RT in User namespaces

Prakash Sangappa

ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

User Namespaces

- Enables non root users to create namespaces

- Non root user mapped to root user(UID 0) inside.

- Gets root privileges/capabilities inside the namespace including CAP_SYS_NICE

- Capabilities not effective in changing/setting RT priority

User Namespaces

- Capabilities only applicable to resources inside namespace
- Restriction also on other capabilities like IPC_LOCK, SYS_TIME, MKNOD etc, affecting global resources.
- Mapping root user from init namespace(UID 0) into User namespace still has same restrictions.
- Deal with them on case by case basis?

RT priority in User Namespaces

- Usecase: Multitenant Oracle DB – Uses User namespace
- Multitenant Oracle DB requires running some processes with RT priority inside namespace, but cannot.
- Same limitation with Linux(lxc) unprivileged containers.

Multitenant Oracle DB

- Architecture to enables Oracle Database to be multitenant Container Database(CDB)

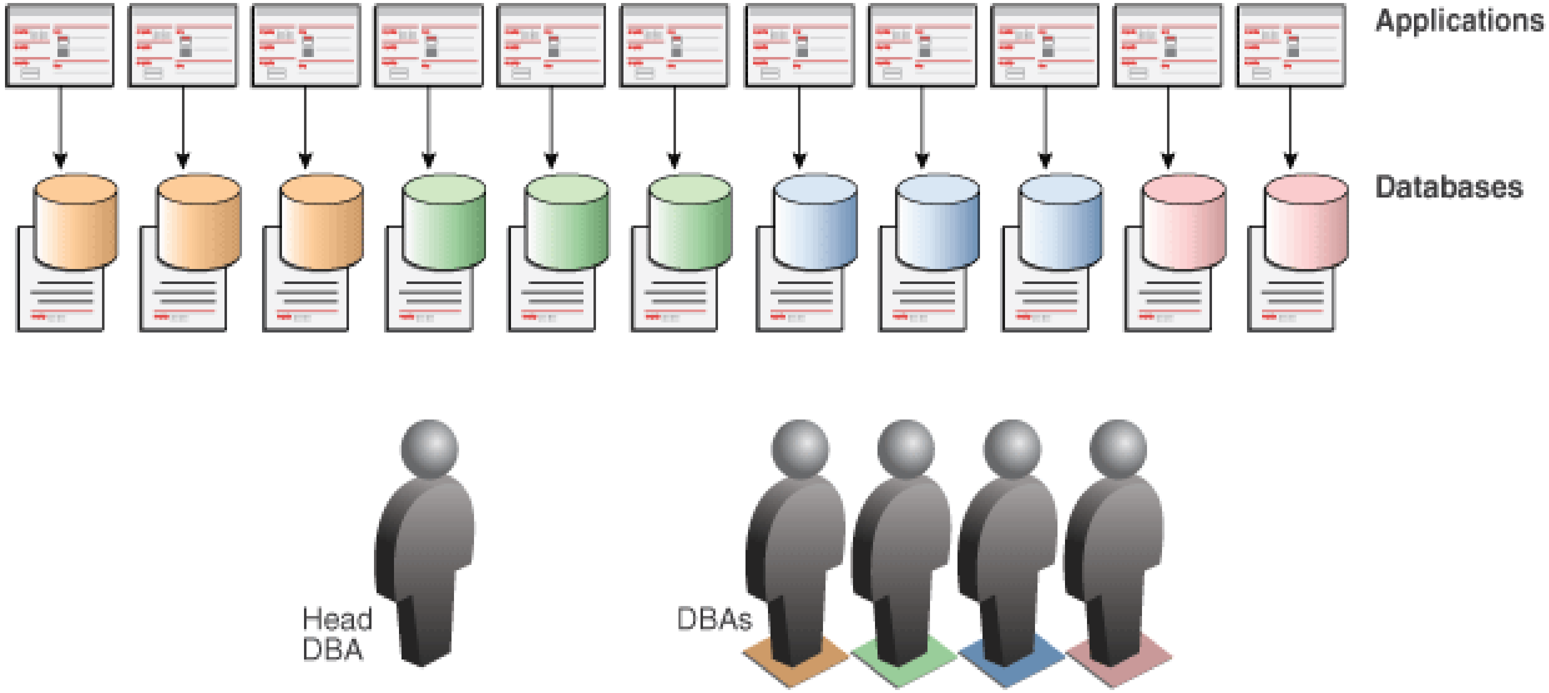
- CDBs have zero or many customer pluggable databases(PDB

- Benefit - Database Consolidation.

ORACLE®

Multitenant Oracle DB

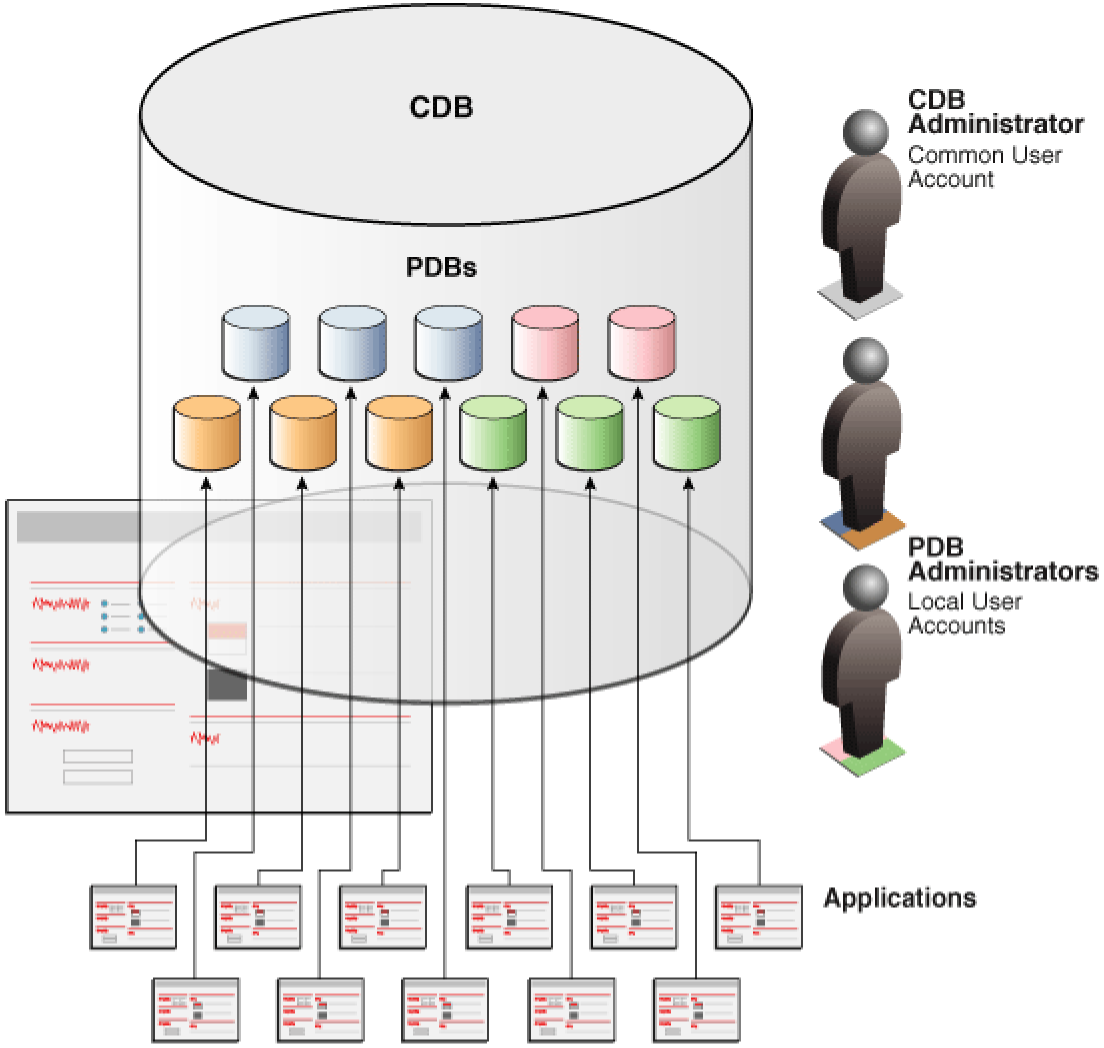
- Database environment before Database consolidation
  Multiple databases on each server



Multitenant Oracle DB

- Database consolidation
  CDB/PDB

- Easier management
  Cost reduction



Multitenant Oracle DB

- There can be multiple CDBs running on a system.

- For Security & Isolation, PDBs and CDBs are sandboxed using Usernamespace + other namespaces(PID, mount, Net)

- CDB runs in top level Usernamespace. PDBs are in nested User namespaces inside.

- Multiple CDBs on a system

Multitenant Oracle DB

- There can be multiple CDBs running on a system.
- For Security & Isolation, PDBs and CDBs are sandboxed using Usernamespace + other namespaces(PID, mount, Net)
- CDB runs in top level Usernamespace. PDBs are in nested User namespaces inside.
- Multiple CDBs on a system

Multitenant Oracle DB

- CDBs have critical processes(Ex Logwriter) that need to run with highest priority(above all other user process priority).

- Critical processes are run with RT priority

- CDBs unable to set RT priority due to User namespace restrictions

- A solution – use help of a daemon/process in parent(init) namespace to change/set RT priority – not convenient

Possible Approaches

- Allow root(uid 0) from init namespace mapped into User namspace to set RT priority (/etc/subuid – Testuser:0:1)

- Permit CAP_SYS_NICE capability if an User namespace is tagged.

- With use of cgroup controls, allow RT priority privileges to UID 0 in User namespace.

- Alternative – Have a fixed high priority scheduler option, above all user priority – avoid RT.

Thank You!