第三届 eBPF开发者大会
www.ebpftravel.com

基于eBPF的全系统PGO优化方案

任玉鑫，华为， openEuler Valuable Professional

中国·西安

PGO背景

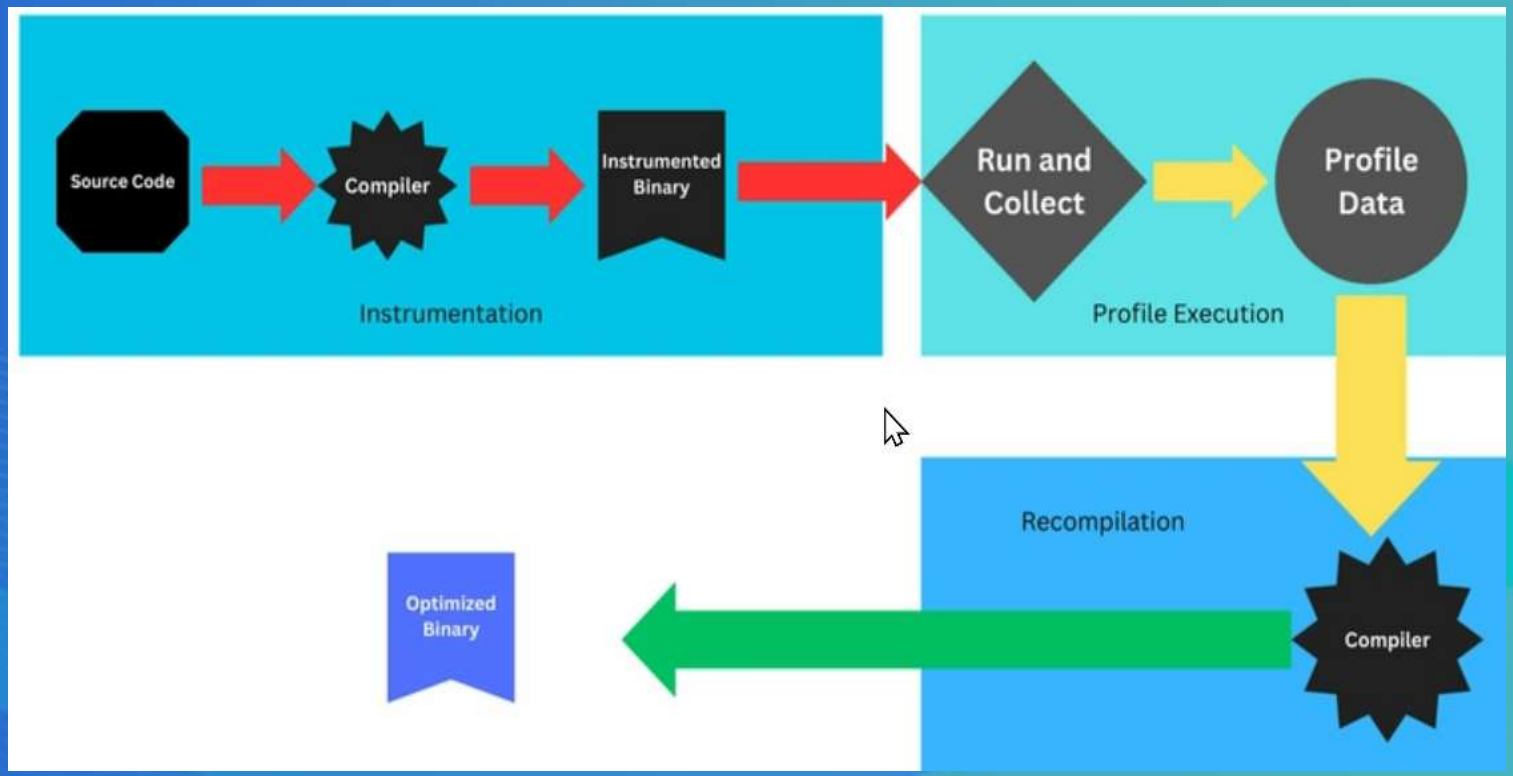
- PGO：Profile Guided Optimization 有效的编译优化手段

- 核心思想
  - 数据越多，优化效果越好
  - 数据约体现应用特性，优化针对性越强

PGO背景

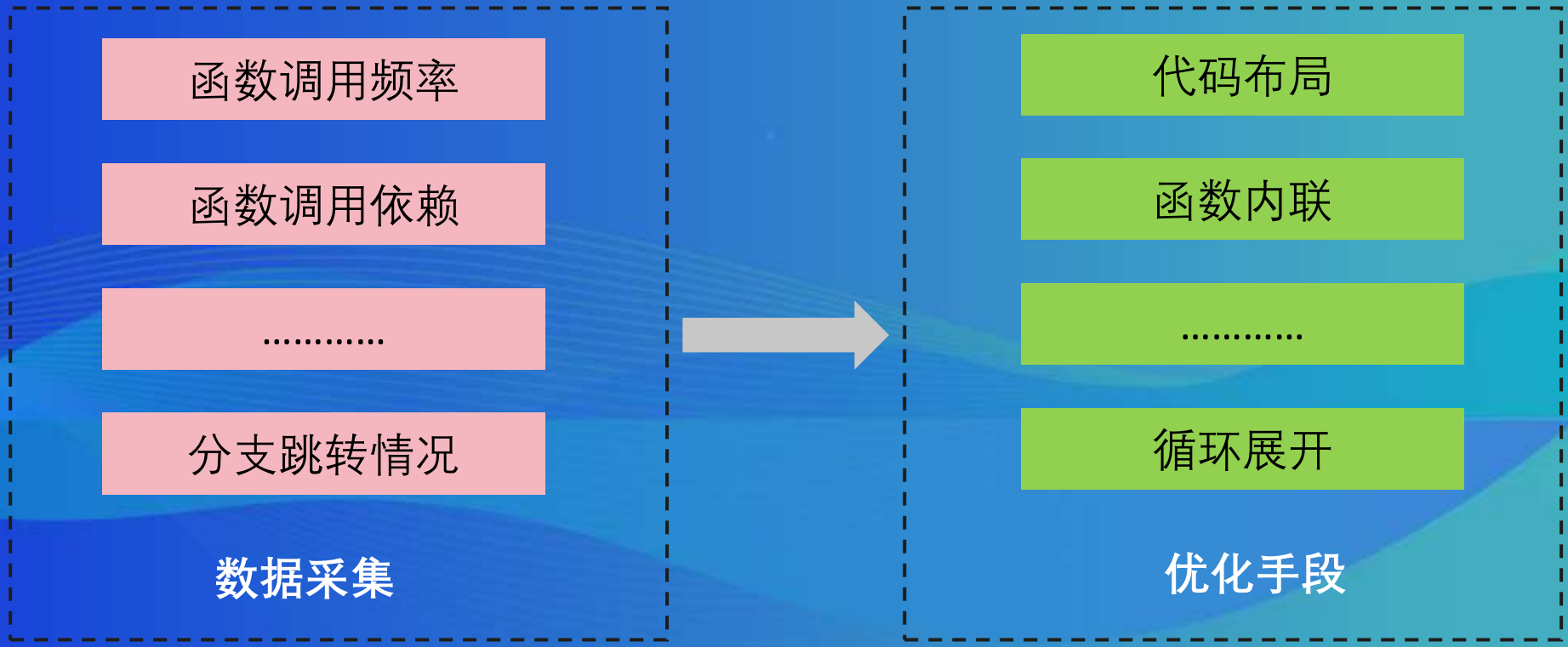
工作流程

- 编译器在源代码插桩
- 运行插桩后的应用，并采集数据
- 根据采集到的数据，尝试相应优化技术
- 不断迭代重复，直到优化效果达标



PGO背景

采集与优化的例子



PGO背景

- **为什么需要PGO**
  - 编译阶段没有足够的应用运行时信息
  - 很多优化不通用，需要在具体应用上尝试
- **PGO优势**
  - 极致性能
  - 确定性优化，针对应用定制
  - 自动化，避免手动尝试各种配置
- **PGO劣势**
  - 优化只针对单个应用环境，每次部署都需要重新优化
  - 数据采集不全面，无法覆盖所有代码路径和程序输入
  - 只能采集优化单个二进制，无法优化动态库

PGO背景

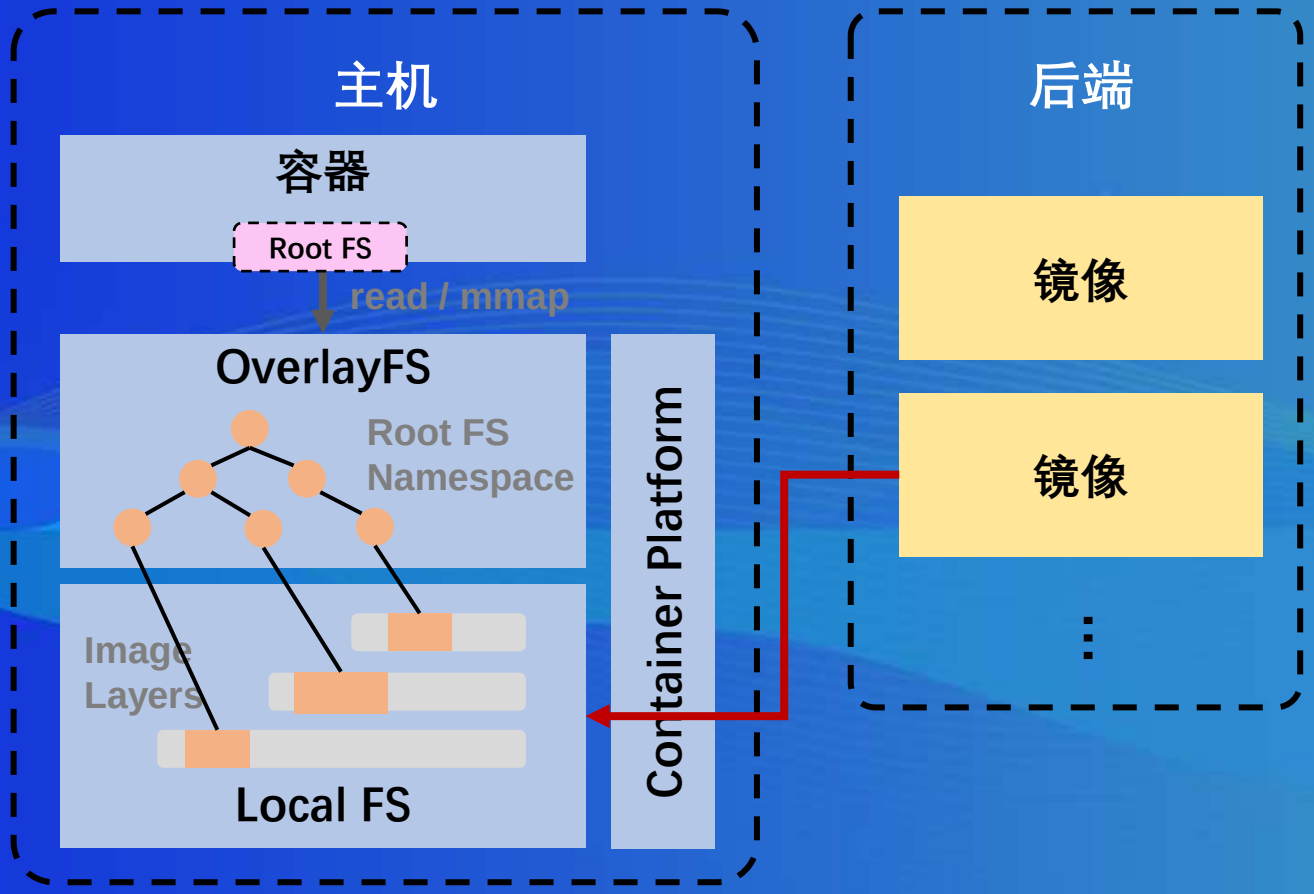
- **无法采集优化应用间、应用与系统之间的交互，难以实现全系统的性能最优**

- 系统性能
  - 交互
  - 中断
  - 系统调用
  - 通信

- **仅仅依靠编译器无法采集优化**

容器IO启动

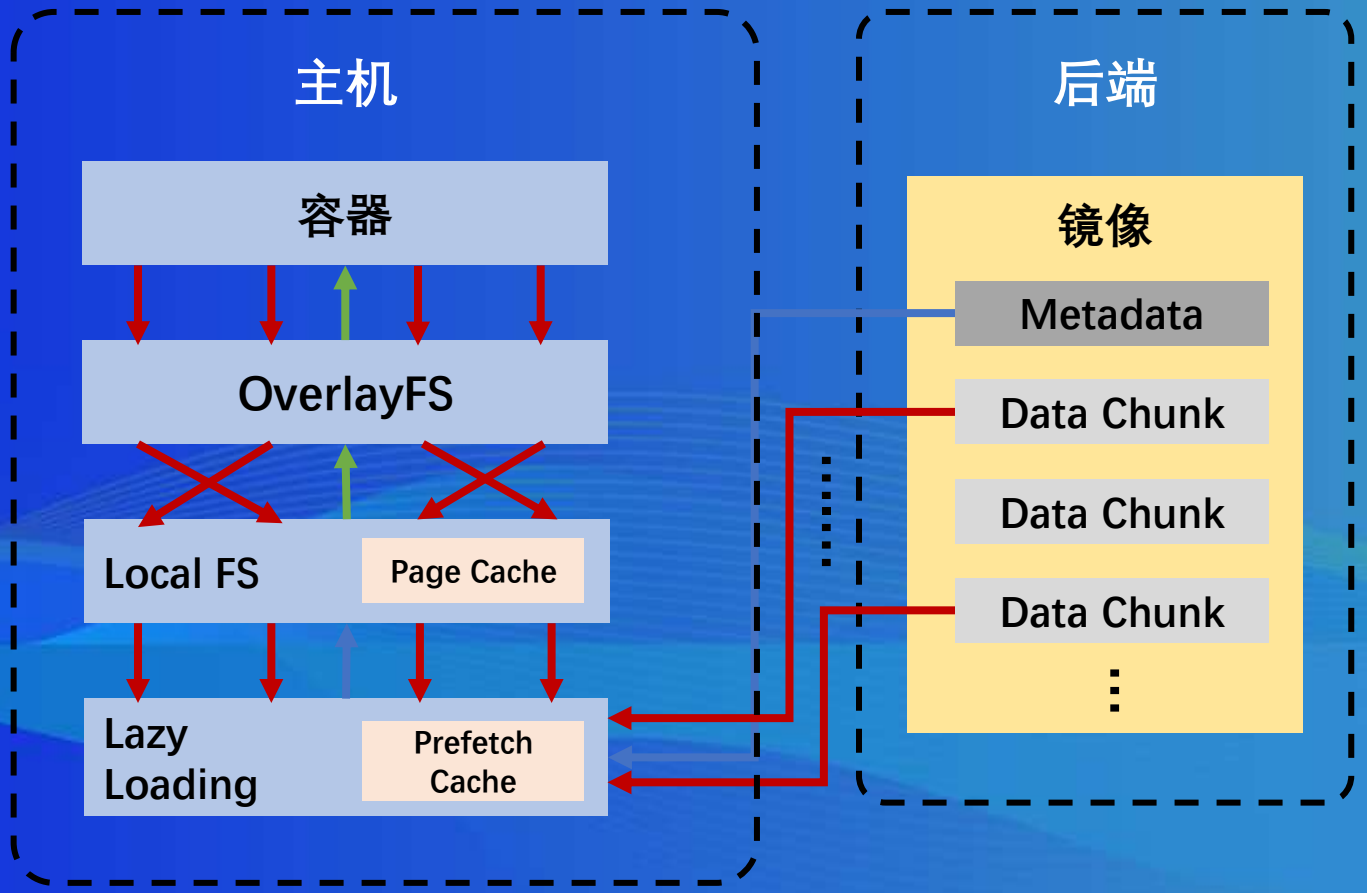
- 容器是云原生计算的基础设施，通过轻量、弹性、模块化等实现高资源利用率和高性能。



容器启动是重要性能指标

- 现有方案是全量加载，但很多数据不需要，造成性能浪费
- AI时代，问题尤为凸显。Pytorch启动达分钟级！

容器IO启动流程



阶段 1: Deploy (→)

获取镜像元数据

阶段 2: Running (→)

创建容器运行时，例如*cgroup*

阶段 3: Ready (→)

启动容器内服务应用

容器IO启动问题

现有方案	时延分解				I/O行为	
	Deploy	*Running*	*Ready*	*Total*	*I/O Amp.*	*Net. Pkg.*
Full Image	124.6s	1.6s	1.7s	127.9s	47.5X	573K
CRFS	1.8s	1.2s	24.1s	27.1s	1.8X	99K
Nydus	0.8s	2.9s	21.4s	25.1s	1.6X	90K
DADI	0.6s	2.6s	17.0s	20.2s	3.1X	171K
DADI-Trace	0.7s	2.2s	17.1s	20.0s	3.0X	166K

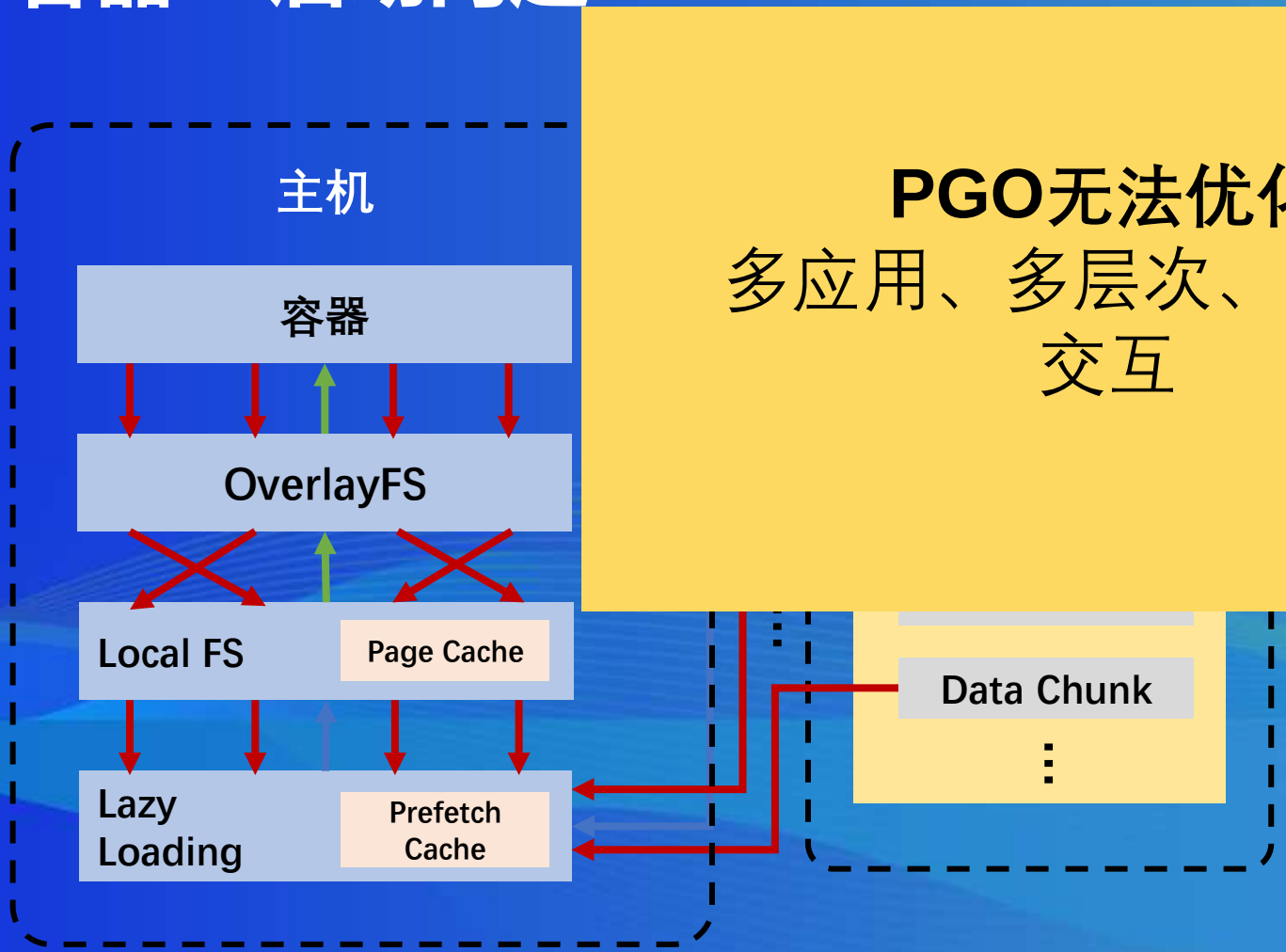
性能瓶颈

- IO的数据量和次数
- 操作系统IO处理和page cahe性能低效

现有方案不足

- 懒加载虽然加速了“*Deploy*” 阶段，但是在“*Ready*” 阶段引入大量开销
- IO放大和频繁网络IO带来不必要的性能开销

容器IO启动问题



PGO无法优化

多应用、多层次、多节点交互

eploy (→)

元数据

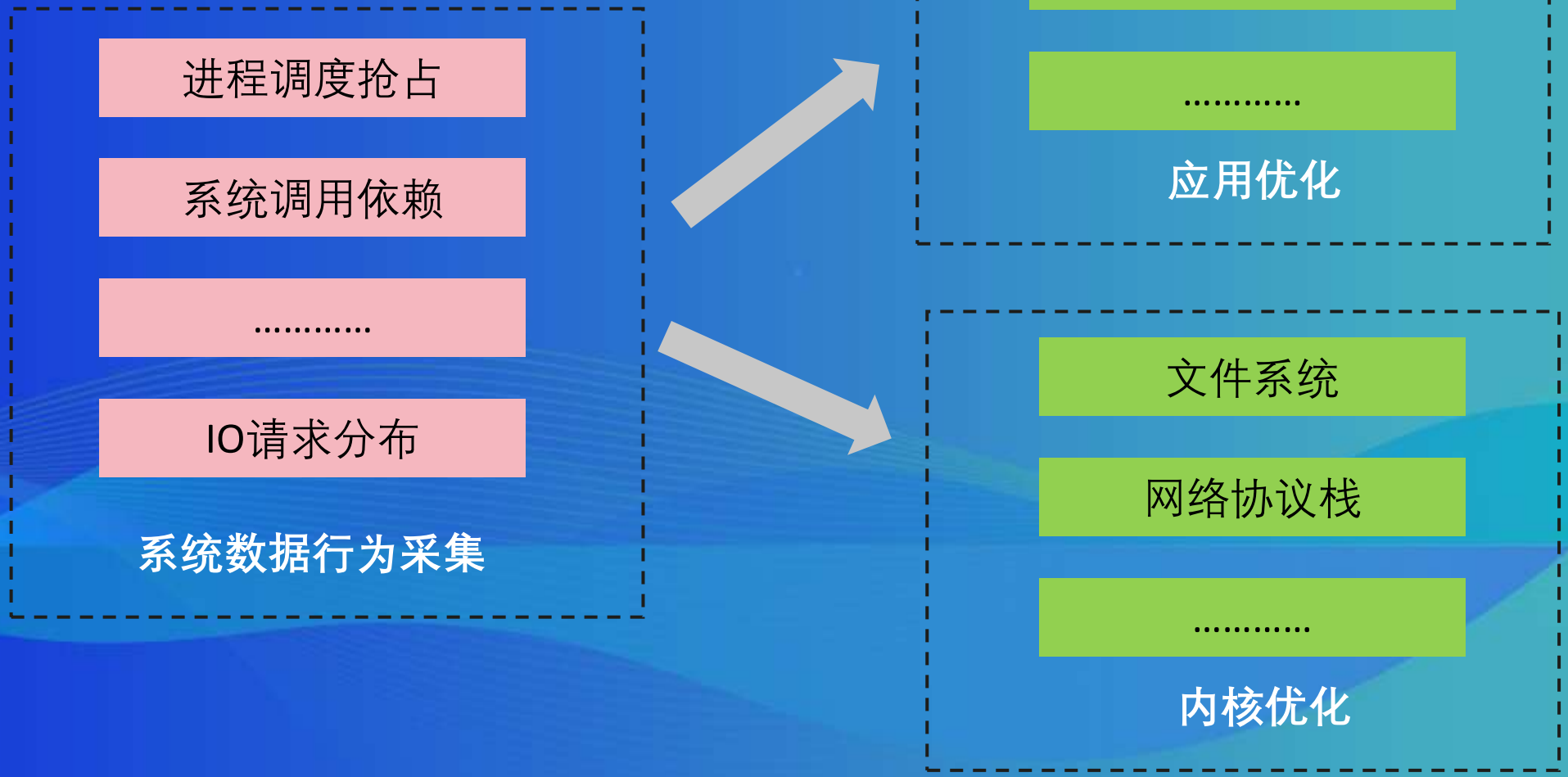
unning (→)

运行时，例如*cgroup*

阶段 3: Ready (→)

启动容器内服务应用

全系统级PGO方案


进程调度抢占
系统调用依赖
.....
IO请求分布
系统数据行为采集
内存布局
多核并行
.....
应用优化
文件系统
网络协议栈
.....
内核优化

基于eBPF全系统级PGO方案

系统数据行为采集

内存布局

多核并行

.....

应用优化

文件系统

网络协议栈

.....

内核优化

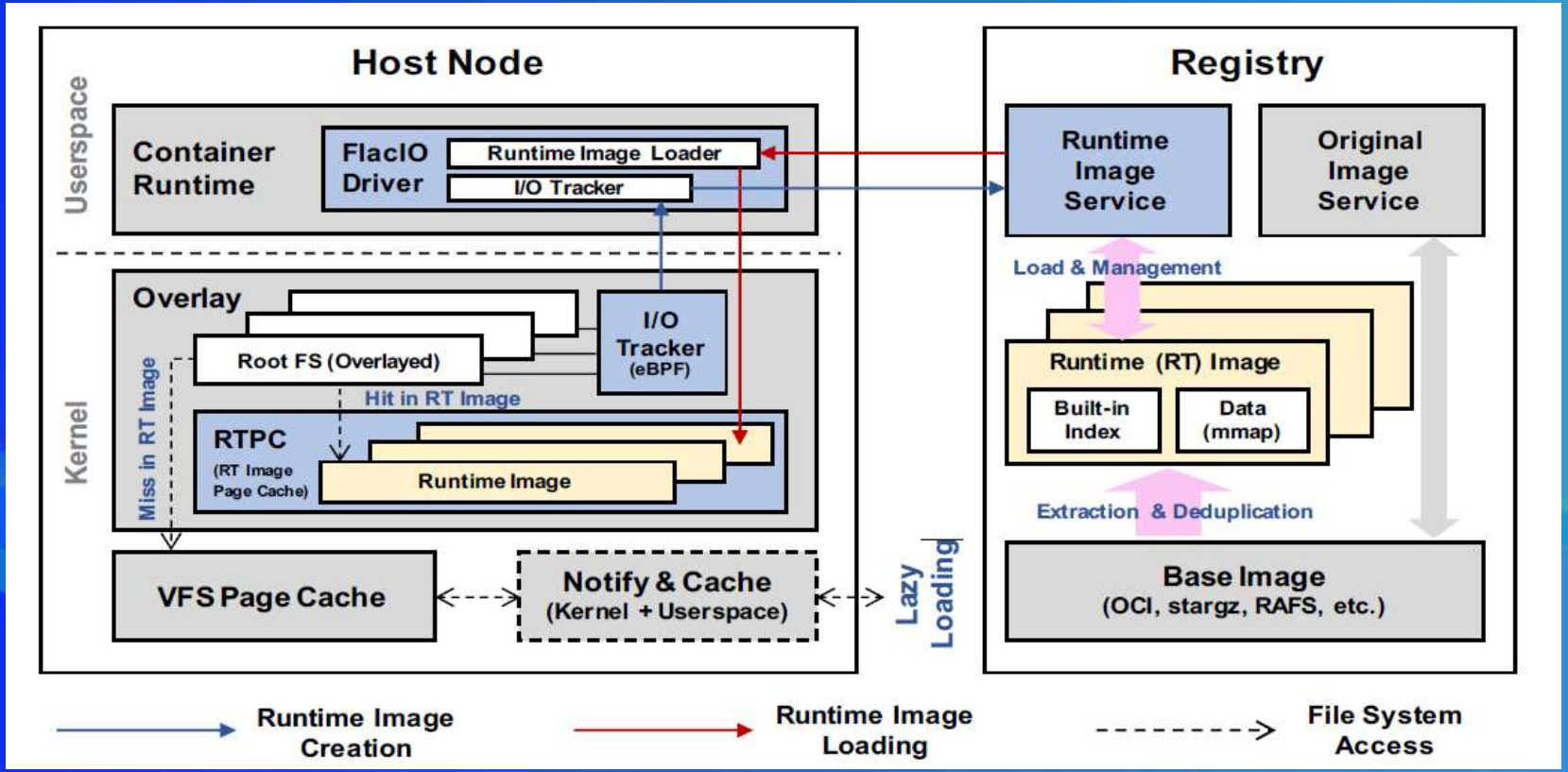
基于eBPF全系统级PGO方案

打点观测  数据分析

短路/下沉  指导重构

实践：容器IO启动加速

FlacIO (FLAt and Collective I/O)

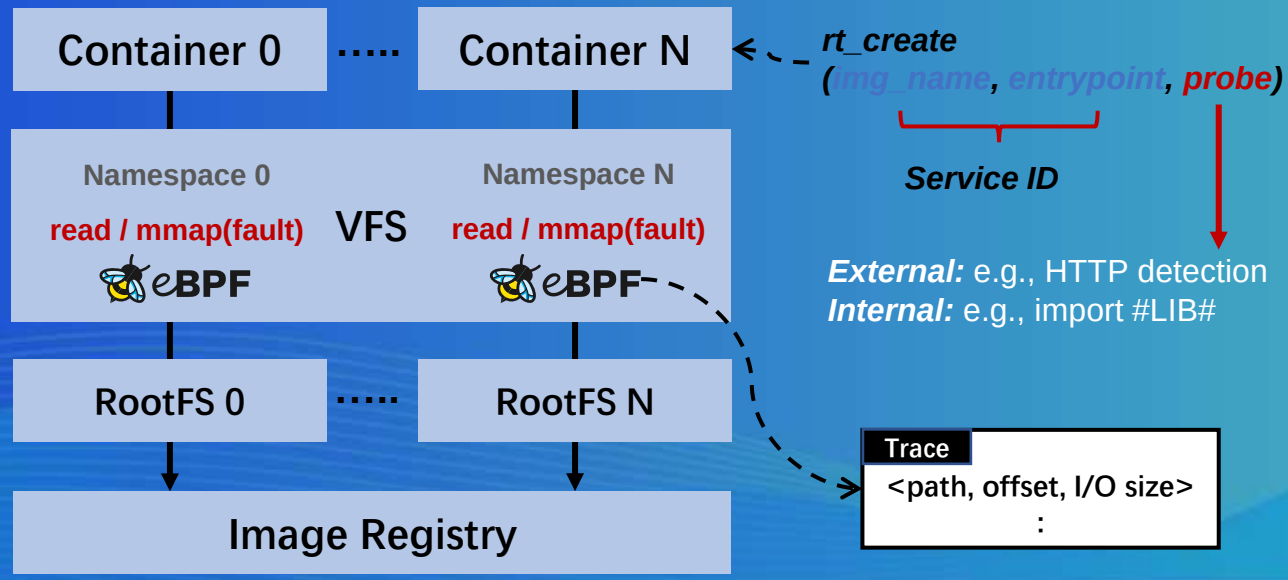


多端IO协同加速

- 细粒度IO行为数据采集
- IO数据聚集的新型镜像格式
- 增强page cache，实现镜像数据直通
- 增量加载，去除多容器间冗余数据加载

容器IO启动加速：采集

Probe-based File-Level I/O Tracing

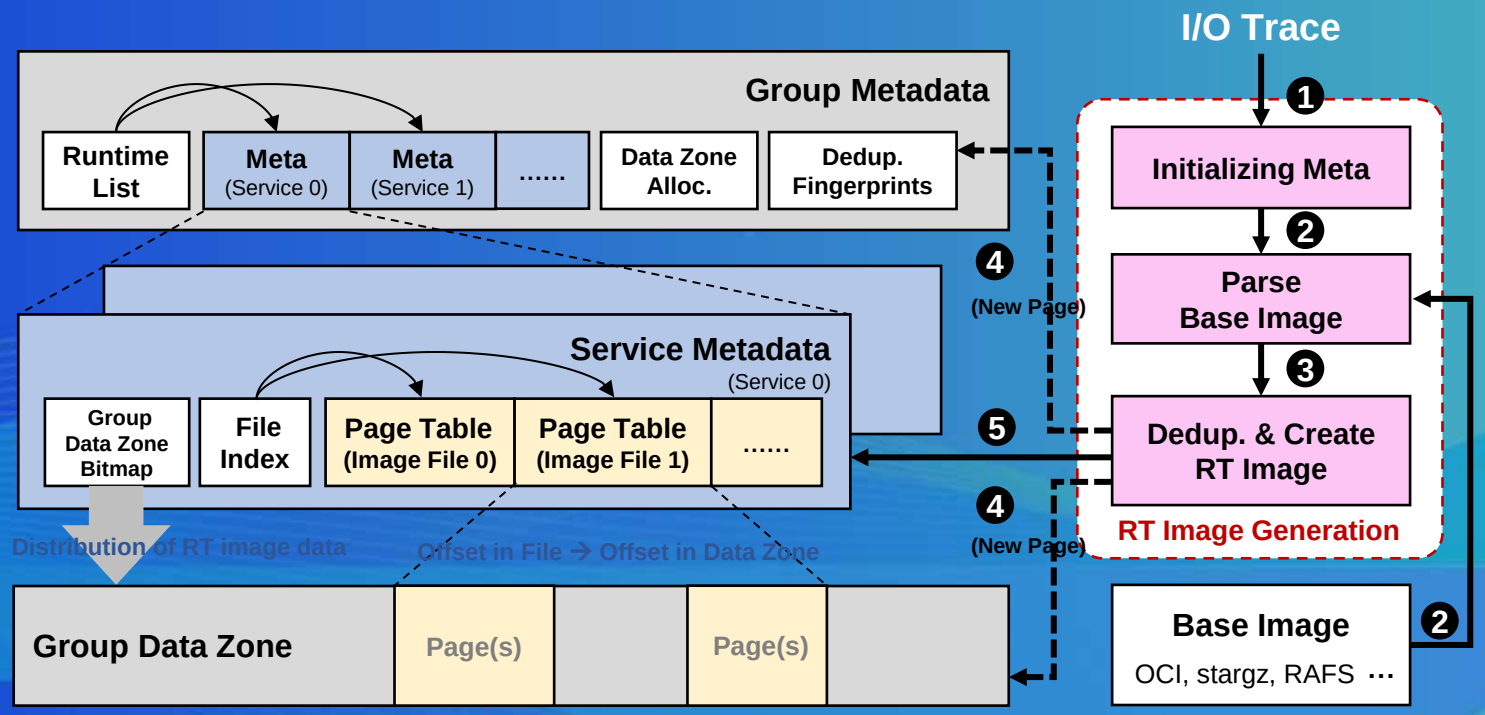


Advantages

- Stop tracing when the probe captures the corresponding event
- Suitable for any lazy loading system

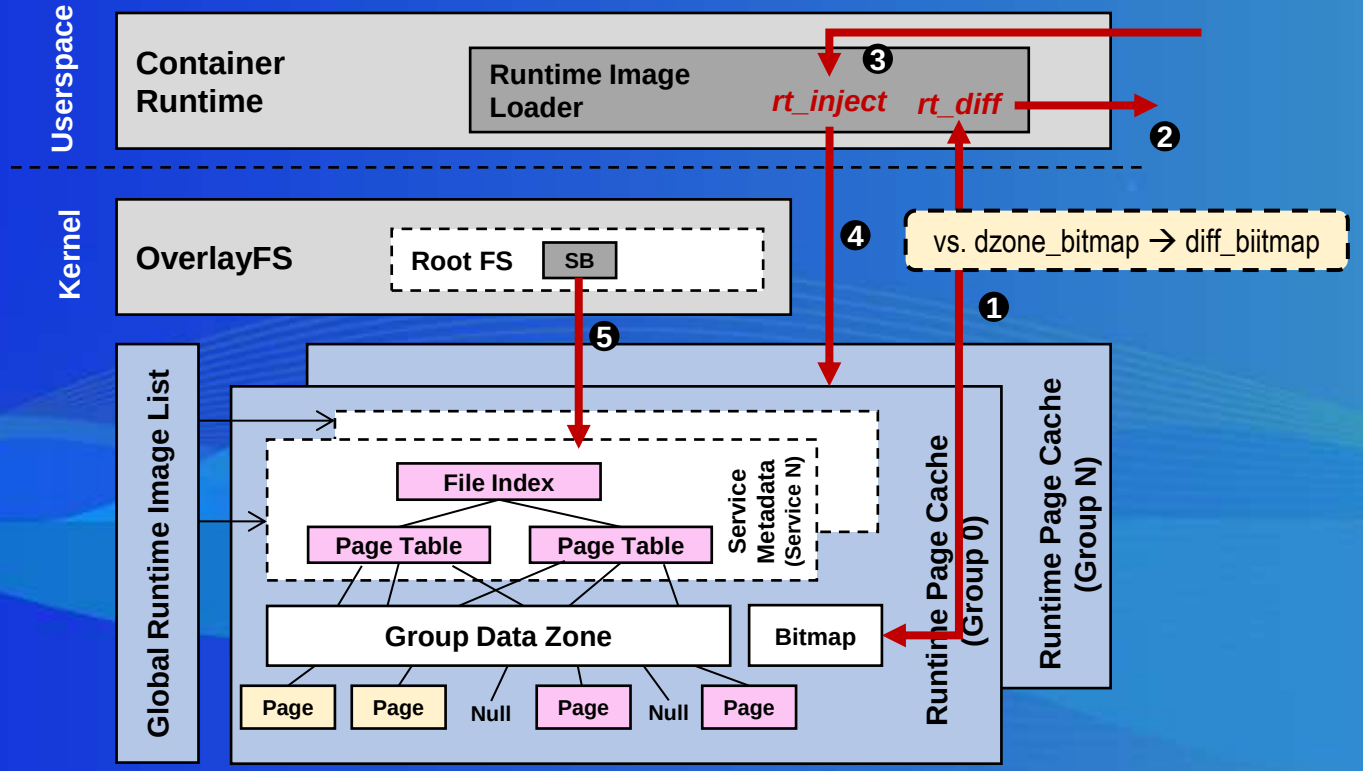
容器IO启动加速：分析

RT Image Generation Process (offline)



容器IO启动加速：重构

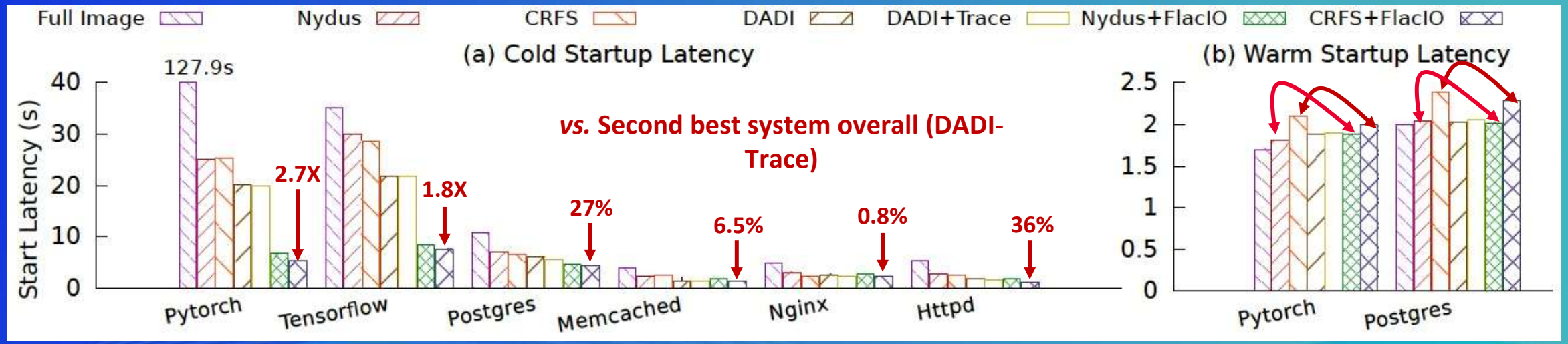
Runtime Page Cache



新的page cache框架，从RT image 构建容器RootFS

支持增量加载和去重

性能效果

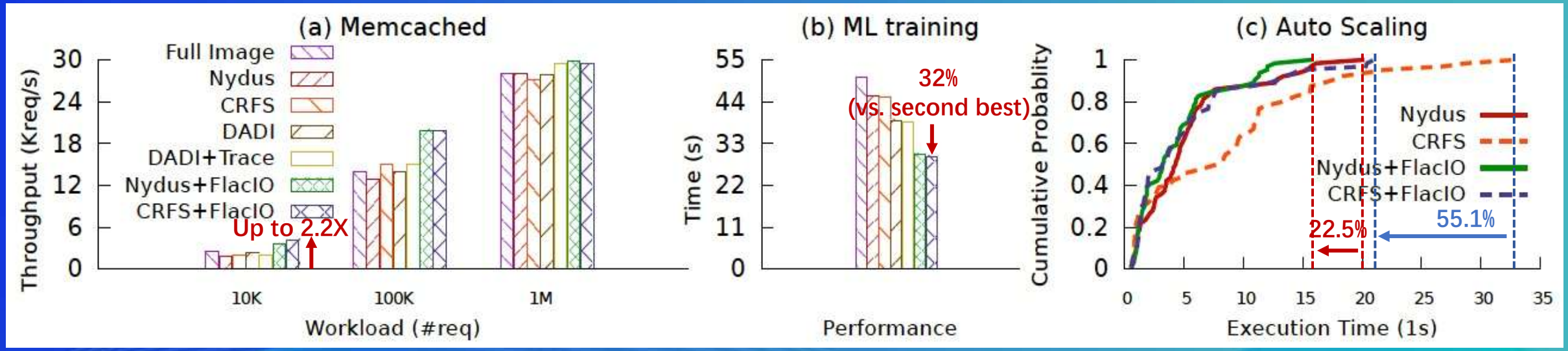


冷启动性能提升2.7倍

FAST'25

FlacIO: Flat and Collective I/O for Container Image Service

性能效果



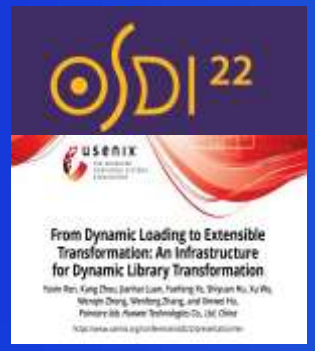
KV store吞吐性能提升2.2倍

机器学习训练性能加速32%

弹性扩容效率提升55%

openEuler：面向数字基础设施开源操作系统和社区

研究驱动、聚焦创新



跨层级垂直整合优化

中间件	Database、K8S、vLLM、Web
框架	ROS、Spark、DPDK、MindSpore
系统服务	container、systemd、sysdig、 windowing
语言	C、Rust、Java、GO、JS
基础库	Libc、OpenSSL、Math、QT
内核	Linux、RTOS、Microkernel、Hypervisor

new OS distribution → open source community → open innovation platform