kubernetes #73097

k8s会使用taint和toleration来为pod分配node。如果一个node具有某种taint，但是等待分配的pod没有具备该taint对应的toleration则该pod不能分配到这个node上。

k8s增加了TaintNodeByCondition来限制pod的分配，例如一个pod无法被分配在一个有NotReady taint的node上。根据kubelet对节点的更新，NotReady状态传回至Node controller，Node controller会将NotReady taint标记在node上，但是可能存在NotReady taint还没有被标记时，scheduler已经将pod分配给该node，从而导致逻辑上的错误。

根据#73097

The problem though is that an unready node may get the taint with some delay. In such scenarios some pods may be scheduled on NotReady nodes. Pods by default tolerate NotReady nodes for 300 seconds. So, some services that may be scheduled on such nodes can remain unresponsive for 5+ minutes.

其中300秒是默认NotExecute时间。

在原来的实现中，NotReady是在node update时（由kubelet）更新的，因为网络等原因会导致node controller获取NotReady状态有延迟，从而为该node标记NotReady taint有延迟，导致scheduler无法马上知晓该node上有NotReady taint，从而在node是NotReady的情况下仍然在该node分配pod。

修复的方式是，增加一个api-server的admission plugin来检查新node创建和更新时的消息（admission plugin拦截尝试创建、修改和删除资源的请求），并且检查该node是否有NotReady taint，如果没有则为它添加上NotReady taint。