CCF-华为胡杨林基金 系统软件专项-2022年操作系统课题指南介绍

华为技术有限公司 ICT操作系统领域架构师 胡欣蔚

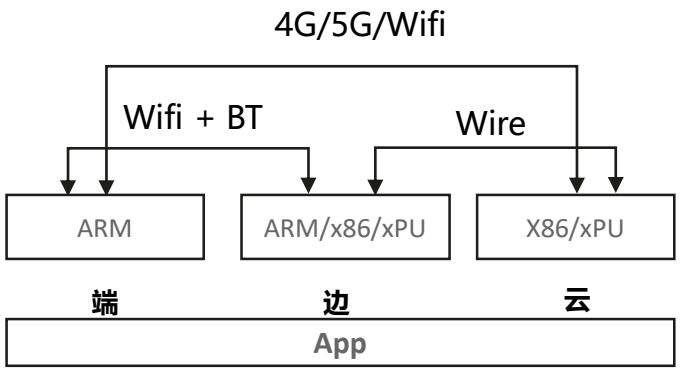
2022年5月15日

系统软件专业委员会
Technical Committee of Systems Software

HUAWEI

面向云网边端协同的操作系统

产业需求和挑战



1. **云网边端协同的操作系统**：极轻量的设备也可以具备富生态能力；以人为中心的多方节点协同操作系统
2. **管理海量异质泛在资源**：云网边端不同的链接方式与不同的硬件架构，提供协同的应用开发、执行和维护；保障以人为核心的系统的安全可靠

潜在技术方向

数字世界：泛在应用

数字化智能化

应用接口
场景化、领域化的软件服务

系统执行、保护、管理
多样性资源调度；无损信息数据通信；多层次可靠性保障。

环境接口
快速、全面感知理解外部环境

可靠性安全性

物理世界：人、机、物

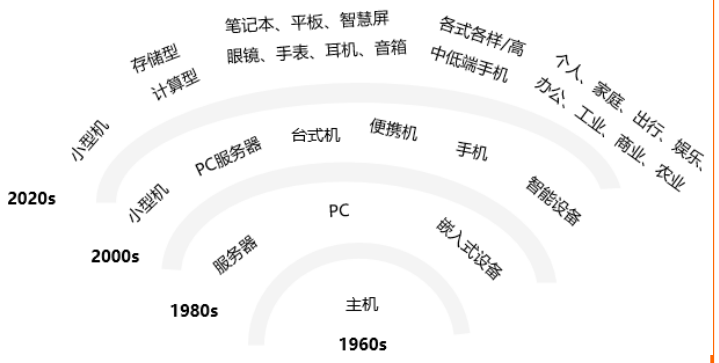
1. **泛在操作系统**
2. **软件统一总线**
3. **程序自动分割与卸载**

技术挑战

1. **分布式协同的操作系统架构**
   - OS架构支持组件运行在多个设备，极轻量设备无需运行完整OS
   - 基于WIFI等近场网络实现 RPC 与内存访问语义
2. **算力运筹**
   - 计算负载的低成本跨节点协作和迁移
   - 根据可用硬件资源与网络状态，确定分割与迁移的最佳方案
3. **多层次可靠性保障**
   - 以人为中心向外逐层提供冗余备份、隐私加密、攻击防御等安全可靠保障系统

面向多样化场景、又具备通用性的新型OS架构

产业需求和挑战



1. “**昆虫纲悖论**”：一方面，5G和AIOT推动了千行百业创新加速，硬件和应用形态快速“进化”、“跨界”和“演变”，带来万亿数量级的巨大市场；另一方面，由于各种形态间的差异化大，碎片化导致难以规模复制，也就难以形成巨大市场。
2. **如何打造一个OS架构，使得它既可以满足多样化场景定制化，又可满足通用性规模复制?**

潜在技术方向

千行百业数字化创新

生态接口
面向异构算力和差异化编程的统一接口

乐高化的OS架构
组件化、弹性可组合，极简、乐高化地生成千行百业所需的基础软件栈

硬件抽象
跨异构算力、跨芯片平台、分布式

多样化的 硬件 & 芯片

1. **乐高化、弹性可组合的OS架构**
2. **面向南向的新型硬件抽象**
3. **面向北向的统一应用生态接口**

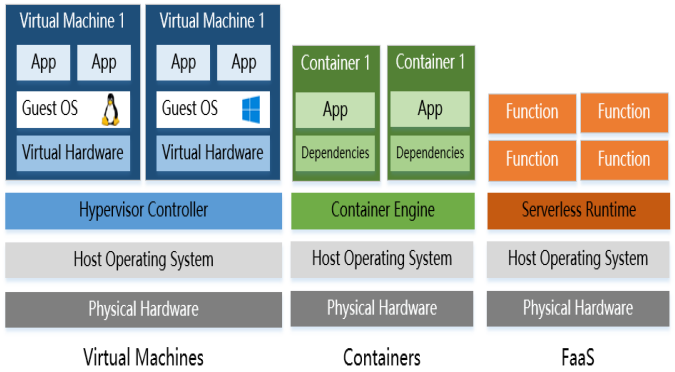
技术挑战

1. **新型OS架构突破**
   - 要求可根据差异化场景的资源大小、功能、性能、安全、生态等千差万别诉求进行量身定制
   - 同时具备通用性，可规模复制
2. **面向南向的新型硬件抽象**
   - 要求能够跨异构算力、跨主流的芯片指令集，支持分布式多设备协同等
3. **面向北向的统一应用生态接口**
   - 要求能够屏蔽异构算力、异构指令集、差异化编程框架

HUAWEI

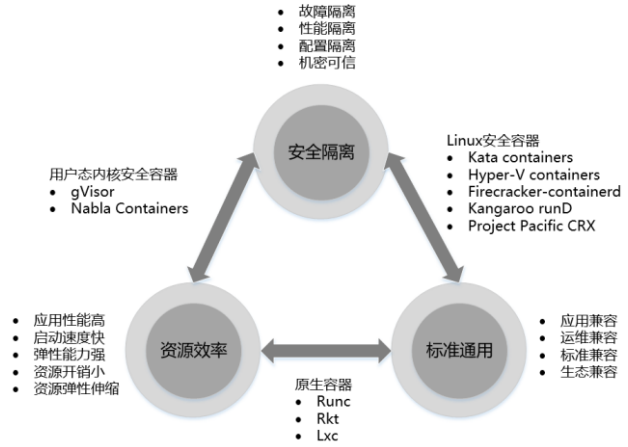
虚拟机&容器的下一跳计算抽象

产业需求和挑战



1. 云计算推动了虚拟机&容器的快速发展，但**万物互联的智能时代呼吁下一跳计算抽象**
2. **函数计算、原子化服务等新的编程范式**使得虚拟机&容器运行时显得臃肿和迟缓，面向服务的更高效运行时技术成为未来焦点
3. **短寿命：业界调查显示22%的容器 < 10 秒，一周内停止率高达8%**
4. 传统容器难以面向多租户构建安全隔离环境

潜在技术方向



图片参考网络：《容器技术20年：容器引擎与江湖门派》

1. 安全容器+Unikernel：通过安全容器实现负载间更好的隔离；而与Unikernel的结合来进一步降低系统资源开销，减少攻击面，并提升运行性能
2. 通过硬件隔离机制创新

技术挑战

1. 虚拟化和容器直接结合，更安全但性能不理想：

- 基于虚拟机或用户态内核构建安全容器，减少容器逃逸的风险；但也导致应用性能下降和底噪升高
- 虚拟化层导致IO性能显著下降。若通过设备直通提升性能，则大幅增加服务成本；
- 虚拟机的资源使用底噪对于短时容器过于臃肿。

2. 需要创新的软、硬件内存安全隔离技术

成熟技术存在较多较多限制

- SFI通过编译插桩实现内存隔离，不支持动态新增内存隔离区；
- Intel MPX支持的内存分区数量有限，只支持16个keys标识隔离区；
- Intel SGX硬件可控制的内存大小有限，应用适配复杂，进出隔离区开销大；
- KASAN会引入额外大量开销，占用额外的CPU资源。

新技术的隔离能力和结合方式需要探索

- 包括但不限于CHERI, TDX, CCA 等

参考文献

[1]. Sultan, S., Ahmad, I., & Dimitriou, T. (2019). Container security: Issues, challenges, and the road ahead. IEEE Access.

[2]. Watson, R. N., Woodruff, J., Neumann, P. G., Moore, S. W., Anderson, J., Chisnall, D., ... & Vadera, M. Cheri: A hybrid capability-system architecture for scalable software compartmentalization. 2015 S&P (pp. 20-37). IEEE.

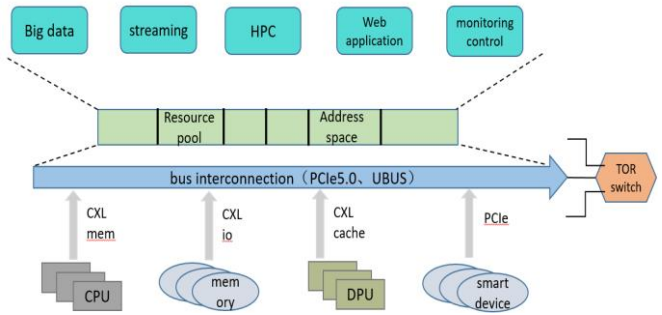
[3]. CHaOS: CHERI for Hypervisors and Operating Systems. Computer Laboratory, University of Cambridge.

[4]. CloudCAP: Capability-based Isolation for Cloud Native Application. Dept of Computing, Imperial College London.

HUAWEI

多样性算力的任务调度

产业需求和挑战



硬件结构的变化：

- 单节点从host-device结构，演变为更对等的结构
- 集群从分布式结构，演变为机柜级池化共享的结构

算力多样性对系统设计和任务调度产生挑战：

- 影响因素多
- 决策目标多
- 使用限制多

潜在技术方向

1. **协同调度**
   - **新型调度框架**：实现可扩展到更多内存层次、更多样化算力和更大互联节点规模的调度框架。
   - **新型调度算法**：实现复杂度为O(1)的调度算法，支持不同硬件形态和互联结构。
2. **异构OS**
   - **系统抽象和接口**：自动将业务模型转化为运行实体。
   - **多样化算力管理**：OS自身可以充分利用异构算力，优化性能；OS功能平滑扩展到越来越大的异构算力规模

技术挑战

将多样性算力当做一个整体资源进行调度

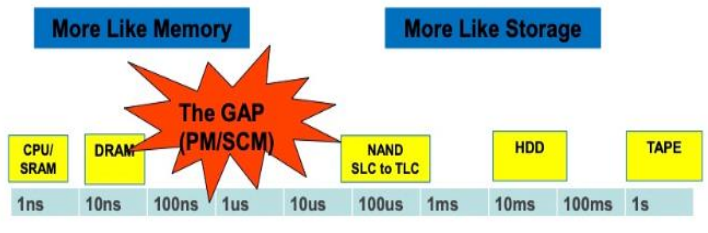
1. **收益-成本模型复杂**
   - 多样性算力的对等架构，复杂组网和异构能力，使得调度收益的建模、估计、分析极其复杂。当前通用模型和策略不再适用。
2. **融合调度**
   - 如何实现数据调度和业务调度的协同；实现不同XPU间调度的协同，充分发挥算力以及拓扑的局部性优势。

参考文献：

- Placement of Virtual Containers on NUMA systems: A Practical and Comprehensive Model. In USENIX ATC 2018
- Breaking the Boundaries in Heterogeneous-ISA Datacenters. In ASPLOS 2017
- Thread and Memory Placement on NUMA Systems: Asymmetry Matters. In USENIX ATC 2015

面向持久内存的Single-Level存储架构研究

产业需求和挑战



随着内存工艺不断升级，一方面内存容量不断变大，且逐步具备持久化能力，但另一方面，内存在生命周期内出错概率也呈倍数上升。

基础软件需要提供合适的方案，一方面为上层应用提供持久内存存储服务，充分发挥内存性能，另一方面通过对内存故障的预测、管理，降低内存出错对应用和系统整体的影响。

潜在技术方向

- **Single-Level存储系统：**简化合并操作系统对内存和存储的管理，精简软件栈的同时，释放持久内存在延迟，高效持久化，CPU直访等方面的优势，规避其在带宽干扰，并发等方面的劣势
- **高效的持久存储原语：**探索新型持久存储原语的设计与实现，区别于传统的使用文件系统接口管理持久存储设备的方法，新持久存储原语结合Single-Level存储系统可以为上层提供更为高效易用的持久存储服务
- **精准的内存故障预测机制：**通过运行过程中硬件和系统的一些特征参数预测内存故障，并据此决策该内存是否做容错机制。
- **透明智能的内存可靠性管理机制：**面对多元化介质和多样性的业务场景，智能感知业务内存数据特征，在线评估内存可靠性风险，根据重要性分级分区存放，并进行局部内存镜像实现数据透明容错。

技术挑战

- **Single-Level存储系统缺少架构设计。**

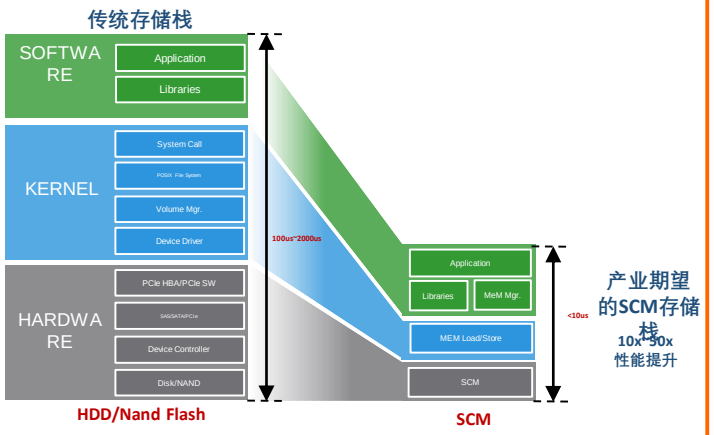
当前在学术界，软件栈优化的相关研究多将存储栈置于用户态以Bypass内核软件栈，但在性能、安全性、通用性方面存在不足；而现有针对内核存储软件栈的研究大都面临场景受限的问题，且设计有较大优化空间；

- **故障内存预测精度不足：**

通过软件对预测到的故障内存提前隔离和重映射，可以按需动态调整内存镜像区域，达成内存可靠性和内存容量开销的权衡。依赖内存UCE预测的精确度，但目前业界的一些方案的内存故障预测的精度比较低，如何高效准确地在业务运行过程中根据一些特征识别内存颗粒可能出现无法修复的错误。在多元化介质下，如何在线自适应不同的介质，达到高预测准确率。该预测机制最好能在线自适应不同的介质，准确率达到80%。

面向新介质的高性能高可靠文件系统

产业需求和挑战



- 非易失内存进入存储领域，存储介质时延进入微秒级别，传统软件存储栈过于厚重，严重影响系统端到端性能；
- 当前主流文件系统已过于复杂，Bug层出不穷且不收敛，其中会造成数据损坏的严重Bug比例大；
- 文件系统设计需考虑掉电一致性，在关键数据/元数据写入时进行加锁及数据同步操作，影响端到端访问性能及并发性能

潜在技术方向

- **面向SCM的新型高并发数据一致性保护技术：**突破新型一致性技术方法，实现轻量级、高并发的一致性保护技术，做到事务延迟相比传统技术降低一半，FxMark并发性能超业界前沿工作（如KucoFS）1倍
- **高可靠文件系统技术：**探索新可靠性技术，在性能不低于传统文件系统基础上，保证文件系统可靠性，解决因软件逻辑导致的数据损坏问题

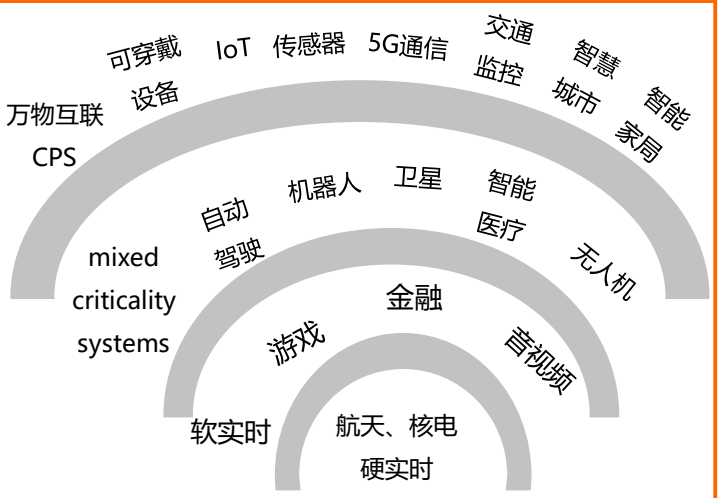
技术挑战

- **新的一致性保护算法设计：**传统一致性保护算法实现流程复杂，采用了大量数据拷贝及同步机制，严重影响业务性能
- **文件系统的可靠性验证：**形式化方法可有效保障系统的可靠性，但文件系统的复杂性使得完全验证的代价难以接受
- **安全编程语言与文件系统开发的结合：**基于Rust等类型安全语言开发文件系统，只能解决内存相关问题

参考文献：
[1] Scalable Persistent Memory File System with Kernel-Userspace Collaboration. In FAST 2021
[2] HashFS: Rethinking File Mapping in Persistent Memory. In FAST 2021
[3] Soft Updates Made Simple and Fast on Non-volatile Memory. In USENIX 2017
[4] Using Crash Hoare Logic for Certifying the FSCQ File System. In SOSP 2015
[5] Verifying a high-performance crash-safe file system using a tree specification. In SOSP 2017
[6] High Velocity Kernel File Systems with Bento. In FAST 2021

面向高性能硬件、多功能软件的CPS操作系统

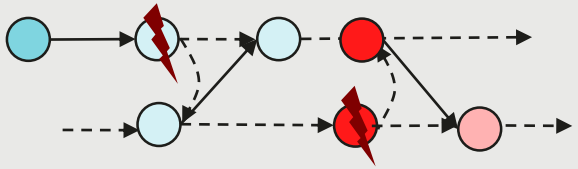
产业需求和挑战



1. **下一代CPS系统**：随着硬件技术不断提升和软件功能飞速发展，CPS系统渴求突破原有的固定硬件和封闭软件，以适应现有的软硬件环境；但因其直接影响物理世界和人身安全，其可靠性、确定性和正确性必须保证。
2. **如何打造一个CPS系统，使得它既可以使用高性能硬件和多功能软件，又可满足CPS对可靠性、确定性和正确性的要求?**

潜在技术方向

Global调度（动态迁移）



- **多核多任务的硬实时调度：**嵌入式高性能硬件往多核方向发展；软件功能复杂化，往多任务方向发展；需要能够保障多核多任务的硬实时调度能力。
- **混合关键性调度：**在传统的实时调度的基础上，结合不同场景的实时、功能安全诉求差异，支持在单一硬件上保障不同关键性等级OS的指标达成。

技术挑战

■ **多核硬实时支持：**

- 任务拆分至多核并发缩短执行周期，提升单位时间算力，多核任务编排满足硬实时约束的量化复杂度急剧上升。

■ **混合关键性系统需要平衡资源利用率与不同关键性等级隔离：**

- **资源利用率有待提高：**在满足嵌入式场景约束的前提下，如何通过机制、算法实现资源共享，提升OS部署密度
- **难以满足不同关键性等级的需求：**实现资源共享后，如何通过以OS为单位的混合关键性调度，保障不同安全性等级达成，特别是高实时、高功能安全OS

参考文献：

- 《DAG Scheduling and Analysis on Multiprocessor Systems: Exploitation of Parallelism and Dependency》
- 《Priority Assignment on Partitioned Multiprocessor Systems with Shared Resources》

融合数据驱动的系统设计与运行管理

产业需求和挑战

长期以来，OS只建立了从资源信息到资源管理的单一链接。实际上，**作为“中间层”，OS也能从它的视角看到数据、计算等要素，却极少收集和使用它们。**

目前，OS的角色仅是在来自上层的数据和计算，以及来自下层的资源都已经确定的情况下，完成好自己的管理衔接工作。那么当三方面的复杂度皆呈爆炸性增长时，OS是否应该保持“袖手旁观”？它是否应该利用能够看到三方面要素的独特性质，走出现有的职能范式，整合这些信息，对上层的数据及计算模式和下层的资源结构主动提出改良建议，形成“外溢”效应？

潜在技术方向

1. **学习型操作系统：**基于数据驱动，重新思考系统设计，保证动态调整，持续演进；
2. **系统智能运维：**准应用灰度故障模型与立体故障树实时构建技术；
3. **系统调优：**根据业务负载系统自动调优，获取的性能数据进一步辅助芯片设计；

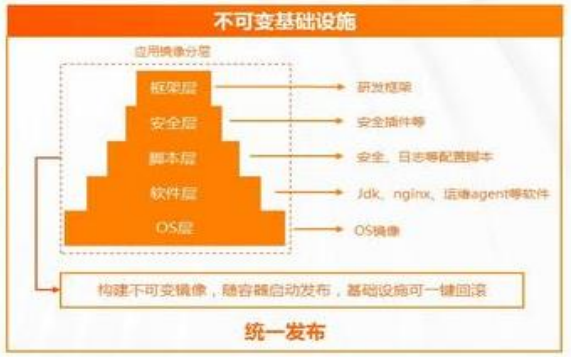
技术挑战

- 业务多样化，数量海量化，资源布局复杂化，作为联结的操作系统如何帮助上层应用与自身，动态调整自身设计，以达成状态最优?
- .....
- **基于应用、系统KPI的异常检测无法识别灰度故障**：软件设计的模块化原则导致应用/系统之间必然存在观测差异，所以现有技术（有/无监督类异常检测算法等）均只能片面评估应用/系统健康度。
- **基于相关性、知识图谱构建调度故障树准确性、可解释性不足**：现有技术（包括知识图谱类、事件相关性类）无法综合平衡准确性、可解释性；进程/线程间隐式动态故障扩散，进一步对故障树构建实时性提出更高要求。

参考文献：
[1] hotos 17 Gray Failure: The Achilles’ Heel of Cloud-Scale Systems
[2] OSDI 18 Capturing and Enhancing In Situ System Observability for Failure Detection
[3] NSDI 20 Understanding, Detecting and Localizing Partial Failures in Large System Software
[4] IEEE 21 Graph-based Incident Aggregation for Large-Scale Online Service Systems
[5] IEEE 21 Predicting gray fault based on context graph in container-based cloud

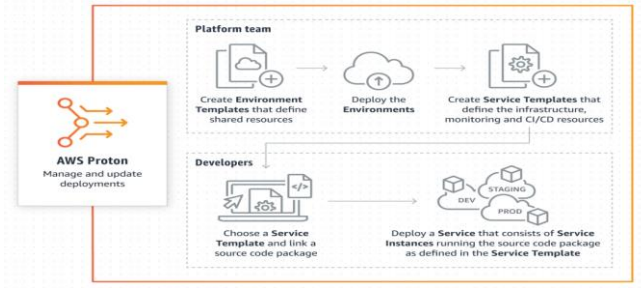
支持全栈敏捷迭代的基础设施

产业需求和挑战

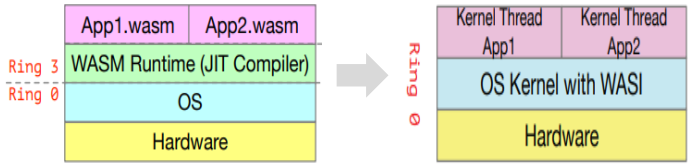


1. **“可变”->”不可变”->?**：不可变基础设施类似于程序设计中”不可变”理念，屏蔽了不同服务器间配以依赖差异，但也要求应用的无状态化；随着云原生步入中场，各类应用均在走向云原生化，不可变基础实施带来的整体一致与敏捷创新产生了矛盾
2. **传统Devops仅仅面向应用框架与应用层，但实际上为了获取极致最优的体验、性能等，当前更多的需要全栈协同的敏捷迭代**

潜在技术方向



AWS Proton定义预置、部署、监控，包括计算、网格、代码管道、安全性和运维等原子能力，创建整体”堆栈”，利用API Model声明式定义操作系统



Wasmachine 尝试利用通用IR达成边缘场景下“可编程式”内核

1. **最小不可变+ 组合式 OS新形态**
2. **灵活可编程的OS Kernel**

技术挑战

1. **OS新形态**
   - 原子化组合方式带来组合状态爆炸，对操作系统以及其上的软件堆栈的升级、调试、运维等带来管理困难
   - 不同的场景是否需要确定不同的最小不可变范围，是否需要一种范式来便捷推广到更多场景
2. **可编程内核**
   - 灵活与安全稳定是矛盾体，如何界定安全、非安全的边界，在保证稳定性不下降的前提下进行灵活定义Kernel。

欢迎向 OpenHarmony 和 openEuler 两个操作系统社区提交开源实现

gitee.com/openharmony

gitee.com/openeuler

HUAWEI

把数字世界带入每个人、每个家庭、每个组织，

构建万物互联的智能世界

Bring digital to every person, home and organization

for a fully connected, intelligent world

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI