Why Oracle Linux Is the Best Platform for Oracle Database and Oracle Cloud

[PRO5797]

Dhaval Giani
Manager, Linux Kernel Development
Ravi Thammaiah
Director of Software Development
Oracle

October 25, 2018

ORACLE OPEN WORLD

ORACLE


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.

Program Agenda

1. Our Focus Areas
2. *db*Nest
3. Story Time – PID Namespaces

Our Focus Areas

The kernel is a big playground

- We work upstream but strive to deliver value to our customers rapidly
- Oracle Database Enablement
- Oracle Cloud Enablement
- Unbreakable Enterprise Kernel
  - https://github.com/oracle/linux-uek
- Ksplice

Where We Are Focused

- Memory Management
- Observability
- Arm Platform Enablement
- Scheduler
- Scalability and Performance
- Resource Management
- Networking
- Filesystems

Memory Management

- Pages are originally 4k in size
- With main memory rapidly increasing in size, this is leading to many issues
- Older programs assume pages are 4k in size!
- Solved with the use of larger page sizes – 2 MiB, 1GiB

Memory Management

- Transparent Huge Pages (THP)
- Hugetlbfs
- Persistent Memory
  - Filesystem DAX
  - Device DAX
  - What does the future hold?

Memory Management

- lru_lock scalability
  - Lock that protects the LRU data structure
  - Uses an innovative approach
  - lru_lock is one part of the problem, also need to resolve it for zone_lock
  - Linux kernel community adopted similar approach for zone_lock

Security

- Smatch
  - Static analysis
  - https://repo.or.cz/w/smatch.git
- libseccomp
- High impact vulnerability mitigations
  - Ksplice → rebootless updates

Scheduler

- Improve scalability of scheduler load balancing
- Capacity aware scheduling
  - Avoid CPUs with high interrupt and RT load
- Adaptive pipes with busy waiting capability
  - Make spin time a tunable

Performance

- Ktask
  - VM bootup performance improvements
  - DB shutdown improvements
- Bootup improvements
- gettimeofday improvements
- General system improvements
- All of these make the Database faster!

Oracle Linux and Oracle Database

Areas of Innovation - highlights

- Memory Management
  - Shared Memory
  - Private Memory
  - Semi-shared Memory
  - NUMA
  - Persistent Memory
- File System
  - XFS new features
- Scheduler
  - MP/MT/MPMT/UT
- System Security
  - Namespace/dbNest
- Synchronization
  - Unified wait
  - Post wait optimization
- Networking
  - High speed networking
  - Infiniband/RDMA/RDS
  - RoCE
- IO
  - Subsystem and new API

*db*Nest

*db*Nest

- Isolation
- Resource Management
- Restricted System Calls
- Lightweight Virtualization
- Filesystem Isolation
- Hierarchical user privileges

*db*Nest

- DB Nest will provide an isolated environment so that
  - Instance external environment is sandboxed
  - Instance runs in a protected environment
  - Hierarchical structure provides environment protection
    - Instance/App-PDB/PDB/User level
  - Hierarchical resource management for nests

*db*Nest

- DB Nest part of RDBMS
- Protections include
  - PID, User ID, File-system and Mount Points, Network
  - All other resources
- Resource management include
  - CPU, Memory

*db*Nest – Technology inside

- User namespaces for OS component Isolation
- Control groups for Resource management
- Pivot root/bind mount for File system isolation
- Network namespaces for isolating network stack
- Capabilities for controlled access
- Seccomp for system call filters

Story Time – PID Namespaces

October 22–25, 2018

SAN FRANCISCO, CA

ORACLE OPEN WORLD

#OOW18

oracle.com/openworld

ORACLE