第三届 eBPF开发者大会

www.ebpftravel.com

操作系统性能优化：性能工具及工程化实践

国科础石操作系统部
石泉
2025/04/17

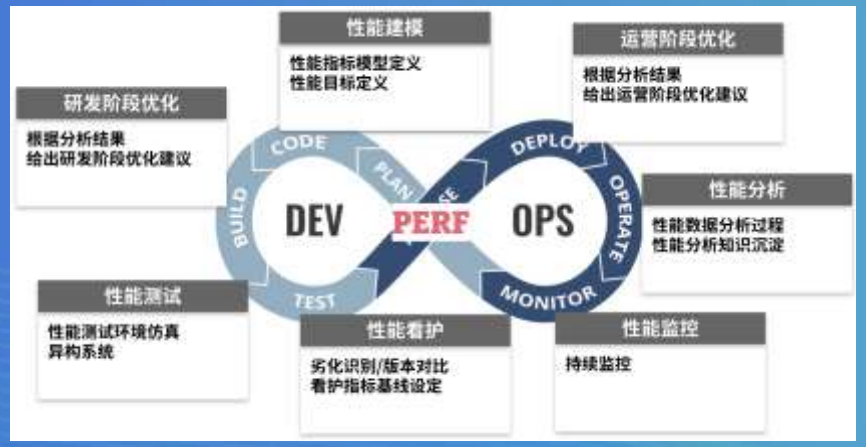
中国·西安

国科础石 KERNELSOFT

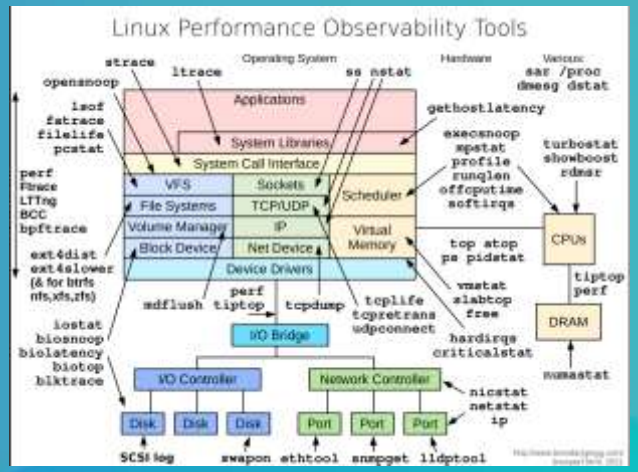
- 性能工程实践背景
- 性能优化工具链建设
- 建立软件性能观测体系
- 将性能管理融入全生命周期

系统性能工程——Have It Both

关联的多学科知识依赖



全流程的需求嵌入

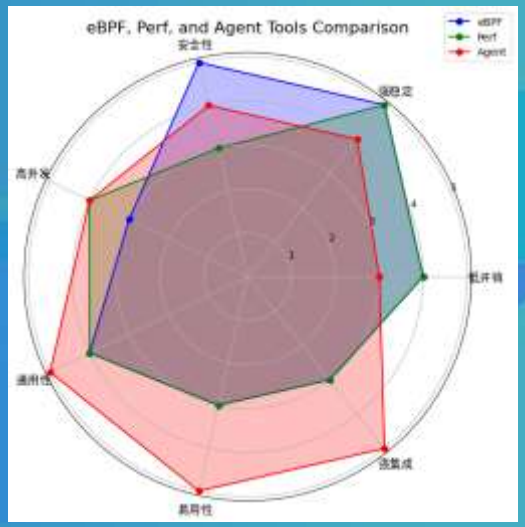


工具本身复杂性

requirement & eBPF Gap Analysis

通用能力需求	汽车领域需求	其他行业需求 以能源行业为例
系统兼容范围	低开销	环境通用
环境通用	实时性可观测	高并发
可视化	通信链路可观测	调度延迟可观测
工具集成程度	稳定性	稳定性
稳定性	可视化	实时性
低开销	工具集成程度	可视化

Expected Functionalities	eBPF Functionalities	perf Functionalities	agent Functionalities
低开销	⭐⭐	⭐⭐	⭐
强稳定	⭐⭐⭐	⭐⭐⭐	⭐⭐
安全性	⭐⭐⭐	⭐	⭐⭐
高并发	⭐	⭐⭐	⭐⭐
通用性	⭐⭐	⭐⭐	⭐⭐⭐
易用性	⭐⭐	⭐	⭐⭐⭐
强集成	⭐	⭐	⭐⭐⭐



理想/实际性能差距

			Jetson Orin			Jetson Xavier NX			Jetson AGX Xavier		
Model Arch	Inference resolution	Precision	GPU (FPS)	DLA1 (FPS)	DLA2 (FPS)	GPU (FPS)	DLA1 (FPS)	DLA2 (FPS)	GPU (FPS)	DLA1 (FPS)	DLA2 (FPS)
PeopleNet - ResNet34	960x544	INT8	330	NA	NA	79	23	23	137	29	29
TrafficCamNet - ResNet18 License Plate Detection License Plate Recognition	960x544 640x480 96x48	INT8	347	NA	NA	85	NA	NA	133	NA	NA
TrafficCamNet - ResNet18	960x544	INT8	1056	NA	NA	289	84	84	490	111	111
DashCamNet - ResNet18	960x544	INT8	1112	NA	NA	276	91	91	465	115	115
FaceDetectIR - ResNet18	384x240	INT8	1145	NA	NA	1142	444	444	1983	608	608

标称性能

电源模式
时钟同步

性能劣化

固件竞争
负载干扰

cyclictest(us)	xavier	orin
min	312	168
max	1152	1648
Multiple	3	10

实际性能

优化后的性能

			Jetson Orin			Jetson Xavier NX			Jetson AGX Xavier		
Model Arch	Inference resolution	Precision	GPU (FPS)	DLA1 (FPS)	DLA2 (FPS)	GPU (FPS)	DLA1 (FPS)	DLA2 (FPS)	GPU (FPS)	DLA1 (FPS)	DLA2 (FPS)
PeopleNet - ResNet34	960x544	INT8	330	NA	NA	79	23	23	137	29	29
TrafficCamNet - ResNet18 License Plate Detection License Plate Recognition	960x544 640x480 96x48	INT8	347	NA	NA	85	NA	NA	133	NA	NA
TrafficCamNet - ResNet18	960x544	INT8	1056	NA	NA	289	84	84	490	111	111
DashCamNet - ResNet18	960x544	INT8	1112	NA	NA	276	91	91	465	115	115
FaceDetectIR - ResNet18	384x240	INT8	1145	NA	NA	1142	444	444	1983	608	608

标称性能

电源模式
时钟同步

性能劣化

固件竞争
负载干扰

cyclictest(us)	xavier	orin
min	312	168
max	1152	1648
Multiple	3	10

特定版本的性能

性能优化

cyclictest(us)	xavier	orin
min	110	165
max	353	529
Multiple	3	3

优化后的性能

更进一步的问题

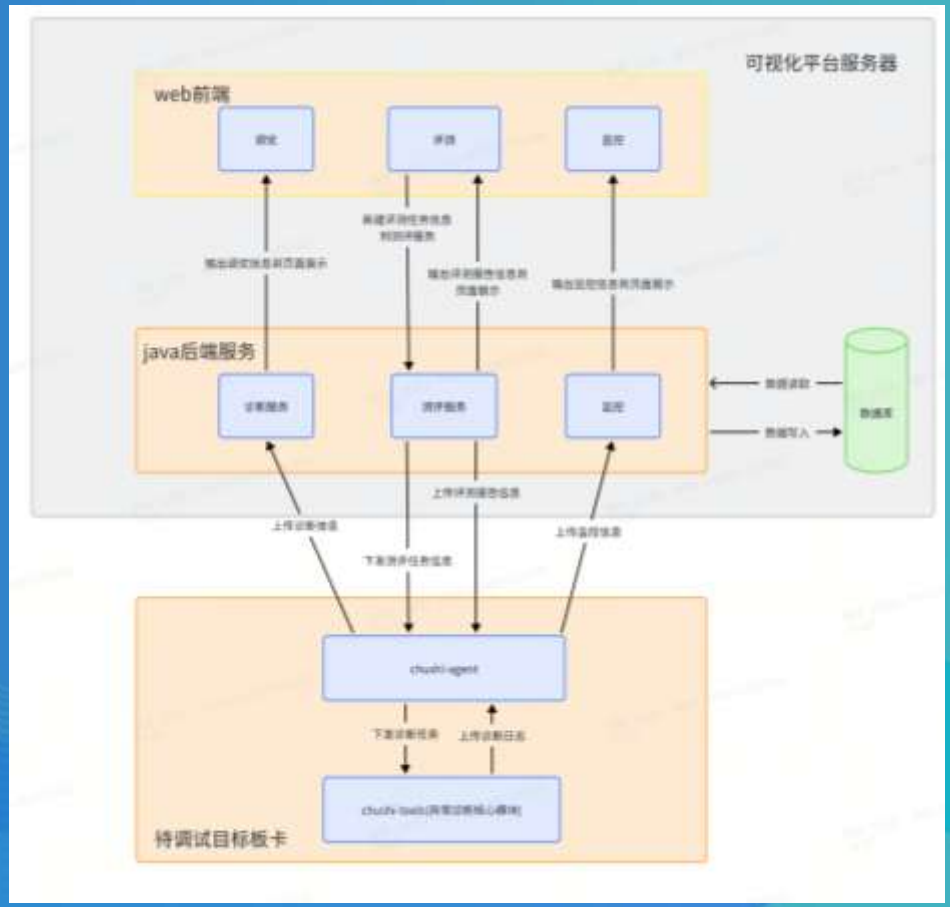
操作系统性能 != 业务实际性能

内核稳定性与业务稳定性之间的关联

如何提升操作系统可靠性

- 性能工程实践背景
- 性能优化工具链建设
- 建立软件性能观测体系
- 将性能管理融入全生命周期

chushi-tools工具链构建



Module+ebpf+agent+web

Have It Both

ebpf/Module超能力 ——> 满足工程需求

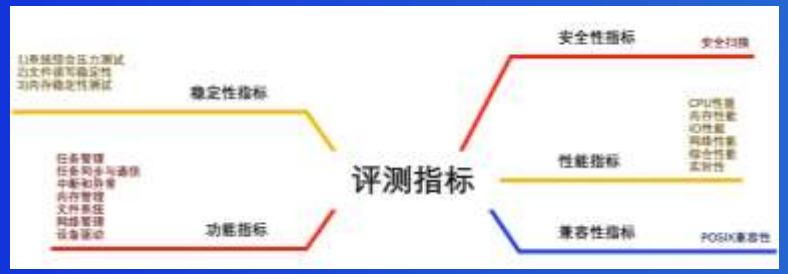
Module能够保证安全的情况下，使用Module实现

Module不能保证安全的情况下，通过ebpf实现

向Module要性能 向ebpf要安全

国科础石 KERNELSOFT

Chushi Metric Customization



评测指标定制

实时性指标定制

```
static int flexcan_read_frame(struct net_device *dev)
{
    struct net_device_stats *stats = &dev->stats;
    struct can_frame *cf;
    struct sk_buff *skb;

    skb = alloc_can_skb(dev, &cf);
    if (unlikely(!skb)) {
        stats->rx_dropped++;
        return 0;
    }
```

can指标定制

More Customization

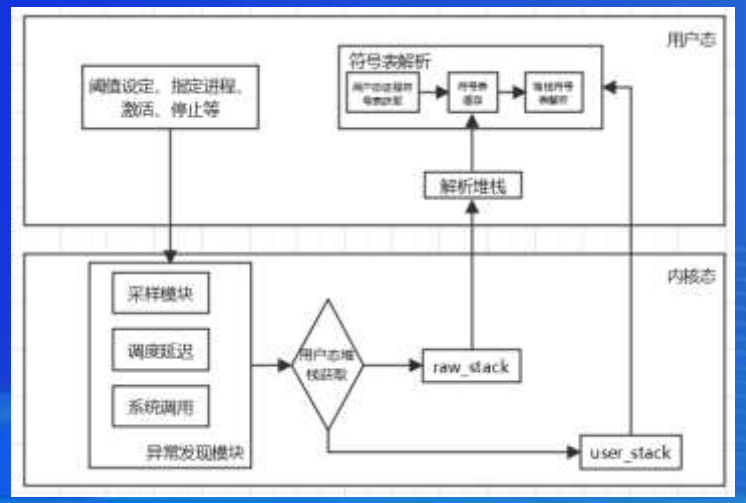
模块定制　　独立可用

兼容性定制　　更大跨度的版本适配 qnx、rtos适配

国科础石 KERNELSOFT

usability

自动化——堆栈解析到自动化评测



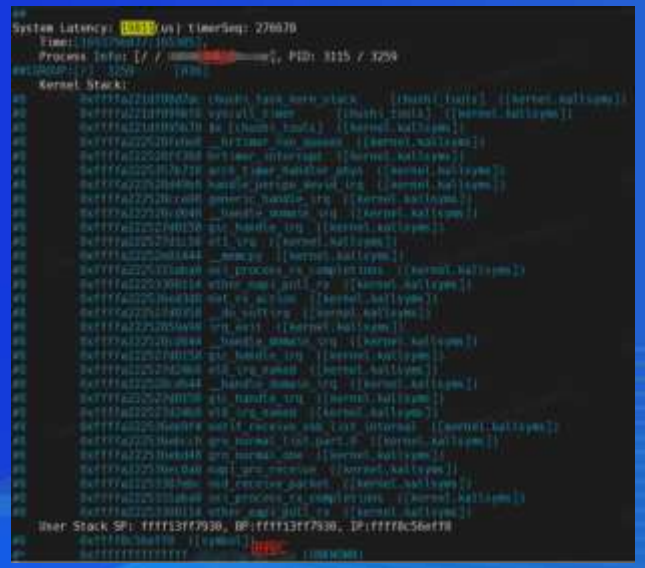
usability

可视化——CLI到web

```
→ ~ chushi-tools  delay-trace --help
  delay-trace Instructions:
      --help  Print the help information
      --config-set   start and config
          level  Detailad information
          delay  Set threshold(MS)
          delay-us  Set threshold(US)
          buf-size-k  Set the cache size for each process information(K)
          timer-us  Sampling Timer Period(us)
          raw-stack  Whether to parse stack frames in user mode
      --config-clear  stop and config clear
      --config-dump  dump config info
      --dump  dump log information
      --test  Function test.
      --set-syscall PID SYSCALL THRESHOLD   Monitoring system calls for specific processes
      --clear-syscall PID   Don't monitor system calls for specified processes
      --flame   Generate Flame Graph
              in  Specify the file name to save after the raw stack is parsed
              inlist  Specify the file name of the file list containing raw-stack to parse
              console  Read the list of files that need to generate a flame diagram from the console
      --uprobe  Set upprobe start and stop tracking points
          tgid  Process to be monitored
          Start-file File the start-probe will inject
          End-file File the end-probe will inject
          Start-offset Offset in the start file
          End-offset Offset in the end file
```

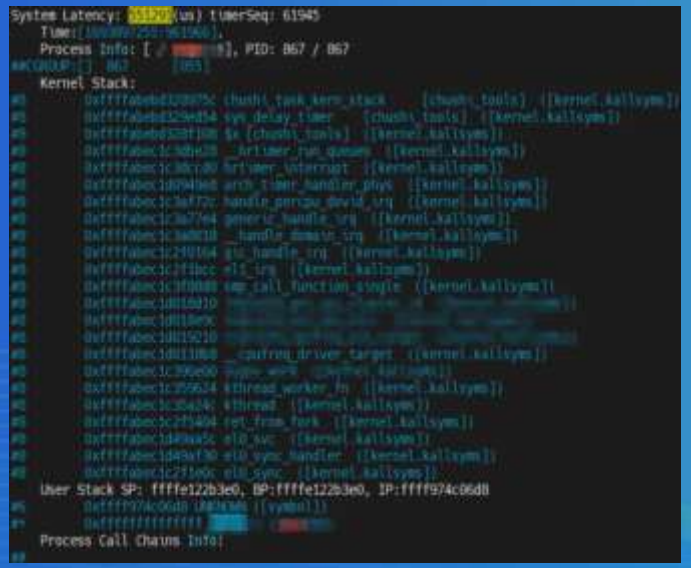
性能问题优化实践

级联失效问题



cpu使用率飙高，应用业务异常
网卡收包软中断中数据拷贝流程
存在19ms的延迟

调频策略问题



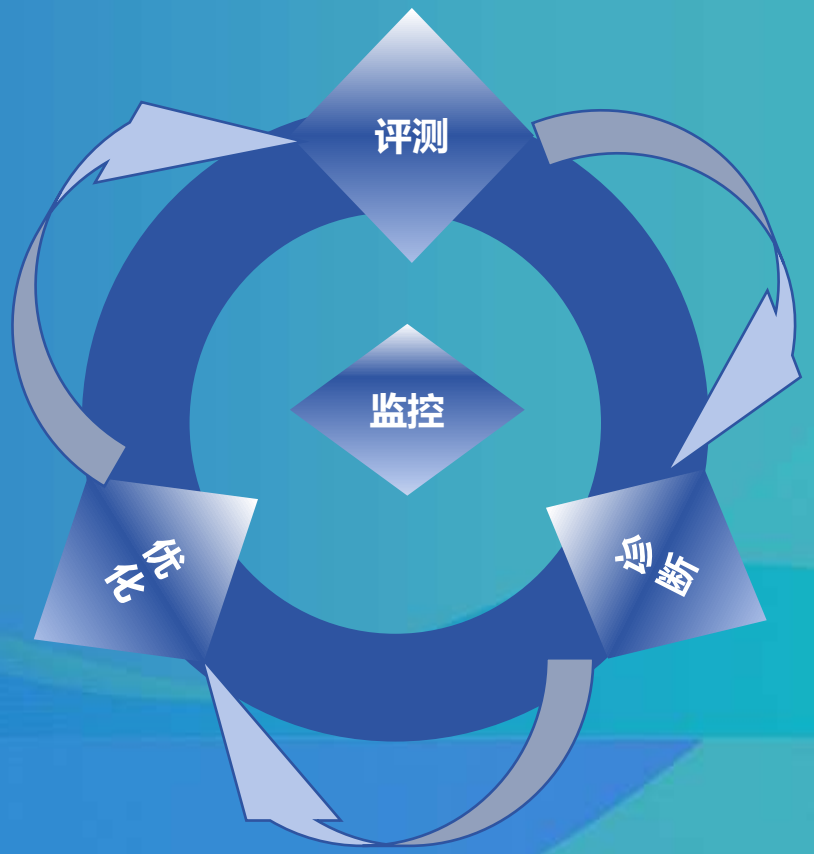
电源调频——导致500ms以上的延迟

间歇性问题

通信服务概率性不响应——异常中断

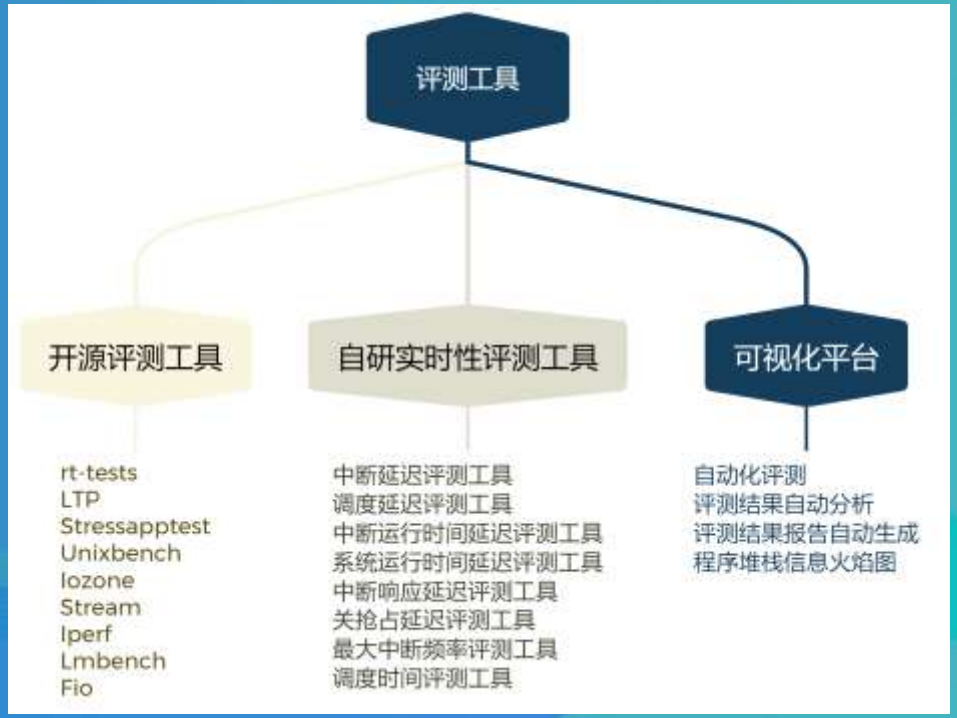
- 性能工程实践背景
- 性能优化工具链建设
- 建立软件性能观测体系
- 将性能管理融入全生命周期

础石自动化性能监控平台

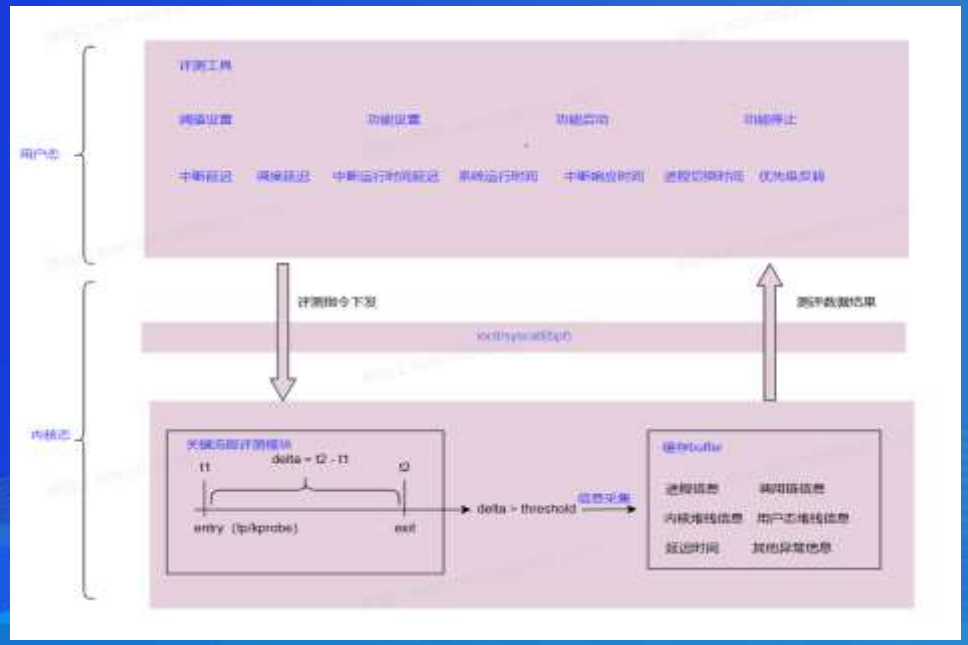


础石自动化评测平台

系统与软件状态模型					
实时性	性能	功能性	安全性	兼容性	…
任务调度	cpu&mem	进程管理	rootkit	信号	…
优先级反转	线程	进程同步	异常主机	定时器	
进程间通讯	io	中断	启动项	消息队列	
线程互斥	访存	定时器	端口扫描	互斥锁	
	iozone	…	…	读写锁	



础石自动化评测平台


评测工具
用户态
评测指令下发
测评数据结果
内核态
delta = t2 - t1
t1
t2
entry (tp/kprobe)
exit
delta > threshold
信息采集


实时性测评、稳定性测评、性能测评...
前端
系统配置、管理
首页
报告管理
报告详情
API
java 后端
controller
service
dao
组件
mysql
redis
minio
gataway
jenkens
测评agent
https/socket
稳定性
实时性
性能
可靠性

础石系统调优诊断工具

定位

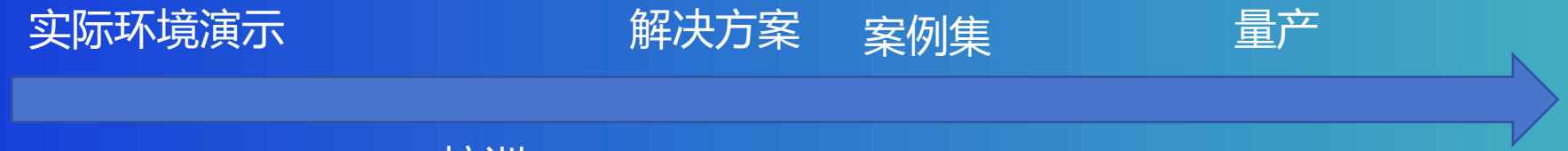
原因分析

优化

问题定位，补丁修复
实时性/稳定性增强
安全增强
漏洞修复
调度负载均衡

- 性能工程实践背景
- 性能优化工具链建设
- 建立软件性能观测体系
- 将性能管理融入全生命周期

尊重工程



制造压力边界，构造极限场景暴露问题
定位具体函数，优化不脱离实际
调整指标涵义，业务需求优先
性能测试+基准测试+诊断优化，形成实践闭环

保留对工具的疑问
重视每一个环节，哪怕再小
记录偶然事件，保留 trace 痕迹
开展培训，提升性能意识

明确性能需求，定位实际价值
形成性能回归体系

感　谢
THANK　YOU