2023
全球数字经济大会
开放原子
全球开源峰会
OPENATOM GLOBAL OPEN SOURCE SUMMIT
</开源赋能 普惠未来.>

# eBPF 内核技术在滴滴云原生的落地实践

滴滴出行系统软件负责人 张同浩

# eBPF 内核技术在滴滴云原生的落地实践


- ❑ eBPF 内核技术简介
- ❑ 生产环境业务痛点
- ❑ eBPF 平台能力建设
- ❑ 业务接入落地实践
- ❑ 未来规划和展望

# eBPF 内核技术简介

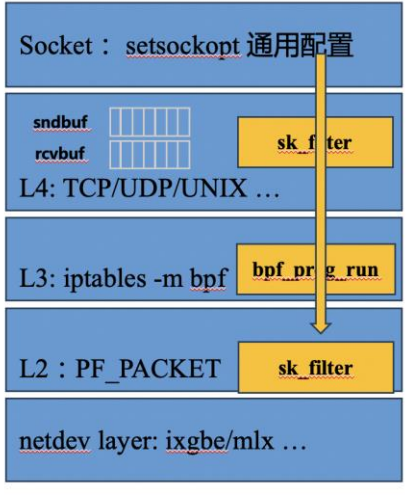
《The BSD Packet Filter: A New Architecture for User-level Packet Capture》

| 元素 | 描述 |
| --- | --- |
| A | 累加器 32bit |
| X | 寄存器 32bit |
| M[] | 数组内存 16x32bit |

**精简指令：加载，存储，跳转，运算四类指令**

```
[root@linux-upstream-dev sk_filter]# cat icmp.asm
ldh [12]            ; Load L2 ethhdr protocol to A
jne #0x800, drop    ; A != 0x800
ldb [23]            ; Load L3 iphdr  protocol to A
jne #1, drop        ; A != 1
ret #-1             ; Accept
drop: ret #0        ; Drop
[root@linux-upstream-dev sk_filter]# bpf_asm -c icmp.asm
{ 0x28,  0,  0, 0x0000000c },
{ 0x15,  0,  3, 0x00000800 },
{ 0x30,  0,  0, 0x00000017 },
{ 0x15,  0,  1, 0x00000001 },
{ 0x06,  0,  0, 0xffffffff },
{ 0x06,  0,  0, 0000000000 },
```

**指令集**




**内核位置**

- ❑ **Seccomp BPF**
- ❑ **Socket filters**
- ❑ **Traffic control cls_bpf**
- ❑ **Team LB 模块**
- ❑ **Netfilter xt_bpf**

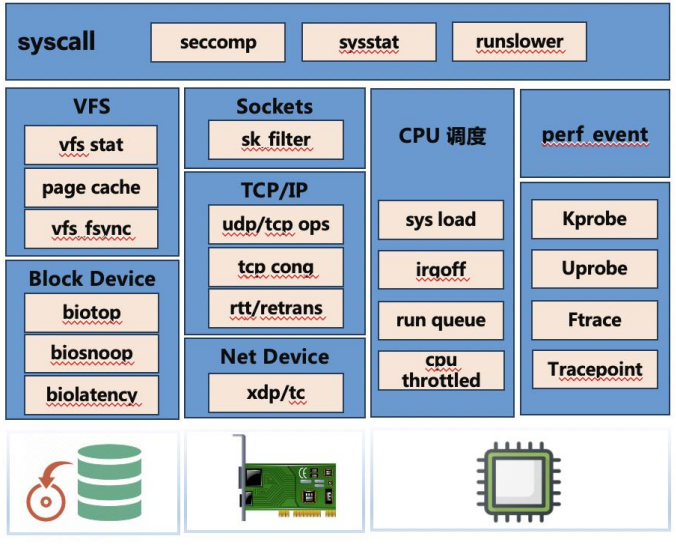
**应用场景**

# eBPF 内核技术简介

| 元素 | 描述 |
| --- | --- |
| R0-R9 | 通用寄存器 10, 64bit |
| RISC | 通用精简指令集 |
| JIT | 宿主原生性能，BPF 设备卸载 |
| bpf_call | 内核交互能力 |
| Map | 数据存储和交换 |

```
$ clang -O2 -g -target bpf -c $(NAME).bpf.c …
```

## 指令集




## 内核位置

- Tracing
- Observability
- Security
- Networking
- Root Cause Analysis
- 开源项目：BCC，Cilium，bpftrace，Falco，Katran，Pixie …

## 应用场景

# 生产环境业务痛点

**服务访问回归**
- ❑ 业务的编程语言，库，版本繁多
- ❑ 定制化难以提升覆盖度
- ❑ 业务有感，稳定性难保障

**服务接口拓扑**
- ❑ 业务种类量大，人工成本高
- ❑ 插桩、特定SDK推广难度大，断链
- ❑ 现有观测能力有限

**容器安全**
- ❑ 用户态难以提升覆盖率
- ❑ 控制访问能力有限
- ❑ 某些场景下性能影响较大

**内核根因定位**
- ❑ 传统工具资源，性能消耗大，常态运行难

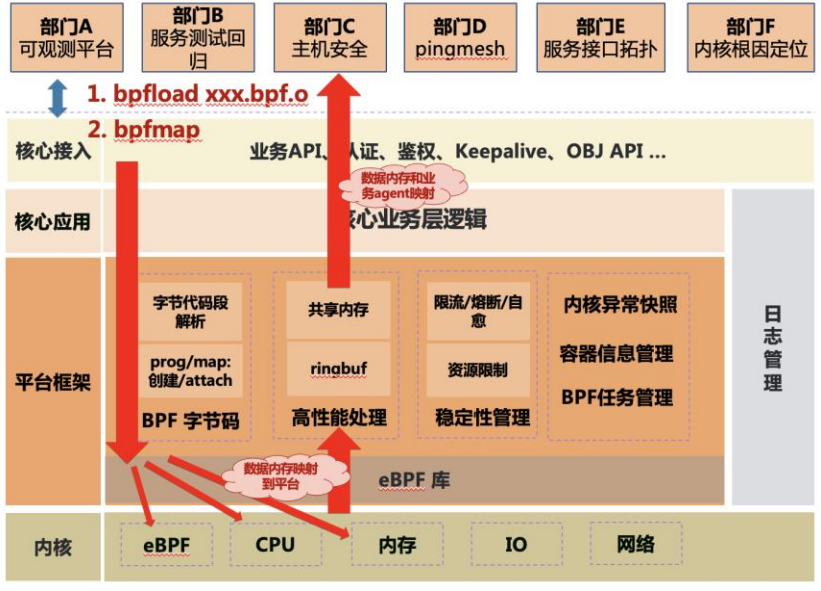
...

**eBPF 能够解决：**
- ❑ **编程语言0侵入：动态插桩，静态插桩等方式**
- ❑ **安全高效性：JIT 机器指令/KV map**
- ❑ **内核可编程：动态追踪无须修改内核**

**落地问题：**
- ❑ **需求多如何提升研发效能**
- ❑ **行业生产环境落地规模小，点到面，如何保证宿主稳定性**
- ❑ **如何统一保障插桩点性能消耗**

**HuaTuo eBPF 平台**

# eBPF 平台能力建设 总架构




**提升研发效能**
- ❑ 屏蔽底层复杂性
- ❑ 37 系统调用 -》2 API
- ❑ 提供BPF编译环境
- ❑ 提供BPF通用函数库

**保障稳定性**
- ❑ 框架层：限流，熔断，自愈
- ❑ BPF 内核层限速
- ❑ 内核热补丁修复BUG

**提供业务视角**
- ❑ BPF OBJ方式
- ❑ 平台和业务逻辑分离
- ❑ 业务决定灰度发布

**保障性能**
- ❑ 事件触发，数据链路零拷贝

# eBPF 平台能力建设 标准化

## 逻辑解耦

- ❑ 平台提供共享服务，通用能力
- ❑ 业务只需要关注自身特定逻辑
- ❑ BPF 字节码，能够解决逻辑解耦，代码分离
- ❑ 轻量级 API 接口，根据BPF 字节码自动化加载

## 标准化

- ❑ 标准化ELF SEC 定时形式
- ❑ 兼容已存BPF字节码，其他字节码无须修改
- ❑ 兼容其他BPF 编译器

```
socket
sk_reuseport
kprobe+
uprobe+
kretprobe+
uretprobe+
usdt+
tracepoint+
tp+
raw_tracepoint+
raw_tp+
tc
classifier
lsm+
syscall
xdp
perf_event
sk_lookup
netfilter
...
```

```
SEC("tracepoint/skb/kfree_skb")
int bpf_prog(void *ctx)
{
        ...
        bpf_perf_event_output(ctx, &perf_buf_map,
                              BPF_F_CURRENT_CPU,
                              &data,
                              sizeof(data));
        return 0;
}

$ clang -O2 -g -target bpf -c $(NAME).bpf.c -o $(NAME).o

$ huatuo-cli bpfobj debug --event perf_buf_map --event-struct perf_data xxx.o
{"cpu":25,"pid":946799}
{"cpu":9,"pid":0}
{"cpu":9,"pid":3210394}
{"cpu":9,"pid":0}
{"cpu":9,"pid":0}
{"cpu":9,"pid":3210394}
{"cpu":9,"pid":0}
...
```

# eBPF 平台能力建设 性能保障

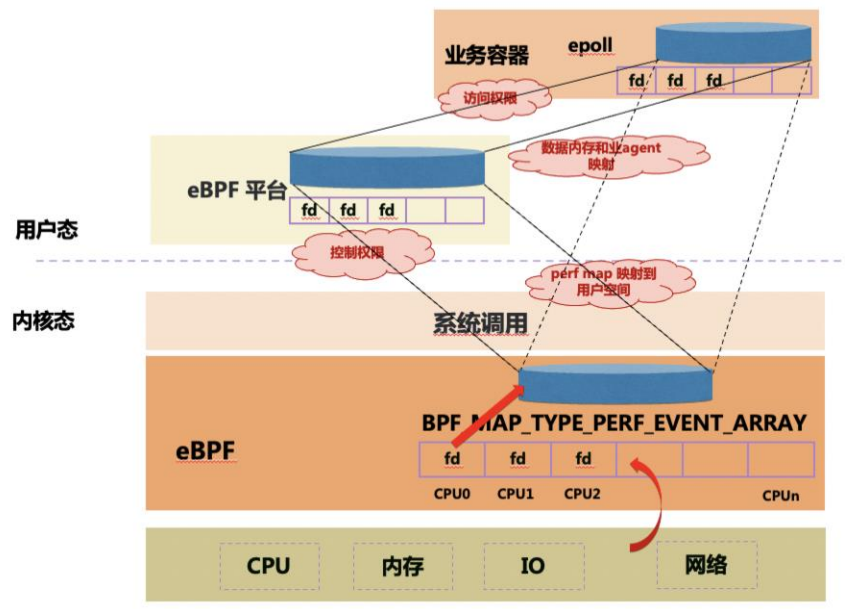
**资源消耗?**

- ❑ 支撑业务类型较多需要各维度保障性能：IO消耗，延迟敏感
- ❑ 性能保障点：
- ❖ 内核层：probe hook 点评测，熔断机制
- ❖ eBPF 平台侧：高性能数据通信，ringbuf 生产消费方式降低延迟

**举例：数据零拷贝**

第一步：创建perf event map
第二步：创建perf event fd 并填充map
第三步：eBPF 平台映射fd 内存到用户态
第四步：sendfds 到业务容器
第五步：业务容器再次映射内存

注：所有步骤只需调用一次API



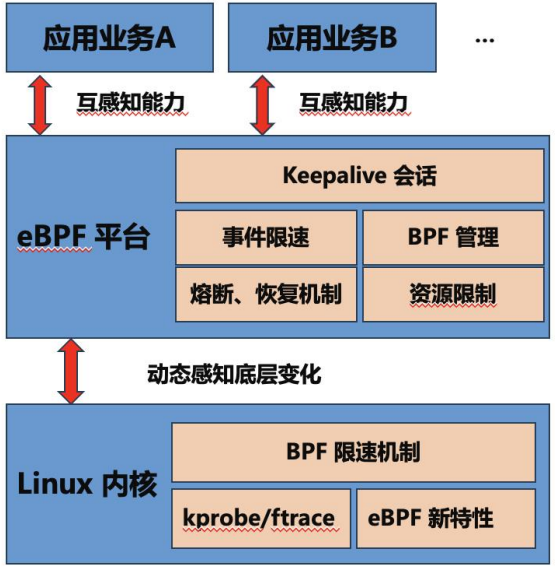
# eBPF 平台能力建设 稳定性

**稳定性保障?**

- ❑ 技术较新落地 不能影响宿主稳定性
- ❑ BPF 运行在内核态 不能影响宿主指标异常

**如何保障稳定性?**

- ❑ 内核层面：控制面，运行时
- ❑ eBPF 平台：承上启下感知能力
- ❑ 应用业务：实现感知，降级



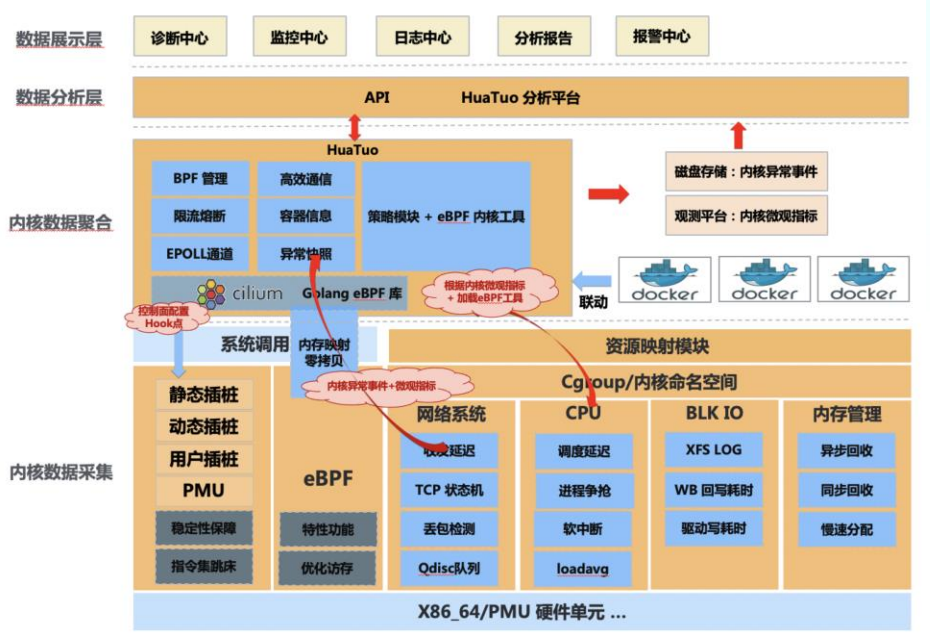
# 业务落地实践

**目前落地场景：**

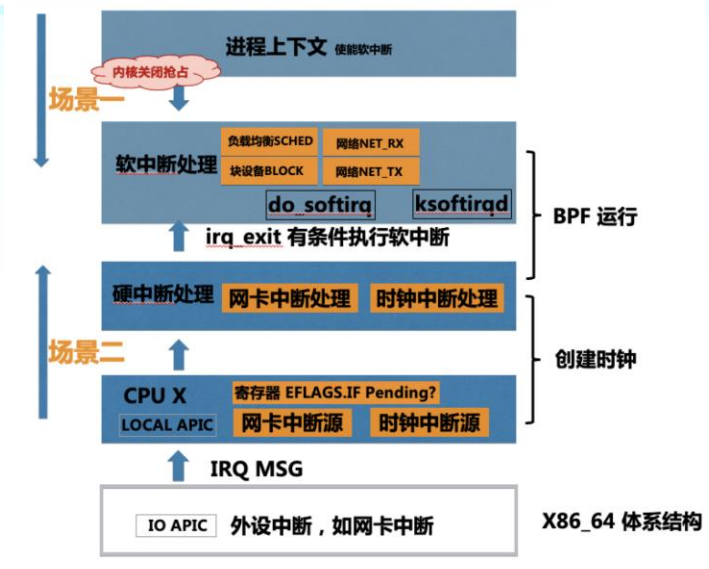
- ❑ 服务测试回归
- ❑ 容器安全
- ❑ 主机安全
- ❑ 服务访问拓扑
- ❑ 网络诊断
- ❑ 内核根因分析
- ❑ 混沌工程
- ❑ …

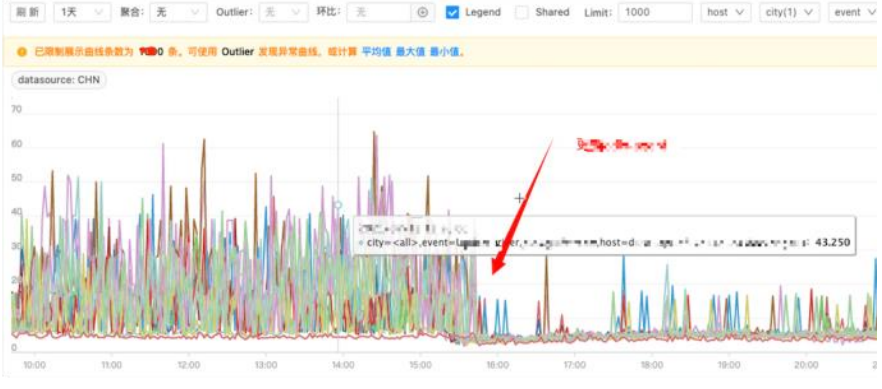
**举例：内核层根因定位**

- ❑ 容器部署密度，超卖率高
- ❑ 深度观测内核，内核微观指标
- ❑ 事件驱动，建立内核异常上下文
- ❑ 常态运行，解决突发毛刺超时等问题



# 业务落地实践




## 举例：中断检测

**背景**

- ❑ 内核通用资源存在争抢，相互影响
- ❑ 容器部署密度大，超卖严重

**影响**

- ❑ 中断，软中断长期关闭导致系统(如网络)响应延迟

**挑战**

- ❑ 中断系统为内核热路径，性能要求高
- ❑ 社区方案 vs 时钟驱动方案

# 未来规划和展望

## HuaTuo eBPF 平台开源计划

- ❑ 开源和业界合作共建： 正规划基金会孵化
- ❑ 反馈开源社区

## eBPF 内核社区

- ❑ 更加精细化BPF 性能监控：监控自身问题
- ❑ 性能优化：支撑更多常态运行需求
- ❑ 更加丰富的应用场景：如云原生混部

# THANKS


群聊: HuaTuo eBPF
该二维码 7 天内 (6月16日前) 有效，重新进入将更新


张同浩
北京 海淀
扫一扫上面的二维码图案，加我为朋友。