LCE SignalContainer

在YARN中ContainerExecutor负责container的初始化、启动以及cleanup。YARN提供了多种ContainerExcecutor，通过属性：yarn.nodemanager.container-executor.class配置：

1. DefaultContainerExecutor，每个Container运行在单独的进程里，但是进程都是由NM的用户启动的，比如NM进程就是用yarn用户启动的，那么所有container进程也是用yarn用户启动。
2. LinuxContainerExecutor，每个contaienr由不同的用户启动，例如A用户提交job的container，由A用户启动。此外支持单独的配置文件及简单的ACL。
3. DockerContainerExecutor，NM将container进程运行在Docker容器中，可以提供自定义，与外部NM隔离的运行环境。

LCE和DCE相比，明显隔离性更好，对于LCE来说：

1）需要Linux Native程序的支持，container-executor程序，使用C来实现。

2）需要一个配置文件container-executor.cfg，这个配置文件和container-executor的二进制文件的相对路径还是固定的。

3）用不同的用户启动container，必须有对应的Linux用户存在，因此增加用户时，在所有的NM节点上都要对应增加

4）container-executor和container-executor.cfg的所有者都必须是root，而且他们的目录上溯到/，所有者必须是root

5）container-executor的文件权限必须是6050（--Sr-s---），因为其执行原理就是setuid/setgid。Group owner必须和启动的NM的用户同组。例如NM由yarn用户启动，yarn用户属于hadoop组，那container-executor必须也是hadoop组。

启动container本质上就是运行一个脚本，可以简单理解DCE是直接*/bin/bash $\**，但是LCE使用的是./container-executor $\*，额外包了一层。

如果直接执行container-executor:

*# ./container-executor*

*Usage: container-executor --checksetup*

*Usage: container-executor --mount-cgroups hierarchy controller=path...*

*Usage: container-executor user yarn-user command command-args*

*Commands:*

*initialize container: 0 appid tokens nm-local-dirs nm-log-dirs cmd app...*

*launch container: 1 appid containerid workdir container-script tokens pidfile nm-local-dirs nm-log-dirs resources*

*signal container: 2 container-pid signal*

*delete as user: 3 relative-path*

下面是NodeManager的进程图：



container-executor.c的执行流程

a）launch\_container\_as\_user <= LAUNCH\_CONTAINER

1. 获取script\_file\_dest，执行脚本laucher\_container.sh的目录
2. 获取cred\_file\_dest，安全认证的文件，例如tokens文件目录
3. 获取exit\_code\_file，退出异常文件
4. 获取container\_file\_source，launch\_container.sh
5. 获取cren\_file\_source，\*.tokens
6. 创建child\_pid，folk一个进程
7. 设置sid，session id
8. 将pid写入文件
9. 如果配置cgroups，将pid写入到task
10. 在所有的磁盘上初始化user目录
11. 在所有的磁盘上初始化log目录
12. give up root privs，change\_user
13. 创建container目录
14. 复制脚本及token文件

b) signal\_container\_as\_user <=CLEANUP\_CONTAINER

1) changer\_user

2) kill (-pid,0)

3) kill(has\_group)

https://major.io/2010/03/18/sigterm-vs-sigkill/

https://www.ibm.com/developerworks/cn/linux/l-cn-rootadmin1/