DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix.net ITPUB

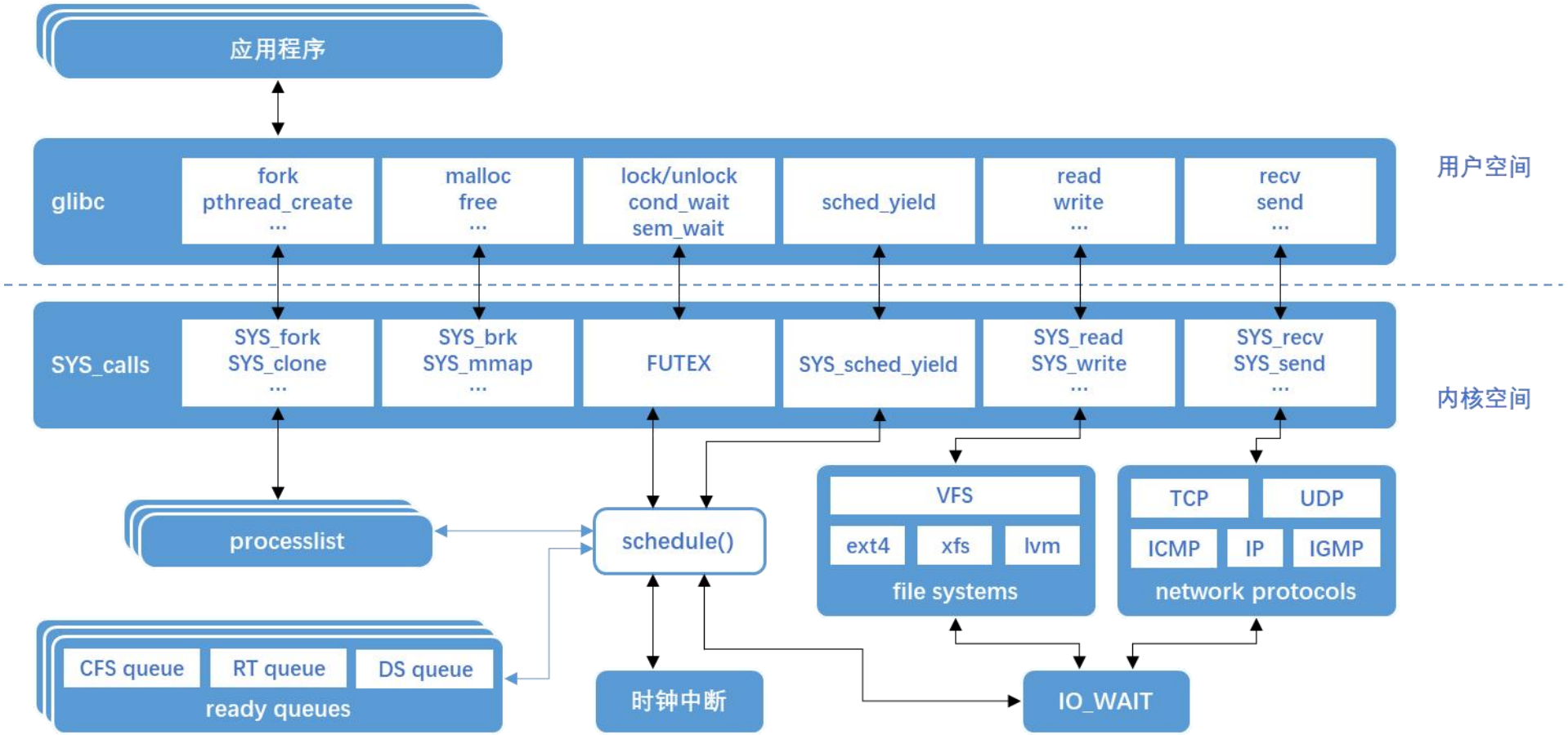
北京国际会议中心 | 2023/8/17-19

StarSQL
打造纯用户态数据库

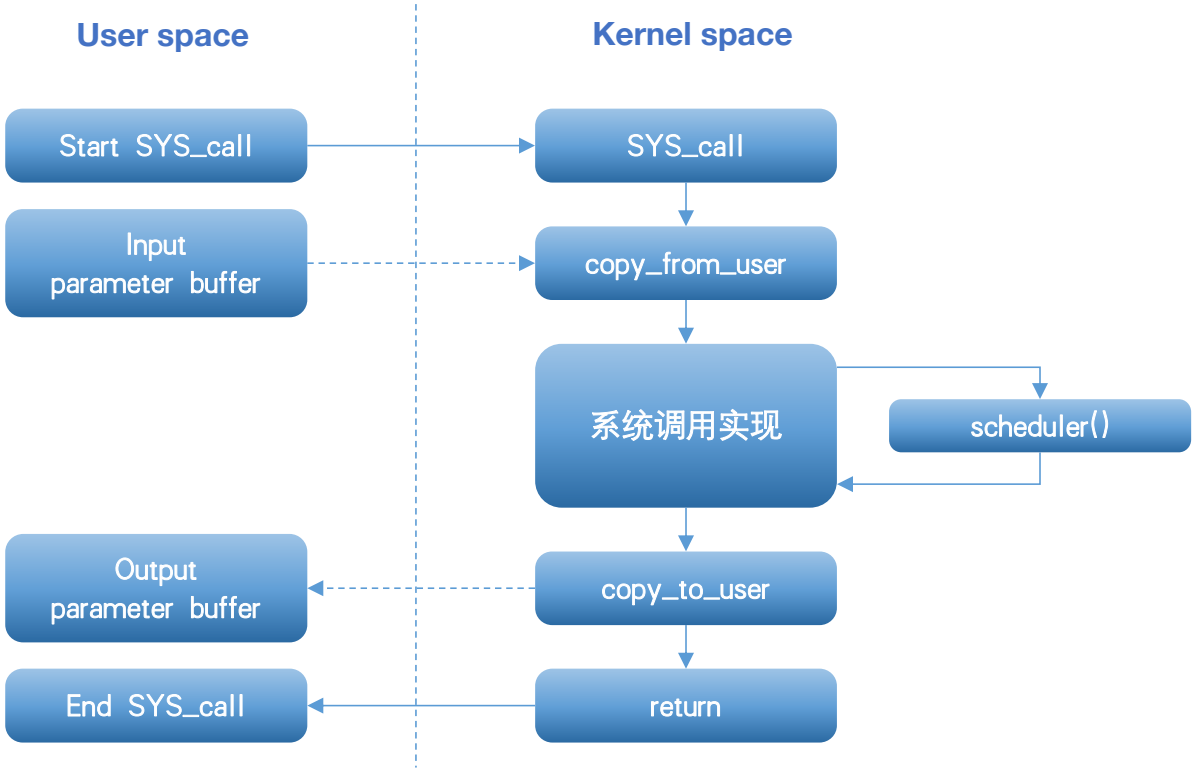
京东科技
数据库架构师
孟祥滨

目录

Linux系统调用和内核调度器

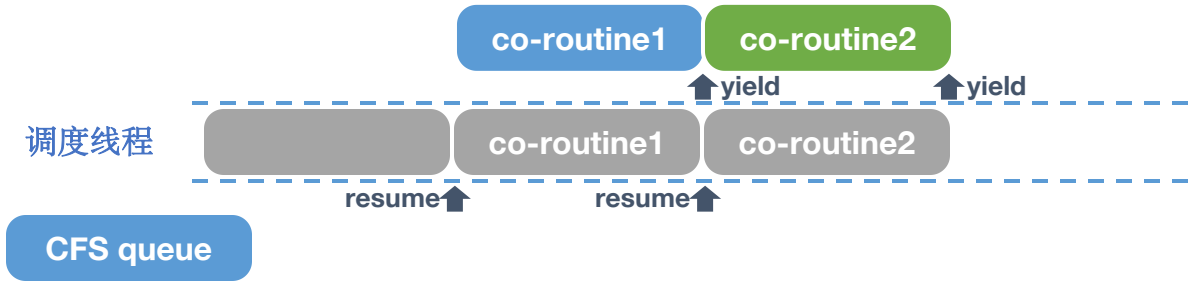


系统调用的执行过程和代价

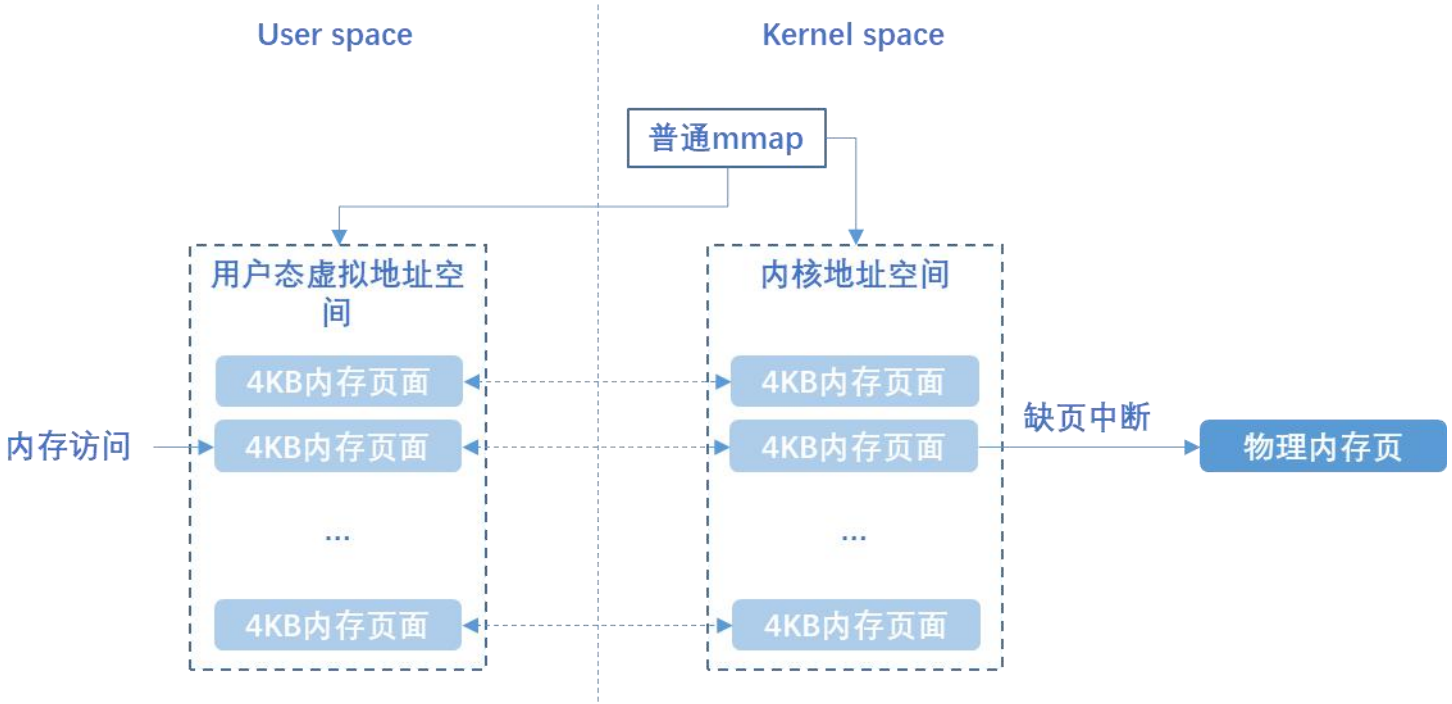


目录

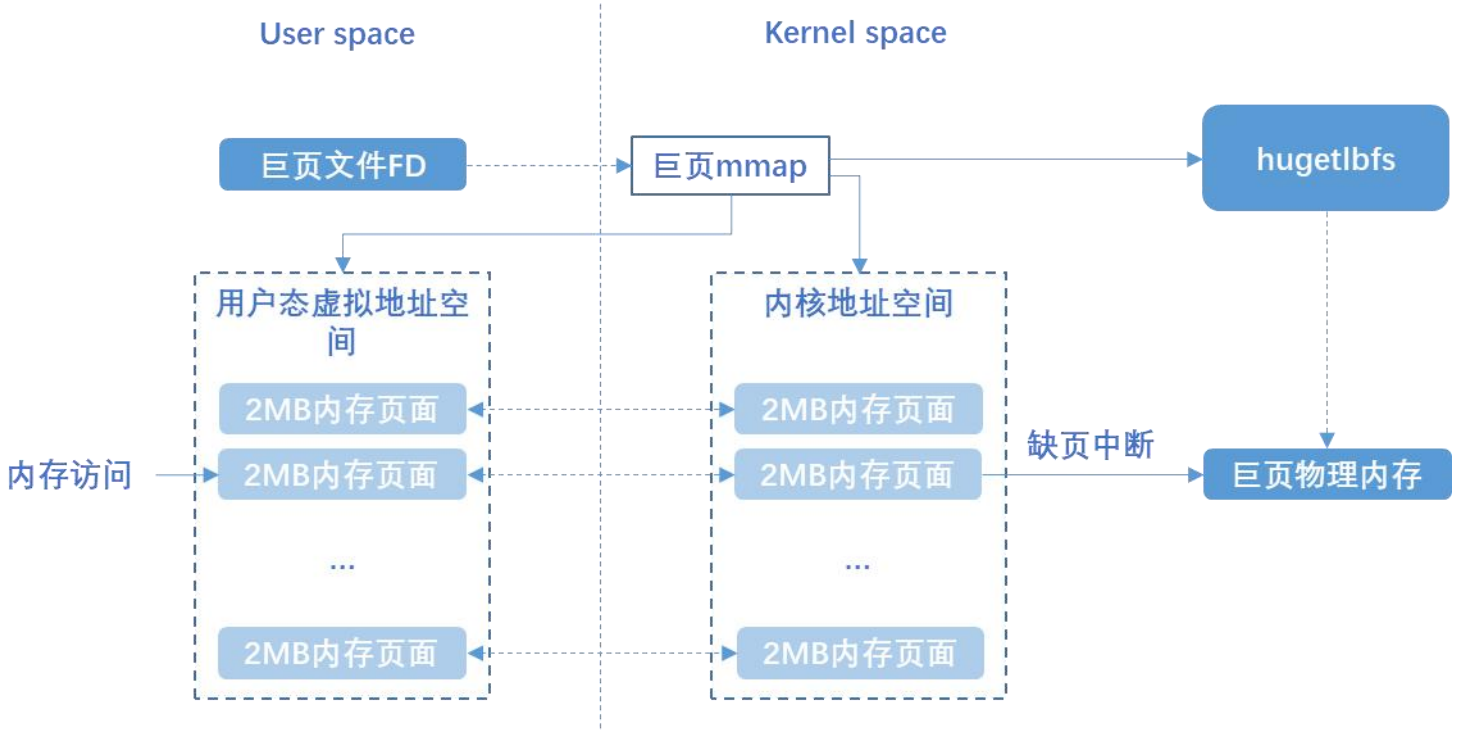
协程化调度


co-routine1
co-routine2
yield
yield
调度线程
co-routine1
co-routine2
resume
resume
CFS queue

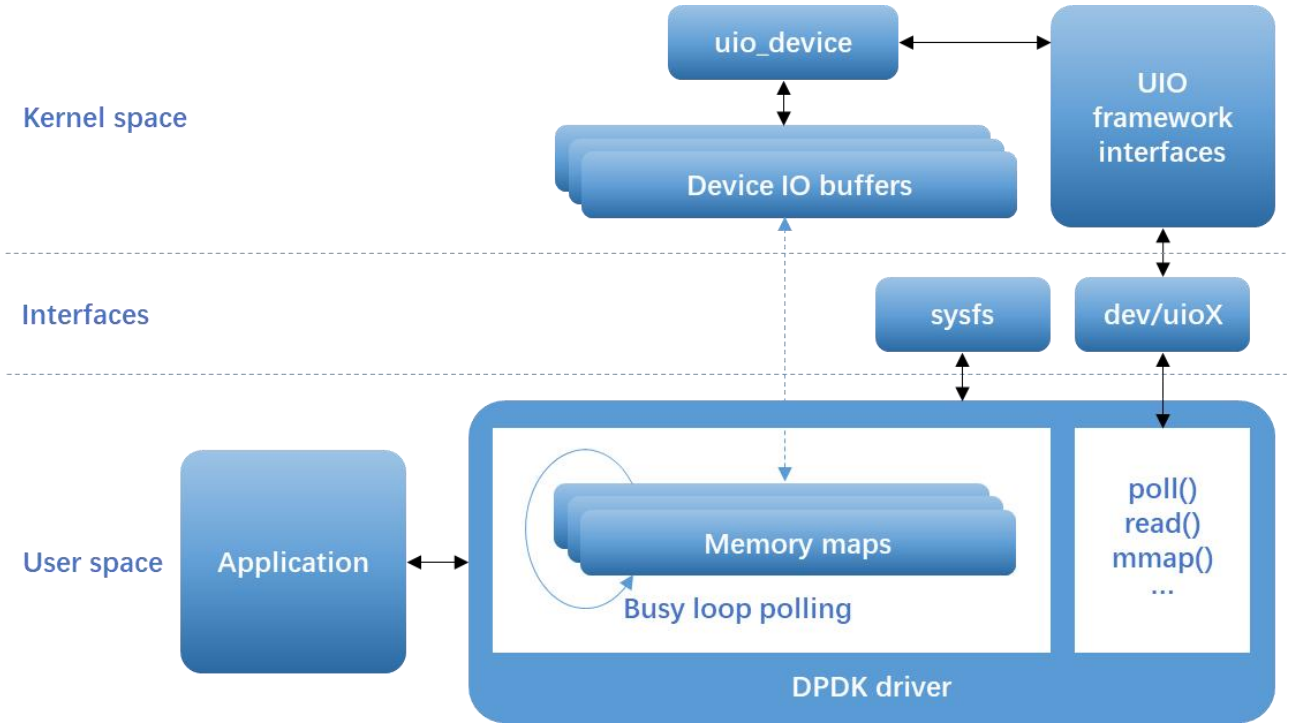
内存映射技术——mmap


User space
Kernel space
普通mmap
用户态虚拟地址空间
内核地址空间
4KB内存页面
4KB内存页面
...
4KB内存页面
4KB内存页面
4KB内存页面
...
4KB内存页面
内存访问
缺页中断
物理内存页

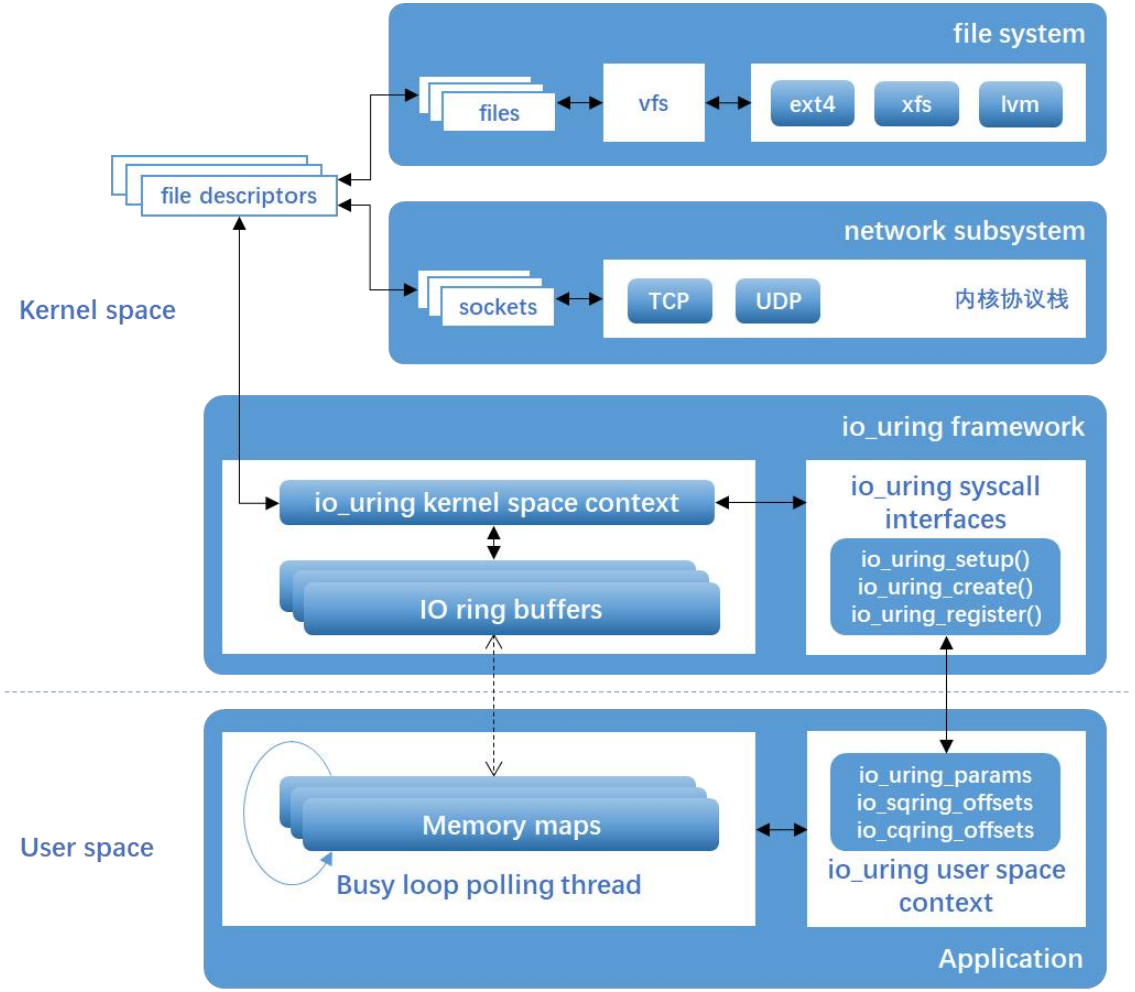
内存映射技术——巨页mmap



用户态驱动（DPDK/RDMA）



异步IO框架： io_uring



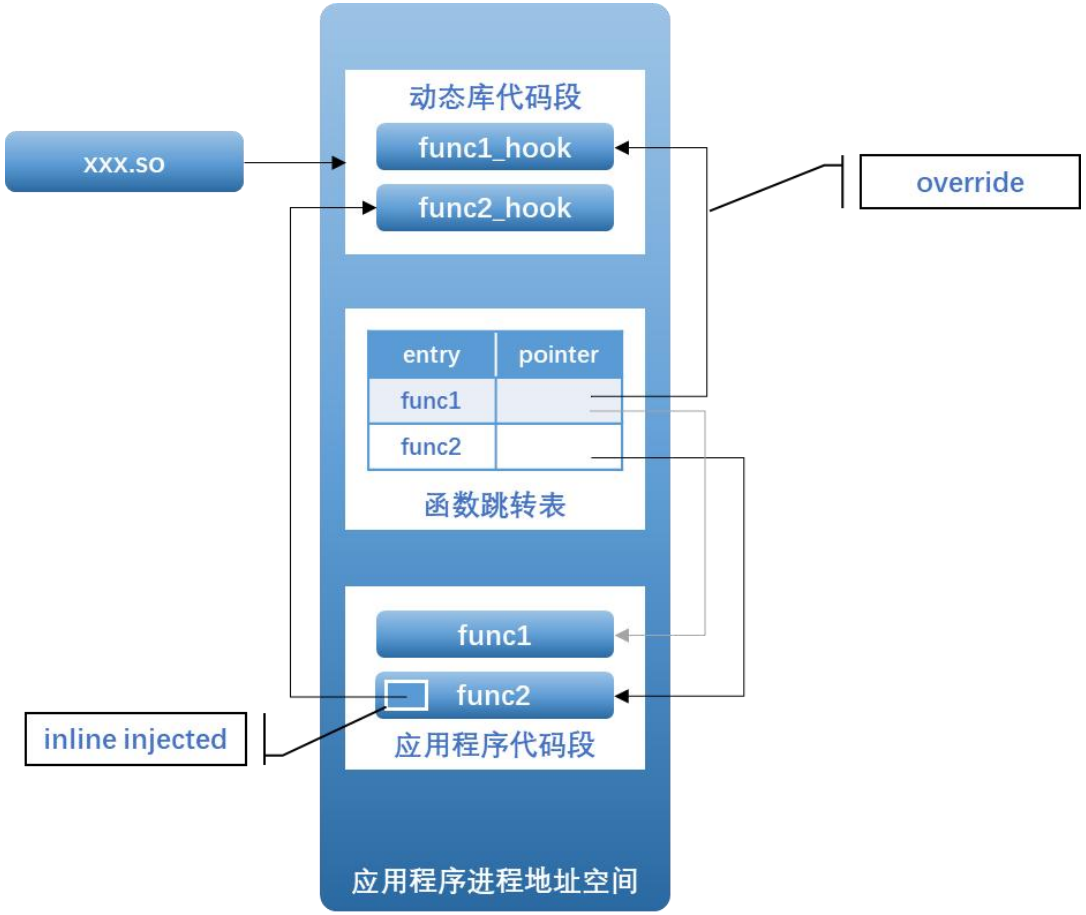
纯用户态编程的特性和技术壁垒

- 诸如mmap和io_uring，本质上都是一种异步系统调用机制
- 应用程序不再直接陷入内核、触发调度器，并以阻塞模式等待内核态代码的执行
- 应用程序利用内存映射直接与内核交互、提交请求，内核态代码则以异步方式执行请求
- 应用程序所在进程几乎不执行任何内核态代码

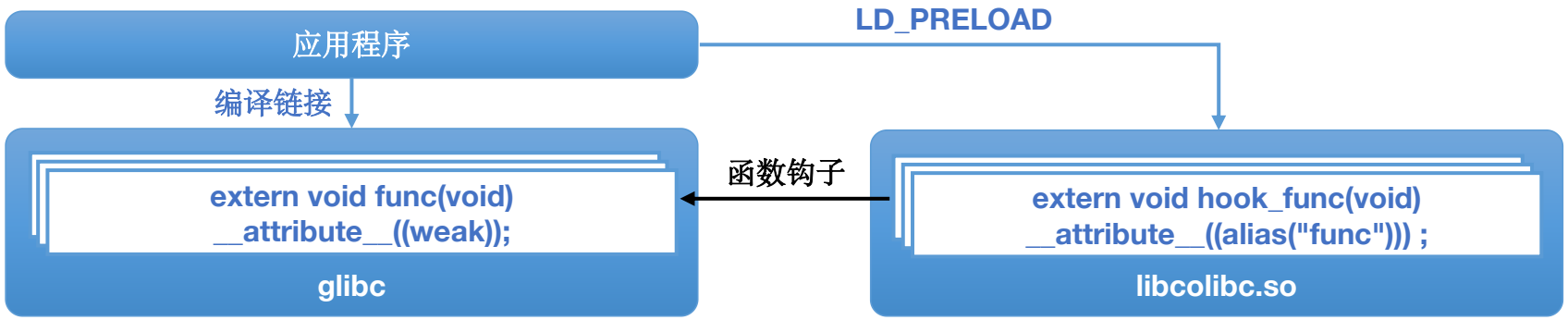
- DPDK/SPDK类用户态驱动技术，缺乏完备的协议栈/文件系统生态
- io_uring框架可以充分利用Linux内核协议栈/内核文件系统生态，但不提供标准POSIX开发接口
- 一个标准的POSIX程序，可以“透明”使用纯用户态技术吗？

目录

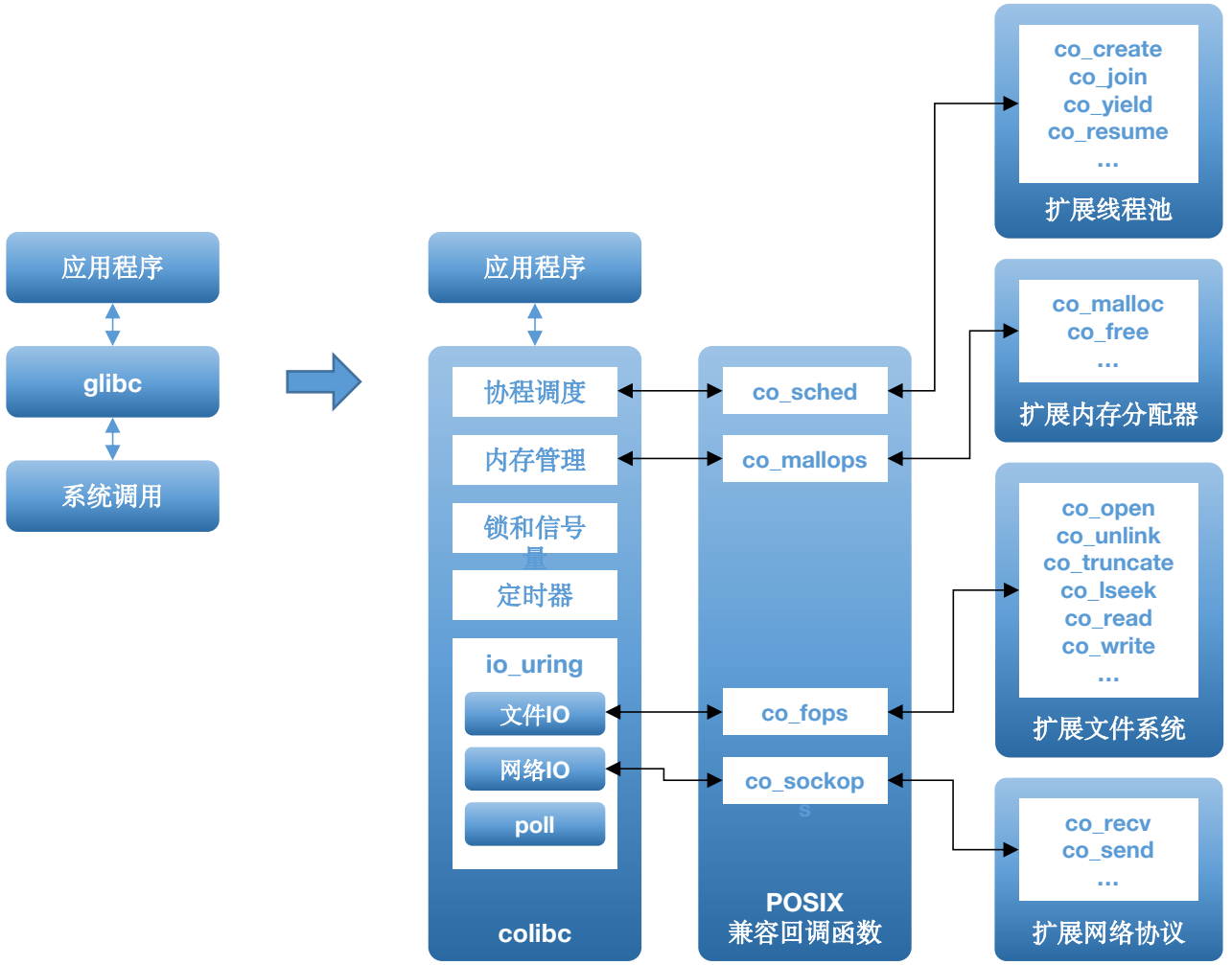
动态库注入技术


xxx.so
动态库代码段
func1_hook
func2_hook
override
entry
pointer
func1
func2
函数跳转表
func1
func2
inline injected
应用程序代码段
应用程序进程地址空间

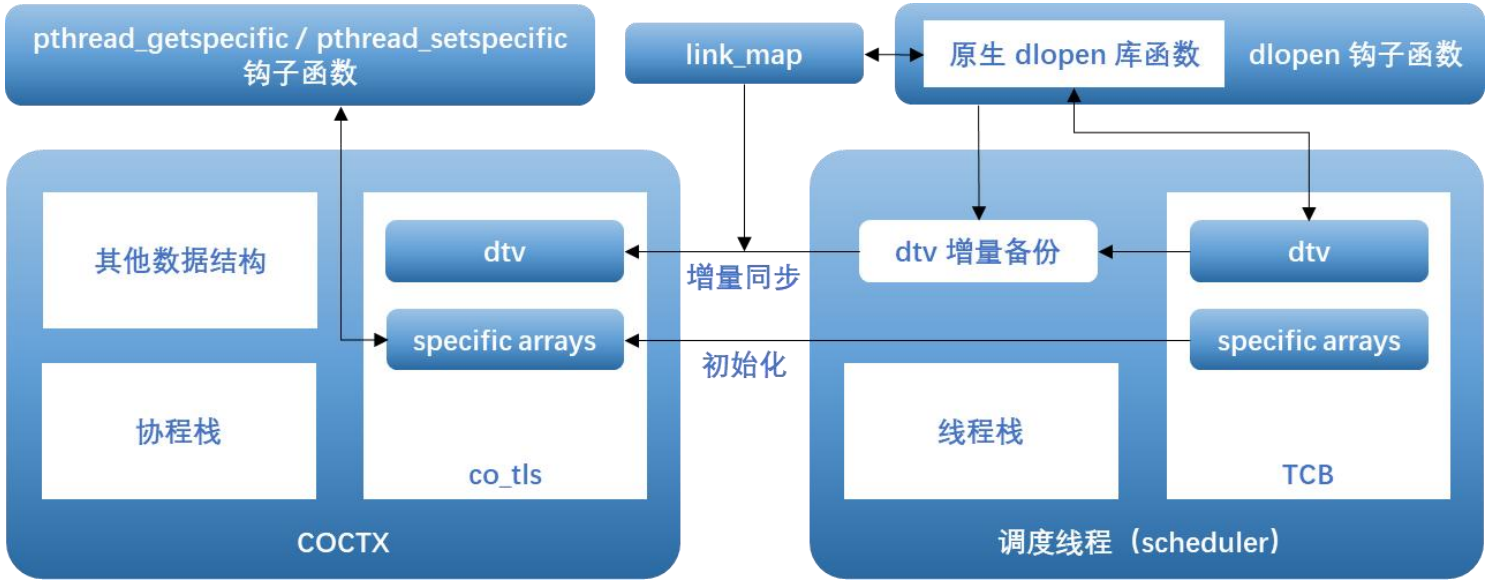
运行时注入——LD_PRELOAD



Colibc的架构和可扩展性



Colibc的关键技术：TLS处理

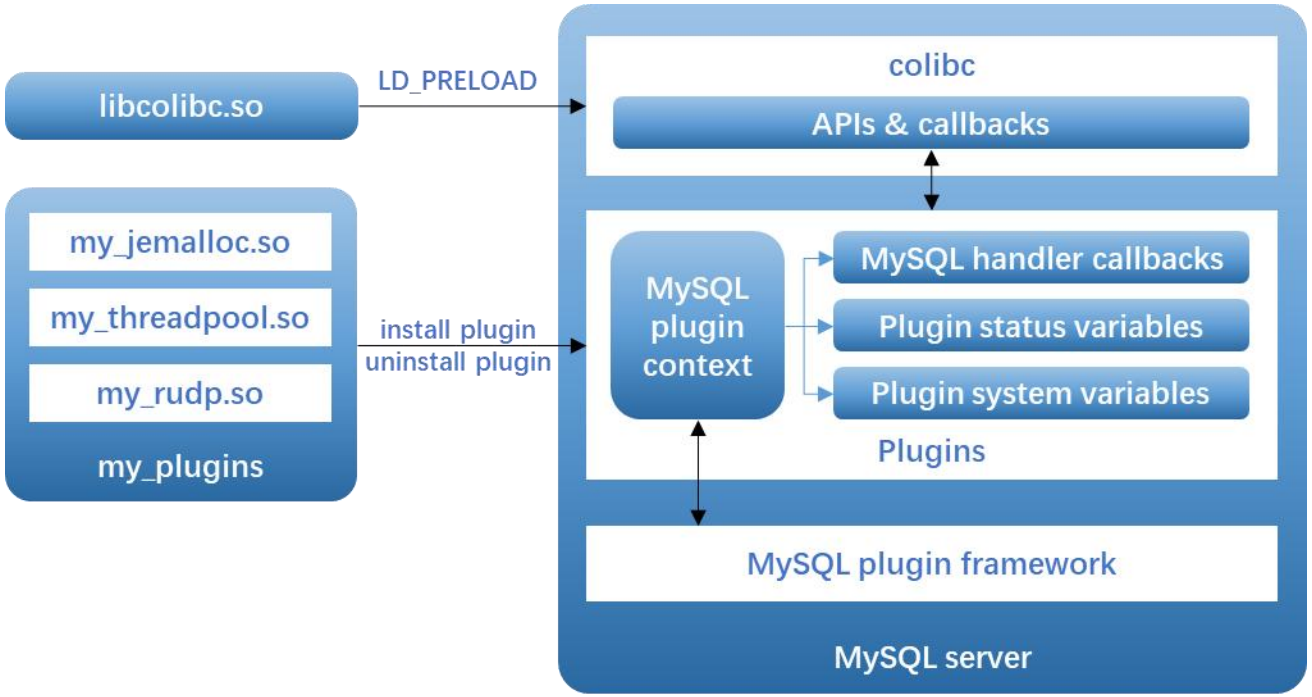


Colibc的关键技术：无锁内存分配

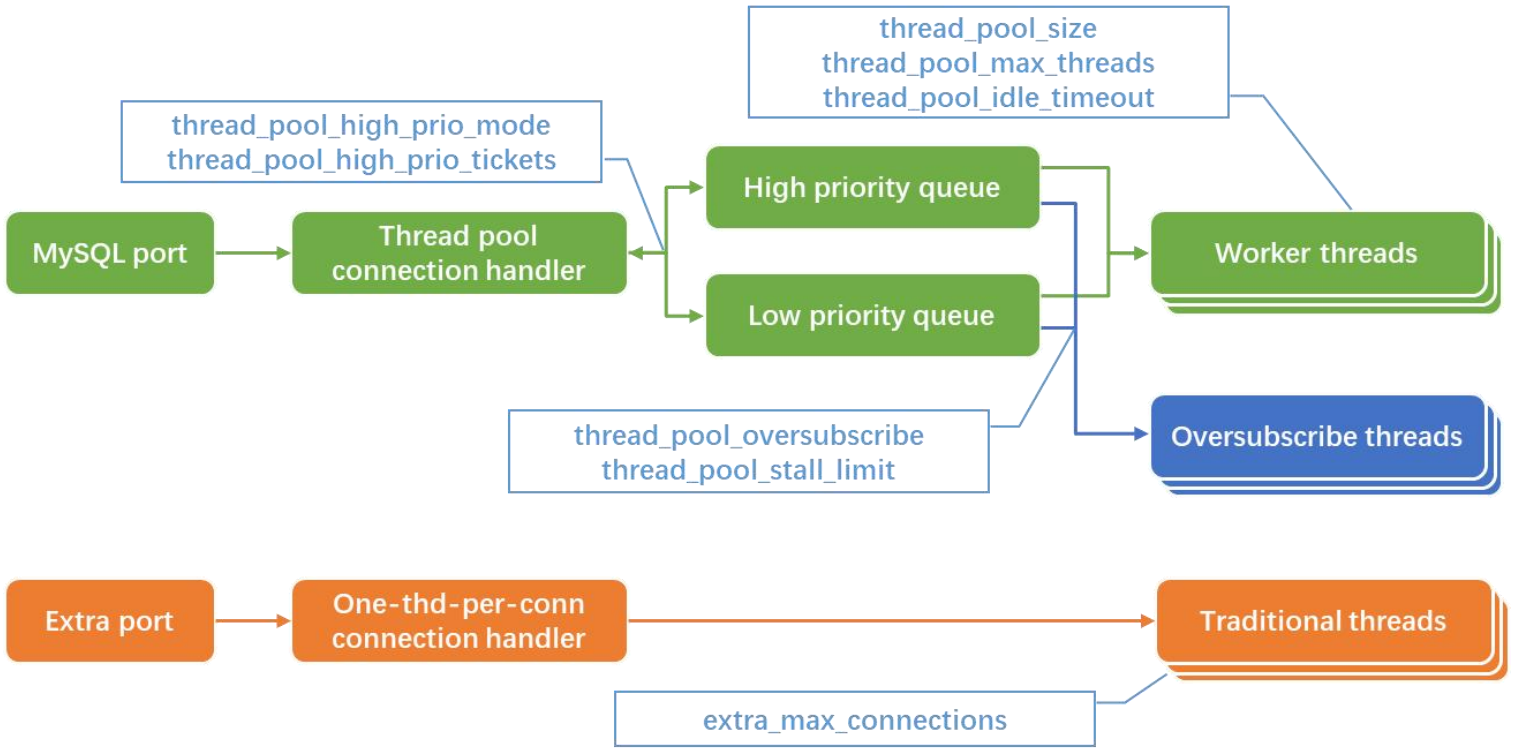
VS.

目录

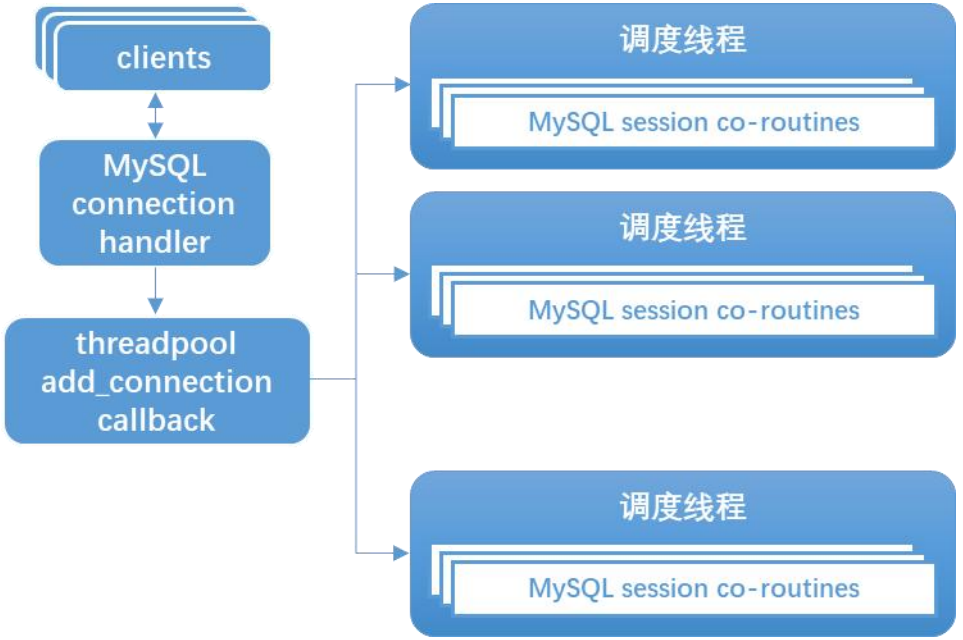
基于colibc的MySQL插件



MySQL现有的线程池方案



基于colibc的线程池插件



线程池插件运行中的状态


```
top - 16:50:13 up 279 days, 20:55,  1 user,  load average: 0.03, 0.05, 0.00
Threads:  27 total,   0 running,  27 sleeping,   0 stopped,   0 zombie
%Cpu(s):  0.1 us,  0.0 sy,  0.0 ni, 99.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 26401875+total, 22593728+free,  2262168 used, 35819304 buff/cache
KiB Swap: 16777212 total, 16777212 free,        0 used. 25898193+avail Mem

   PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+ COMMAND
129439 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.31 mysqld
129448 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129449 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129450 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129451 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129452 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129453 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129454 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129455 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129456 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129457 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129458 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129459 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129460 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129462 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129463 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.01 mysqld
129464 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129465 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129466 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129467 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129468 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129469 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129470 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129471 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129472 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129473 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
129474 matthew   20   0 2027944 293228  17896 S  0.0  0.1   0:00.00 mysqld
```

```
top - 16:48:48 up 279 days, 20:53,  1 user,  load average: 0.12, 0.07, 0.01
Threads:  47 total,   0 running,  47 sleeping,   0 stopped,   0 zombie
%Cpu(s):  0.1 us,  0.0 sy,  0.0 ni, 99.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 26401875+total, 22559376+free,  2605652 used, 35819348 buff/cache
KiB Swap: 16777212 total, 16777212 free,        0 used. 25863844+avail Mem

   PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+ COMMAND
129133 matthew   20   0 2624868 618272  18172 S  0.3  0.2   0:00.07 inno_er_monitor
129099 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.38 mysqld
129100 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 background
129101 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 timer_notify
129102 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129103 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129104 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129105 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129106 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129107 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129108 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129109 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129110 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129111 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129112 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129113 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129114 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129115 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129116 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129117 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 scheduler
129119 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129120 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129121 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129122 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129123 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129124 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129125 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129126 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_io_handler
129127 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_pg_c_coord
129128 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_pg_cleaner
129129 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_pg_cleaner
129130 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_pg_cleaner
129132 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_lock_timeo
129134 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_monitor
129135 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_master
129136 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 background
129137 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_pur_coord
129138 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_pur_worker
129139 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 background
129140 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 background
129141 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.01 inno_buf_dump
129142 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_stats
129143 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_fts_opt
129144 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 inno_buf_resize
129145 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 ack_receiver
129146 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 signal_handler
129147 matthew   20   0 2624868 618272  18172 S  0.0  0.2   0:00.00 compress_gtid
```

目录

总结和展望

- 透明（MySQL源码0侵入）
  - 快速功能回滚（启动时不注入即可）
  - 功能演进独立于MySQL源码，用户可持续集成新的开源社区版本
- 高性能
- 可扩展

- 更多的扩展（开发更多的插件、适配其他数据库）
- 资源管理（内存占用率、CPU占用率、调度优先级、存储生命周期等等）
- 兼容性（POSIX兼容、跨平台、跨系统）
- 虚拟化（完备的运行时虚拟机）

THANKS

DatabaseKernel
TemporalData
TDDL
DistributedTable
CloudnativeData
DBproxy
Hyperbase
Hubble
HBase
DataCenter
AIalgorithm
Cassandra
VisualDataPlatform
PostgreSQL
SSD
ArgoDB
Distributed
MongoDB
Blockchain
GreatDB