What should be PID 1 in a container ?

Ranjith Rajaram,
Sr. Principal Technical Support Engineer,
Red Hat

Agenda: What should be PID 1 in a container ?

- PID namespace
- Which process should be PID 1 ?
- Does it matter which process has pid 1 in a container ?
- Process reaping and quick hack
- Minimal init/systemd inside a container

What should be PID 1 in a container ?

PID namespace

- PID namespaces isolate the process ID number space, meaning that processes in different PID namespaces can have the same PID
- The first process created in a new namespace has the PID 1
- Process created using clone(2) with the CLONE_NEWPID flag

What should be PID 1 in a container ?

Controlling which process should be PID 1

Process section in runc spec file
controls which process is started

Snip from the spec file

```
"process": {
                "terminal": false,
                "user": {},
                "args": [
                "/usr/sbin/httpd",
                 "-D",
                "FOREGROUND"
                ],
```

Httpd container

What should be PID 1 in a container ?

Controlling which process should be PID 1

CMD directive in Dockerfile.

- CMD ["executable","param1","param2"]
- CMD ["param1","param2"] (as default parameters to ENTRYPOINT)
- CMD command param1 param2 (shell form)

Note:
Option 1 is the widely used one

Dockerfile: Option1

```
FROM fedora:latest
USER root
RUN yum install httpd
EXPOSE 80
# Start the service
CMD ["/usr/sbin/httpd", "-D","FOREGROUND"]
```

What should be PID 1 in a container ?

Controlling which process should be PID 1

Containers started with different options

Option 1:args ["/usr/sbin/httpd", "-D","FOREGROUND"]

```
runc exec -t httpd ps alx
F   UID   PID  PPID PRI  NI    VSZ   RSS WCHAN  STAT TTY        TIME COMMAND
4     0     1     0  20   0 236156  5336 poll_s  Ss+  ?          0:00 /usr/sbin/httpd -D FOREGROUND
5    48     9     1  20   0 246572  3556 inet_c  Sl+  ?          0:00 /usr/sbin/httpd -D FOREGROUND
```

Httpd has pid 1

Option 2: args [“/top.sh”]

```
runc exec -t top ps alx
F   UID   PID  PPID PRI  NI VSZ   RSS  WCHAN  STAT TTY        TIME COMMAND
4     0     1     0  20   0  11628  1316 wait     Ss+  ?          0:00 /bin/bash /top.sh
4     0     9     1  20   0  51764  1868 poll_s   S+   ?          0:00 /usr/bin/top -b -d 5
4     0    24     0  20   0  45316  1520 -        Rs+  ?          0:00 ps alx
```

Bash has pid 1

What should be PID 1 in a container ?

Does it matter which process has PID 1 inside a container

Quick demo using two containers. “web.py” is simply python based httpd server

```
"process": {
                "terminal": false,
                "user": {},
                "args": [
                "/web.py"

                ],
```

Container 1 spec file

```
"process": {
                "terminal": false,
                "user": {},
                "args": [
                "/web.sh"
                ],
```

Container 2 spec file

```
#!/bin/bash

/web.py &
PID=$!
trap "kill $PID" INT TERM
wait
```

Container 2: file web.sh

What should be PID 1 in a container ?

Process reaping and quick hack

Role of PID 1

Process ID 1, which is normally the UNIX ‘init’ process, has a special role in the operating system. If parent of a child process dies before it exits, it is responsibility of the init process to reap the child and clear system kernel process table

Side effect of rogue containers

If cleanup of orphaned processes fail, it can fill up the kernel process table. Typically sysctl setting of a system

```
#sysctl -a | grep pid_max
kernel.pid_max = 32768
```

What should be PID 1 in a container ?

Process reaping and quick hack:- contd

Shell as pid 1

‘ /bin/sh’ will reap orphan child processes and prevents zombie processes from filling up the kernel process resource table.

Downside with shell as pid 1

It will not propagate signals properly.

Workaround: Trap the signal and pass it to the application for a clean exit.

What should be PID 1 in a container ?

systemd/minimal init

Minimal init

Tini: https://github.com/krallin/tini

Dumb-init: https://github.com/Yelp/dumb-init

CMD ["dumb-init", "python", "web.py"]

This creates a process tree that looks like:

- docker run (on the host machine)
  - dumb-init (PID 1, inside container)
    - python web.py (PID ~2, inside container)

What should be PID 1 in a container ?

systemd/minimal init

Docker-init

Docker 1.13 adds **--init** flag on dockerd and “docker run” to run a zombie-reaping init process as PID 1 inside the container. We also get **--shutdown-timeout** to shutdown containers gracefully during daemon exit and **--stop-timeout** for individual containers.

```
# docker exec -it devconf1 ps aux
USER   PID %CPU %MEM   VSZ   RSS TTY   STAT START   TIME COMMAND
root     1  0.0  0.0     1148     4     ?     Ss     16:49     0:00     /dev/init /web
root     5  0.0  0.6   110540 13088 ?       S      16:49     0:00     /usr/bin/python
root    30  1.0  0.1    47448  3388   ?       Rs+    16:53     0:00     ps aux
```

What should be PID 1 in a container ?

systemd/minimal init:- contd

Why not Systemd ?

Systemd can run inside a container without the privileged mode.

Additional benefit:

1. Better handling of logging
2. Default service init and handling of order etc

OCI hooks

```
"hooks": {
  "prestart": [
              {
                "path": "/usr/libexec/oci/hooks.d/oci-systemd-hook"
              }
            ]
  },
```

What should be PID 1 in a container ?

Questions

Questions

What should be PID 1 in a container ?

References

Running systemd in a non-privileged container : Daniel Walsh

Issues with running as PID 1 in a Docker container : Graham Dumpleton


CLOUD
NATIVE
CON
Europe 2017
KubeCon
A CNCF EVENT
Thank you !!!!