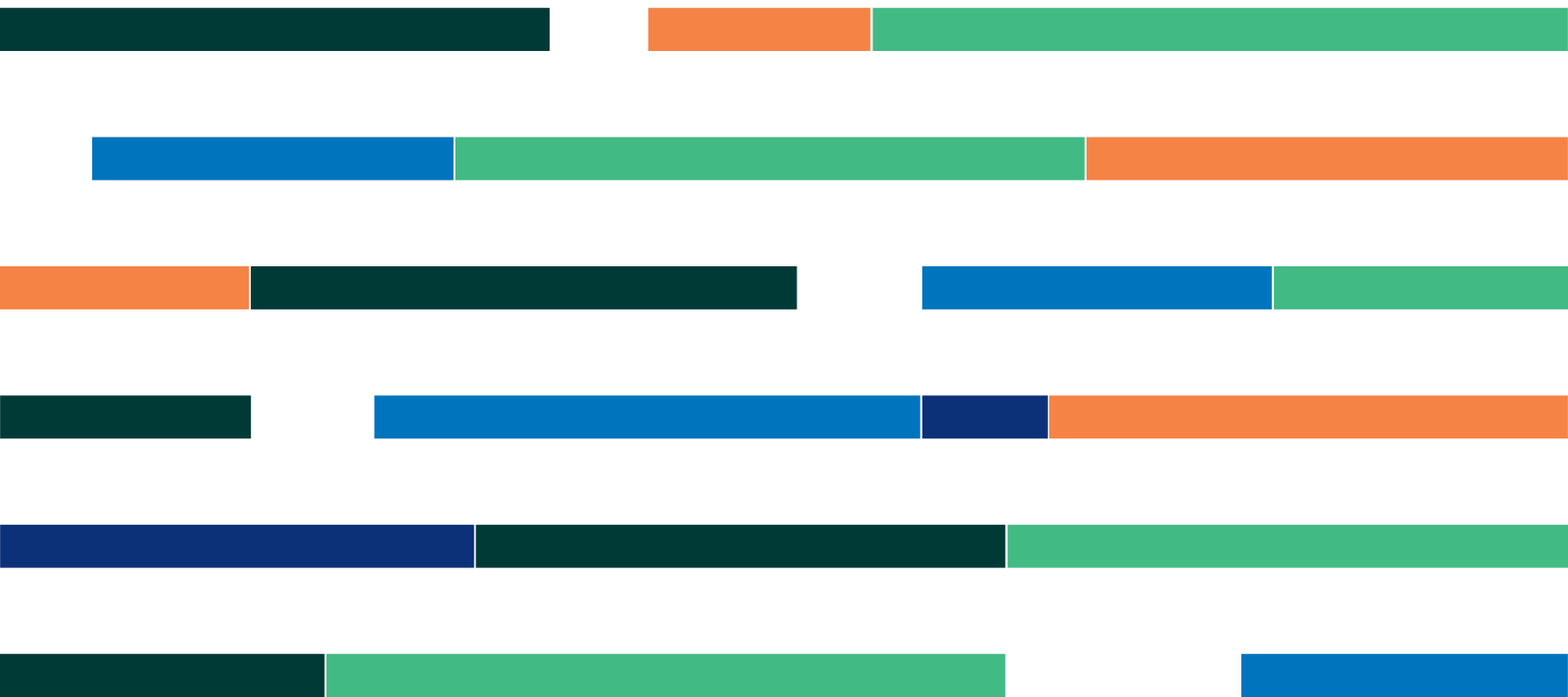
SUSE

白皮书

通过开放灵活的基础设施平台实现面向未来的电信现代化



SUSE 构建了一个适应性强的边缘计算基础设施平台，可支持现在和未来的电信应用程序

SUSE 自适应电信基础设施平台（Adaptive Telco Infrastructure Platform，ATIP）是一款可大规模扩展的管理解决方案，适用于电信级云原生裸机和云边缘基础设施。它支持零接触载入安全设备以及操作系统、Kubernetes 集群和应用程序的自动引导和生命周期管理。SUSE ATIP 融汇了 SUSE 多年来在各行各业深耕的边缘客户经验，并在系统架构方面添加了电信级功能和性能。

边缘的主要挑战：大规模管理

在典型的大型数据中心中，要管理的 Kubernetes 集群数量通常是几十个或最多数百个，而在边缘，集群的数量经常是上千、上万甚至上百万的，而且这是各行各业（包括电信行业）的普遍趋势。为此，用户需要从使用图形 UI “手动”管理集群转向使用 API 的编程方法来进行管理。

SUSE 的 Rancher Prime 提供了市场领先的 Kubernetes 管理解决方案，它以简单、可靠和出色的用户体验著称。

为了应对在边缘进行大规模管理的挑战，我们对 Rancher 进行了两项重要的补充。

1) **提供商中立的编程化边缘集群生命周期管理。**我们集成了 CNCF 的 Cluster API，为整个边缘堆栈（包括裸机和私有云基础设施、操作系统、Kubernetes 集群和应用程序）的编程管理提供了一个提供商中立的集成点。Rancher UI 的扩展让你轻松处理边缘基础设施，而且提供了直观的可视化功能，助你了解运营状态、资源利用率和库存。此外，整个边缘基础设施都以 Kubernetes 相同的声明方式进行管理。所有管理操作都是通过在中心管理集群上创建和操作 Kubernetes 自定义资源来完成的。
2) **基于 GitOps 的基础设施管理。**边缘的规模要求用户使用更有效和更严格的方法来管理所需的基础设施状态。而 GitOps 就成为了这个问题的解决方案。Rancher 的 GitOps 组件 Fleet 结合了 Cluster API，支持使用基于 GitOps 的工作流来管理边缘基础设施。

所需的状态（例如，使用哪个操作系统和 Kubernetes 版本，边缘集群要包含的 Control Plane 和 Worker 数量，要部署的 Ingress Controller 和其他的 Kubernetes 组件）都被定义为 YAML 或 Helm Chart 并存储在 Git 仓库中。此外，SUSE ATIP 也可以与 Flux 或 ArgoCD 等第三方 GitOps 工具结合使用。

通过 Git 使用完善的版本管理工作流，你可以可靠地管理状态定义。

- 可以在合并之前对更改进行同级审核
- 可以通过易于理解的 CI/CD 原则和工具链来实施自动 linting 和合规性检查
- 默认可以使用完整的更改历史
- 回滚容易，而且 100% 可预测

GitOps 引擎会负责将 Git 仓库的状态定义同步到管理集群，从而将它们转化为所需状态的有效声明。然后，SUSE Edge 管理堆栈会根据声明的所需状态协调边缘基础设施的实际状态，例如，将操作系统和 Kubernetes 集群部署到新加入的边缘站点中的裸机主机，更改边缘集群的配置，或在一组边缘集群上同时更新应用程序。

专为边缘场景量身定制：以最低成本获得最佳性能和安全性

图 1：基于 GitOps 用于管理边缘基础设施的工作流及其与 SUSE Edge 系统架构的映射

针对容器优化的 OS：在 Kubernetes 堆栈中，Linux 操作系统通常只需要企业 Linux 发行版的小部分功能。而 SUSE Linux Enterprise (SLE) Micro 是为作为 Kubernetes 堆栈中的主机操作系统专门设计的。

安全：针对容器优化的操作系统更安全，因为它的攻击面要小得多。此外，SLE Micro 有一个不可变的根文件系统，攻击者几乎不可能修改系统设置，也无法添加或替换标准系统路径中的二进制文件。

可靠：SLE Micro 的事务更新机制保证了更新的原子性，即更新要么成功要么失败，如果失败，系统会自动回滚到之前的状态。这样能避免更新失败导致系统处于损坏状态的情况。

使用 NeuVector 实现零信任云原生安全：NeuVector 是市场领先的云原生零信任安全解决方案，不仅提供了传统的漏洞扫描和安全策略合规工具，还具有观察和监控云原生应用运行时行为的独特能力（也称为零信任安全），能抵御传统安全工具无法检测到的大范围攻击媒介，让安全性能达到新的高度。

在边缘使用电信级 OS 和 Kubernetes：轻松可靠

在高数据包吞吐量和可靠访问系统资源（CPU 和内存等）方面，云原生网络功能 (Cloud Native Network Function，CNF) 有特殊的要求。

有时，由于应用程序需要访问 GPU、加速板或类似设备，你需要为集群中的特定硬件节点调度应用程序。

企业 Kubernetes 和 Linux 的典型功能和配置难以满足这些要求。**而 SUSE ATIP 的设计包含了一组专用技术，用于助力电信行业的典型用例。**

Node Feature Discovery

Node Feature Discovery（NFD）是由 CNCF 开发的 Kubernetes 附加组件。它能检测底层硬件的特性并相应地标记 Kubernetes 节点，以便 Kubernetes 调度程序在调度工作负载时能考虑这些特性。你可以在 Rancher Manager 集群工具中使用 Node Feature Discovery，并且通过 GUI 或集群 YAML 定义轻松将其安装到每个下游集群上。

使用 Multus CNI 让每个 Kubernetes Pod 拥有多个 IP 地址

Vanilla Kubernetes 仅支持每个 Kubernetes Pod 拥有一个 IP 地址。有了 Multus CNI，你可以为 Kubernetes Pod 分配多个 IP 地址。Multus 只能与一个额外的 CNI 结合使用，该 CNI 服务于正常的集群内部网络（例如 Control Plane 和 Worker 节点之间的连接）。Whereabouts 用于确保在使用 Multus 时 Kubernetes Pod 无重叠 IPAM。Multus 是 Rancher Kubernetes Engine (RKE2) 中受支持的 CNI，可以在创建集群时自动安装。你可以通过集群定义中的功能开关同时安装 Whereabouts。

单根 I/O 虚拟化 (SR-IOV)

SR-IOV 是用于允许单个物理 PCIe 设备显示为多个虚拟 PCIe 设备的技术，通常用于允许从 Kubernetes Pod 直接访问网络接口，而且没有内核中断处理的开销，也能支持其他类型的 PCIe 设备（例如，加速板、FPGA 板等）。SUSE ATIP 解决方案使用了上游 Kubernetes SR-IOV Operator（通常与 Multus CNI 结合使用）来创建 SR-IOV VF（Virtual Function）并将其公开给 Kubernetes Pod。Rancher 可以在安装时通过集群工具市场或 Helm Chart 控制器将 SR-IOV Operator 部署到下游集群。Operator 支持通过 Kubernetes Custom Resource 创建 SR-IOV VF。

卓越的电信级性能

CPU Pinning 和 IsolCPUs

电信工作负载需要处理大量用户流量，并以确定的性能执行计算量大的操作。Linux 中的普通用户态进程不会一直获得完全确定的性能，这是因为 Linux 内核调度程序可能会将进程移动到不同 CPU 核，或者进程可能必须与其他进程或内核本身共享对物理系统资源的访问，因此其执行速度可能会受影响。

在这种情况下最需要的功能是 CPU Pinning，它确保进程始终被调度到特定的 CPU 核，避免了缓存未命中和其他影响性能的因素。
另一个相关的功能是 CPU 隔离（也称为 IsolCPU），它可以防止其他进程（包括内核本身）在同一个 CPU 核上运行。所有 SUSE Linux 变体中使用的 Linux 内核都支持 CPU Pinning 和 IsolCPU。

Huge Pages

在为数据包 I/O 读写 SR-IOV VF 时，电信 CNF 需要对大量内存进行低延迟访问。Huge Pages 显著减少了内存页查找时间和后备缓冲 (TLB) 未命中的情况。所有 SUSE Linux 变体中使用的 Linux 内核都支持内置的大内存页。Kubernetes 会自动发现大内存页，并以与标准内存和 CPU 相同的方式将它们作为可调度资源提供。

Kubernetes 中的拓扑感知调度

有时，确保 Kubernetes 工作负载分配到恰当的资源集（CPU 内核、内存、PCIe 总线）是非常重要的。为此，Kubernetes 提出了 QoS 类和资源请求的概念。

由于现代多处理器系统的非统一内存访问 (Non-Uniform Memory Access，NUMA) 架构，某些资源可以通过特定 CPU 插槽和核以更高的带宽访问。简而言之，你可以想象为某些资源之间的距离比其他资源“更近”。换句话说，系统中的资源具有拓扑结构。

拓扑感知调度（例如，将作为 SR-IOV VF 公开的 NIC 放在运行指定 Kubernetes Pod 的同一 NUMA 区域，并确保从该区域分配大页内存）是保证性能敏感的应用程序可以在 Kubernetes 上成功运行的重要因素。

从 Kubernetes 1.22 开始，Kubernetes Memory Manager 和 CPU Manager 支持为 Kubernetes Topology Manager 生成 NUMA 感知的亲和性提示，然后根据这些提示及其配置的分配策略将 Pod 分配给 Kubernetes Worker 节点内的资源。因此，要正确设置 NUMA 感知，你只需要正确配置系统（包括应用程序清单）的各个层。
你可以通过 kubelet 标志配置这些管理器，而这些标志可以在创建新集群时通过 Rancher 自由配置。

实时内核

一些电信应用程序对触发事件后多久能完成调度有一定的要求。在一般的 Linux 系统中，触发事件（例如，软件中断）的延迟取决于系统上运行的内容，因此在很大程度上是不确定的。这样的系统可能会提供足够的性能，但却不能保证确定的性能。Linux 内核的 PREEMPT_RT 补丁提供了调度任务的能力，可以计算出访问 CPU 核的可靠上限。包含 PREEMPT-RT 补丁的 Linux 内核通常称为实时内核。SLE Micro 提供了使用 SUSE Linux Enterprise 实时内核的选项。

支持电信相关协议和精确定时

精确时间协议和同步以太网支持

多个本地系统（例如边缘 Kubernetes 集群中的多台裸机主机）需要在亚微秒级别内同步时间时，网络时间协议 (NTP) 等典型的时间协议是不够的。为此，IEEE 标准 IEEE1588（也称为精确时间协议，PTP）应运而生。SUSE Linux Enterprise 为 Linux 中可用的所有相关 PTP 实现提供了包。

同步以太网是用于在以太网物理层上传输时钟信号的 ITU-T 标准（定义在 ITU-T Recommendations G.8261、G.8262 和 G.8264 中），用于为 L2 广播域中的所有节点提供时间同步源。Linux 内核中的同步以太网 (SyncE) 支持目前正在进行中，预计将于 2023 年上半年推出 6.2 版本。然后 SUSE 会将代码向后移植到 SUSE Linux Enterprise 内核中。

流控制传输协议 (SCTP) 支持

从 1.20 开始，SCTP 支持已成为 Kubernetes 中的一项稳定功能，它要求 Kubernetes 部署支持 SCTP 的 CNI。在 RKE2 支持的 CNI 中，支持 SCTP 的 CNI 包括 Calico 和 Cilium。

支持基于数据平面开发工具包 (DPDK) 的容器工作负载

目前 SLE Server 15 和 SLE Micro 5.3 支持 DPDK 1.19。从 SLE 15 Service Pack 5（计划于 2023 年 Q2 发布）开始，SUSE 将支持 DPDK 1.22.x。

SUSE Package Hub 和 openSUSE 中也提供了其他 DPDK 版本。

摘要

我们对下一代电信基础设施的愿景是，让 5G 和边缘计算使能新型应用程序和用例。
SUSE ATIP 通过提供灵活、适应性强的基础架构来支持此类新型应用程序，帮助电信运营商实现网络转型。管理分散在裸机基础设施、公有云和私有云中的大量集群是一项挑战，而 SUSE 能帮助用户应对这一挑战。同时，我们 30 多年来一直提供关键业务的 Linux 解决方案，能满足你所期望的功能集和安全要求。

要了解更多信息，请访问电信行业解决方案、边缘解决方案

或者通过 https://www.suse.com/contact/ 联系我们