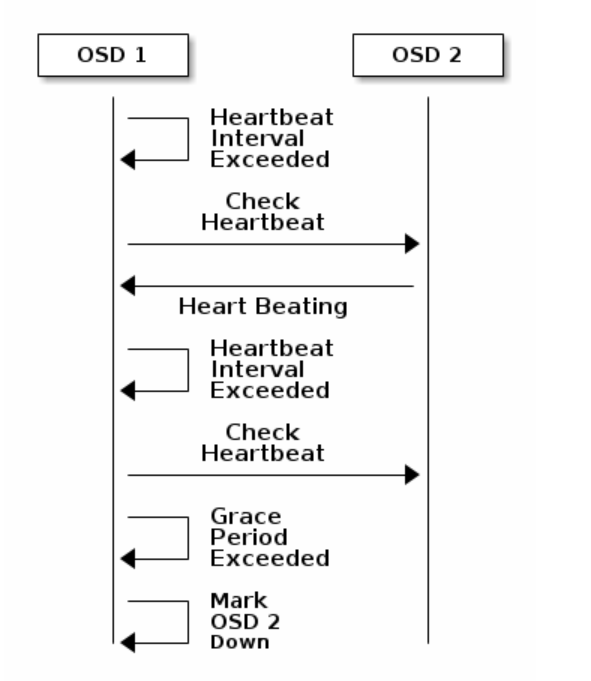
本章将讲解ceph之心跳检测实现原理。

ceph的心跳检测原理，官方文档写的很详细：

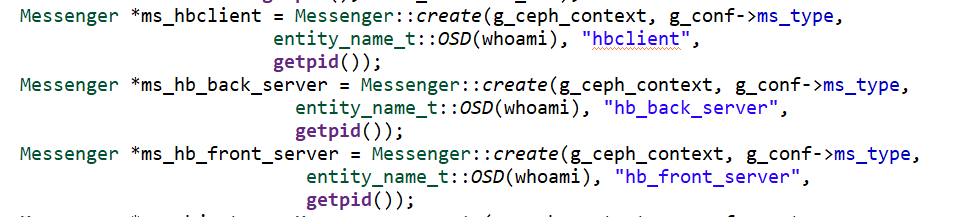
<http://docs.ceph.com/docs/master/rados/configuration/mon-osd-interaction/>

本文从代码级别去解释其实现过程：

osd 与 osd 之间的心跳：



在OSD启动时会建立3个创造了三个messenger用于心跳通信

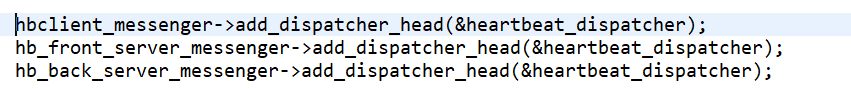


其中ms\_hbclient用来发送ping心跳

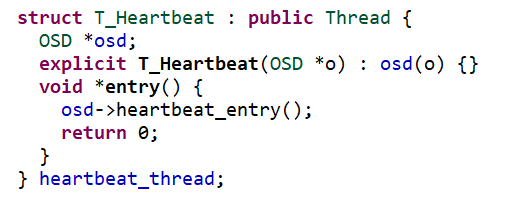
ms\_hb\_back\_server 用于接收来自back地址的ping心跳，即cluster network地址的心跳

ms\_hb\_front\_server 用于接收front地址的ping心跳。即public network地址的心跳。

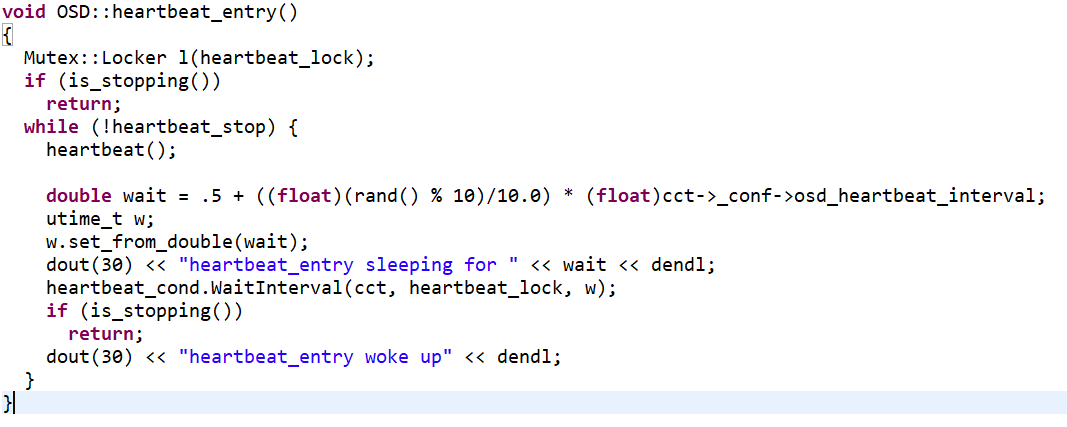
然后注册heartbeat\_dispatcher



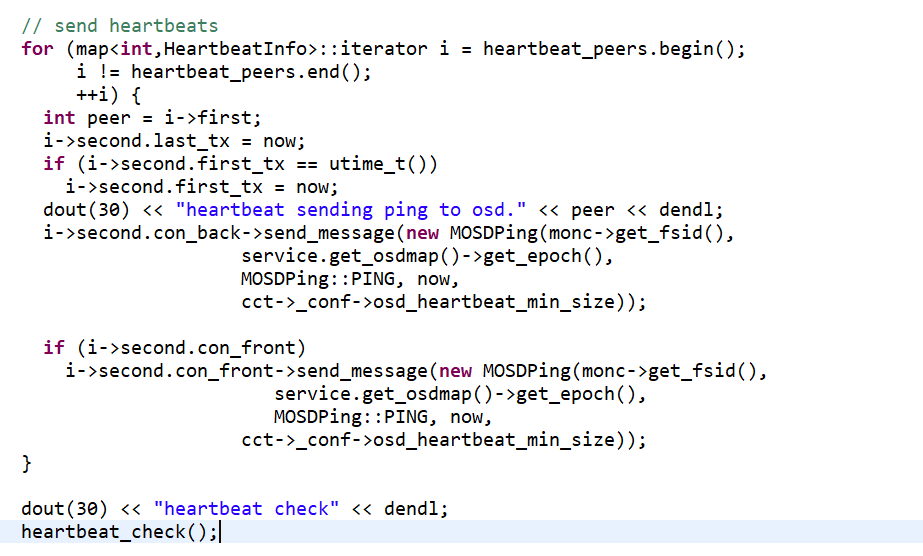
OSD会启动一个独立的线程用于定时向其他OSD发送ping 信息，



heartbeat\_entry()函数定时（0.5~6.5）秒调用heartbeat()发送ping信息

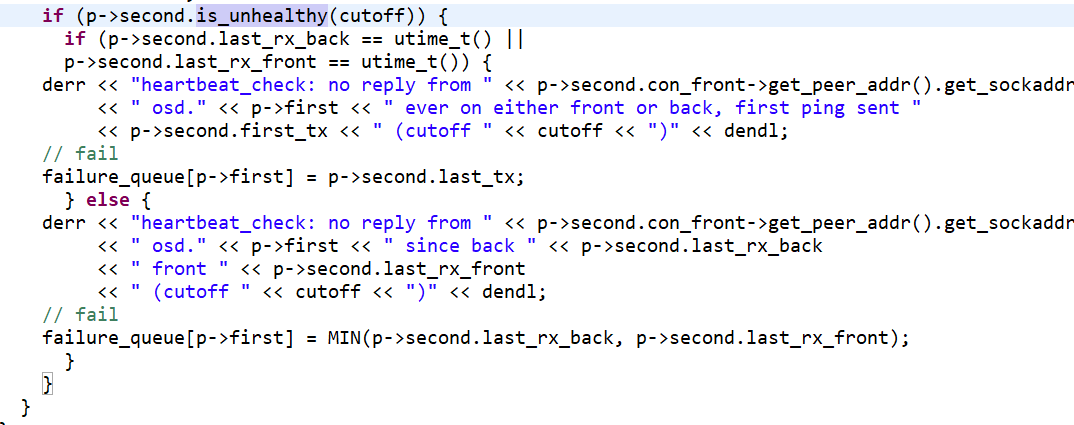


heartbeat()函数就是选择osd发送ping消息：



代码可以看到其分别向back地址和front地址发送ping消息。

同时发送消息后会检查一下是否超时：heartbeat\_check

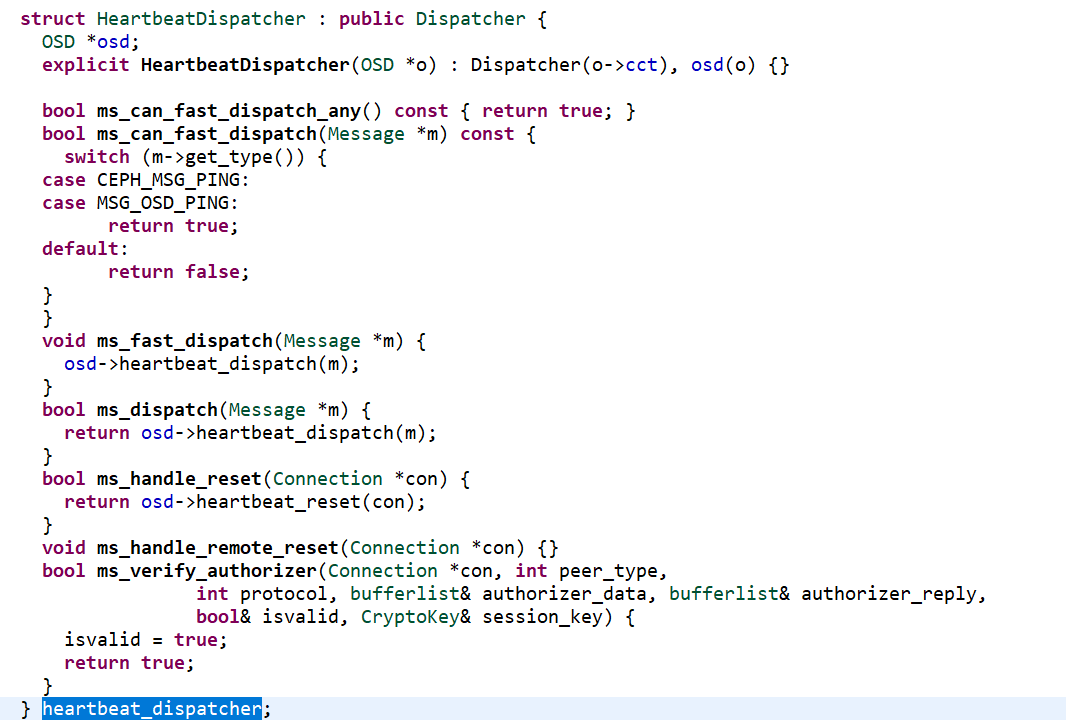


检测超时就加入failure\_queue，等待上报给monitor

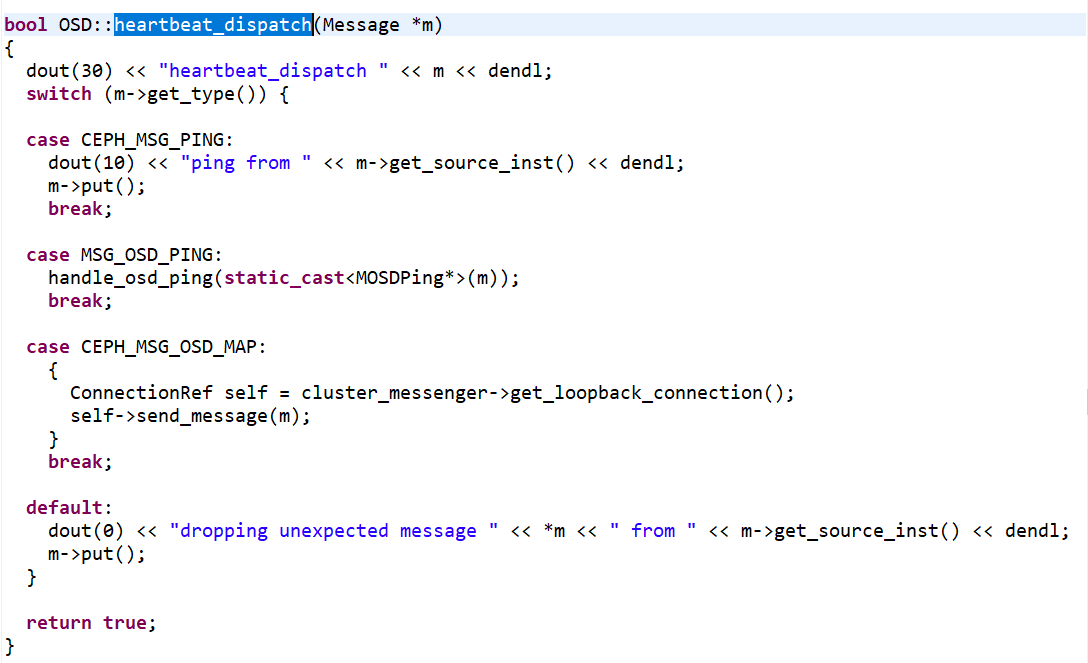
心跳是否超时的检查，另外还有专门的tick线程，也会检查心跳是否超时



上面讲到心跳消息的发送，那对于心跳消息的接收则有专门的注册的dispatcher类处理：



ms\_dispatch(Message \*m)处理:

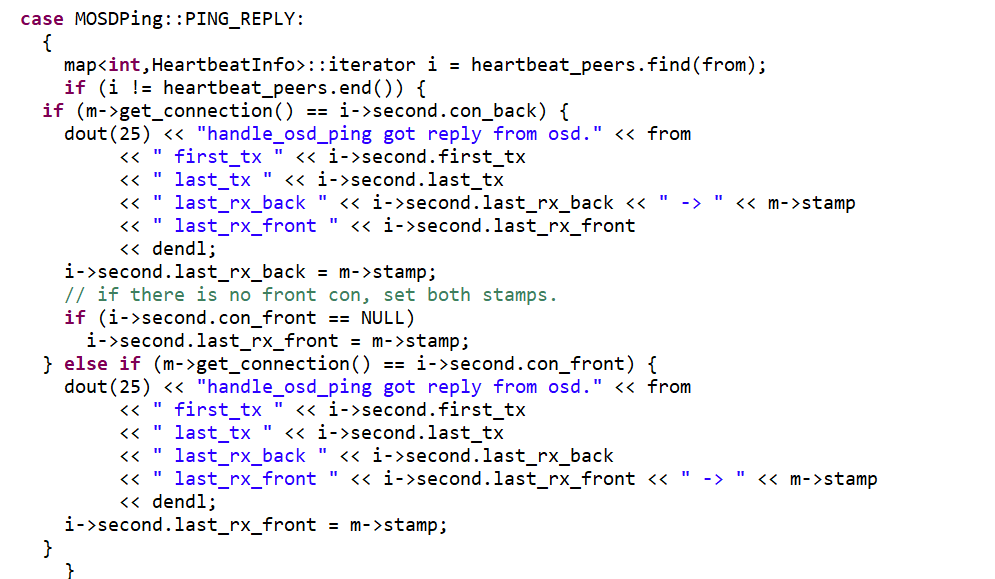


在分支MSG\_OSD\_ PING处理收到的ping消息：

处理心跳信息：

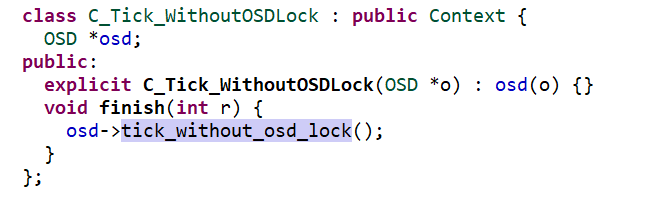


处理心跳回包：

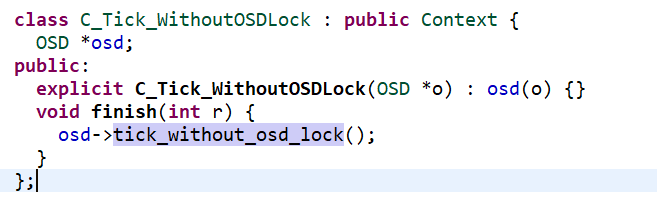


更新时间戳。判断是否超时交由专门的定时检查heartbeat\_check，上面已经讲述了。

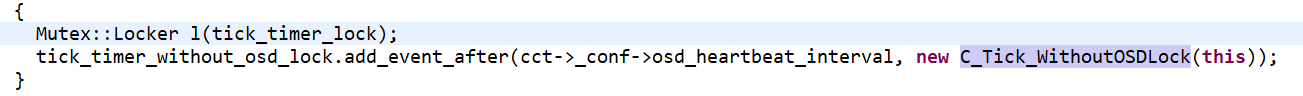
而对于向monitor



心跳超时上报的时候，也是在tick线程内完成



osd有专门的定时器负责检测心跳上报超时检查。



在tick\_without\_osd\_lock()中send\_failures就是向monitor上报OSD状态。

