CCTC 2016 中国云计算技术大会
Cloud Computing Technology Conference 2016
CSDN

Container + TianHe 2

Yusong Tan
Li Wang
2016.5

内容

内容

1. TH2 & KylinCloud概述
2. Container+KylinCloud@TH2
3. 国产飞腾平台支持

天河二号超算平台

- 目标定位
  - 满足高性能计算、高吞吐率信息服务和海量数据处理等多领域复杂应用需求
  - 科学、工程与教育的开放支撑平台
  - 支撑智慧城市建设的公共基础设施平台
- 国家超级计算广州中心，目前系统20000个节点
- 峰值计算性能54.9PFlops，连续六次位居国际Top500榜首

KylinCloud

- 基于OpenStack的增强和定制
  - 为用户提供IaaS、PaaS层次的云服务解决方案，构建公有云和私有云服务
  - 提供虚拟化管理、项目管理、系统资源管理、资源状态监控、告警管理、计费管理、用户管理等功能

KylinCloud
Based on OpenStack

openstack CLOUD SOFTWARE

ceph

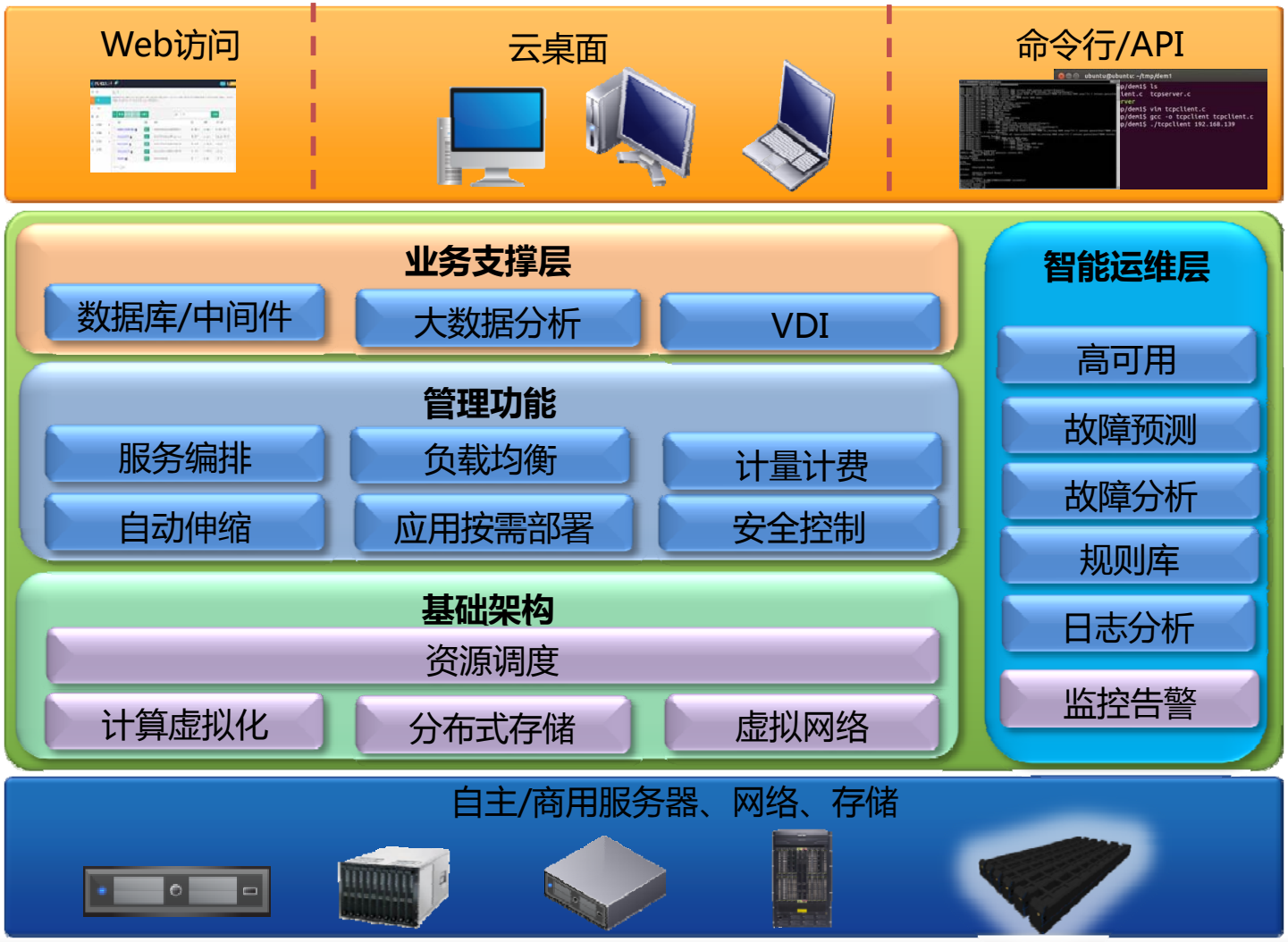
puppet labs

MIRANTIS

docker

ubuntu

技术架构


Web访问
云桌面
命令行/API
增值服务工具套件
IaaS云平台
业务支撑层
数据库/中间件
大数据分析
VDI
管理功能
服务编排
负载均衡
计量计费
自动伸缩
应用按需部署
安全控制
基础架构
资源调度
计算虚拟化
分布式存储
虚拟网络
智能运维层
高可用
故障预测
故障分析
规则库
日志分析
监控告警
自动化运维平台
自主/商用服务器、网络、存储

社区贡献

- 积极参与OpenStack社区并贡献代码
  - 自F版以来贡献代码数接近3万行，1411个Patch
- Liberty版代码commit数排名国内第四，Mitaka版本第六
- 一名Rally的core，一名SearchLight的core

Liberty

团队	Commit数	社区排名
huawei	1074	6
99Cloud	261	15
UnitedStack	93	30
KylinCloud	**57**	**41**
EasyStack	18	68
AWCloud	17	72

Mitaka

团队	Commit数	社区排名
huawei	1073	9
EasyStack	806	12
99Cloud	609	14
UnitedStack	115	37
AWCloud	118	38
KylinCloud	**99**	**43**

*数据来源：http://stackalytics.com/

社区贡献

- 积极参与Ceph社区并贡献代码
  - 向Ceph社区提交多个重要features
    - Rbd in userspace
    - Rbd offline recovery tool
    - Rbd clones copy on read
    - Rbd diff merge tool
    - Cache tiering writeback throttling
    - CephFS quota support
    - CephFS punch hole support
    - CephFS inline data support
  - 多个Ceph版本的代码贡献排名前列
    - 在2016年1月发布的Ceph v10.0.0.2中，团队代码贡献排名第四

TH2+KylinCloud应用情况

- 国家超级计算广州中心
  - 基于KylinCloud搭建了生产环境，最大部署规模达到6400节点，其中目前正在稳定运营近3000个节点提供商业服务
  - 为电子政务、动漫渲染等多个行业和应用提供服务
  - 入选2015 OpenStack SuperUser四强

内容

1. TH2 & KylinCloud概述
2. Container+KylinCloud@TH2
3. 国产飞腾平台支持

Container+KylinCloud@TH2

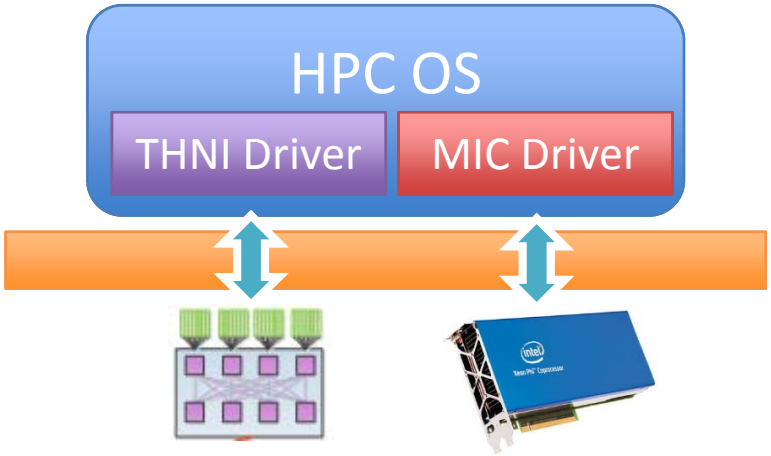
- 为啥要容器？
  - 性能好！
    - 性能折损低
    - 高效能地计算服务：HPC、大数据处理服务
  - 够灵活！
    - 以容器作为资源管理与调度的基本控制单位
    - 提供基于容器的可定制、强隔离、可伸缩的资源调度框架

基于Container的弹性HPC服务

- 目的
  - 将云计算的优势引入到传统的高性能计算领域，使用容器技术为用户提供可定制、隔离性强、可弹性伸缩、界面友好的HPC集群
- 主要需求
  - 接近物理机的性能
  - 充分利用天河节点的加速部件和高速网
  - 拉通HPC软件栈，实现全系统资源统一管理
  - 可定制的软件环境
  - 满足弹性的资源使用需求
  - 更好的资源共享与隔离

基于Container的弹性HPC服务

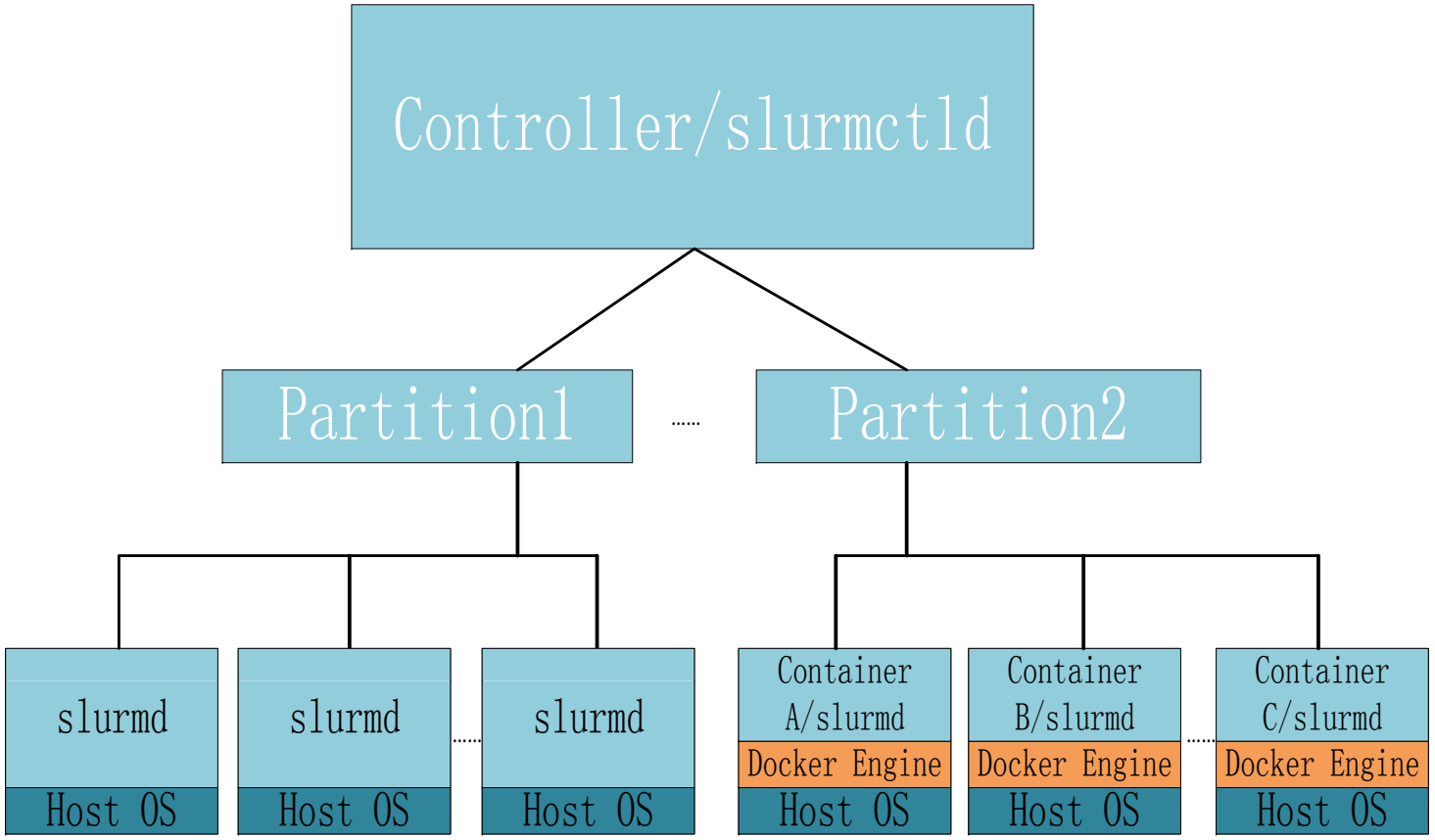
- 主要措施与优化
  - 使用基于容器的轻量级虚拟化
    - 基于HPC环境制作镜像，整合HPC软件栈
    - 虚拟机以Passthrough模式访问MIC加速器和TH-NI高速网
    - 虚拟机直接访问主机挂载的Lustre共享存储
  - 全系统计算资源的混合调度
    - 根据Image属性判断Host
    - 计算节点角色的按需切换



基于Container的弹性HPC服务

- Slurm混合集群统一管理策略
  - 部署过程
    - 管理员根据用户软件栈需求通过KylinCloud选择docker镜像
    - 管理员通过KylinCloud创建docker集群
    - 管理员将docker虚拟节点加入slurm管理
    - 管理员基于docker虚拟节点为用户分配slurm虚拟节点
    - 用户利用slurm虚拟节点进行作业管理

基于Container的弹性HPC服务



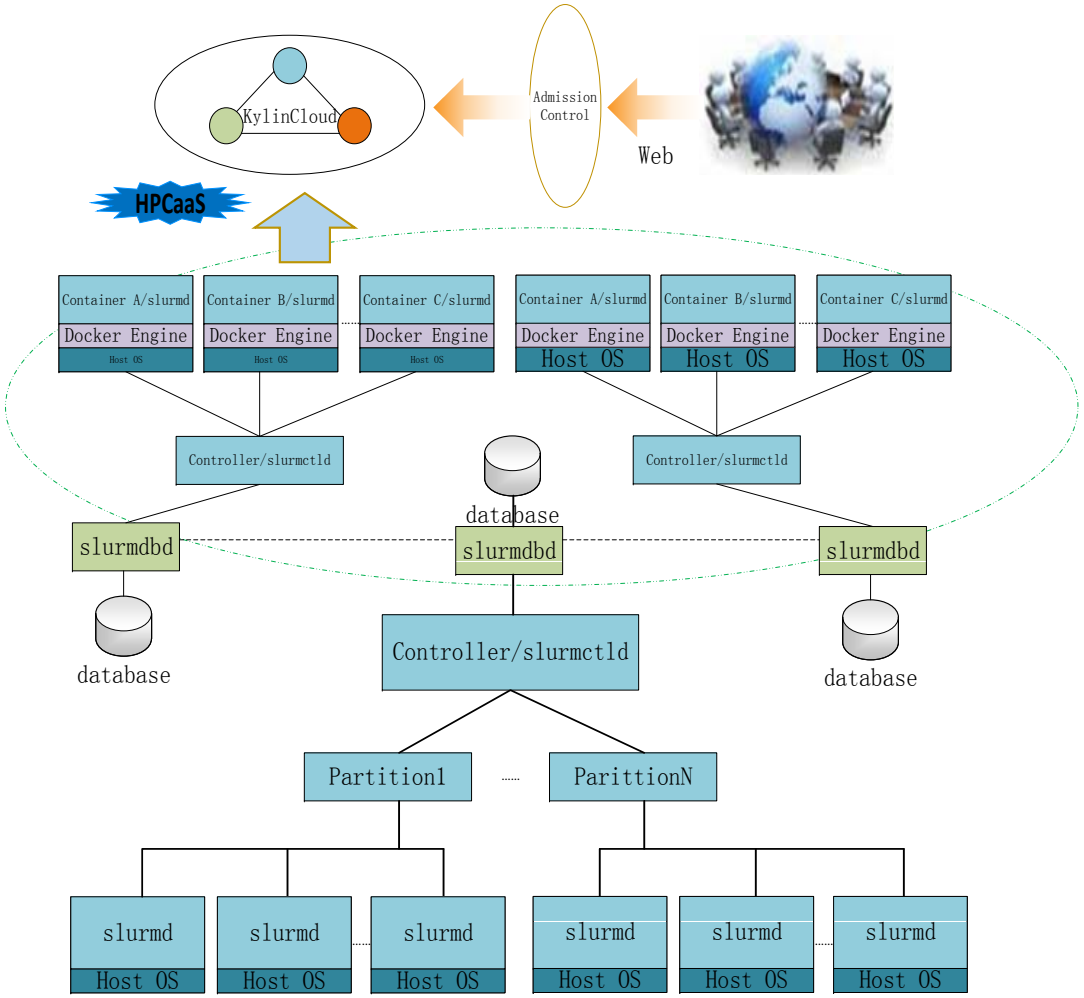
基于Container的弹性HPC服务

- Slurm混合集群统一管理策略
  - 优点
    - 使用同一套slurm数据库来维护状态,资源计费简单
  - 缺点
    - 管理员工作量相对较大
  - 难点
    - 基于虚拟层TCP/IP网络协议，优化实现物理节点和虚拟节点的统一管理
    - 通过低延迟网络和节点状态，优化实现slurm集群管理系统资源分配

基于Container的弹性HPC服务

- 独立管理策略
  - 部署过程
    - 管理员为用户分配slurm物理节点
    - 用户通过KylinCloud选择docker镜像
    - 用户通过KylinCloud基于slurm物理节点创建docker集群
    - 用户利用docker虚拟节点创建slurm集群
    - 用户基于slurm集群进行作业管理

基于Container的弹性HPC服务



基于Container的弹性HPC服务

- 独立管理策略
  - 优点
    - 管理员工作量相对较小
    - 用户灵活度较高,可以自己创建和管理slurm集群
  - 缺点
    - 资源计费相对困难
  - 难点
    - 节点状态弱一致性异步更新
      - 采用不同数据库异步更新方式，保持虚拟集群slurm与物理slurm集群节点状态一致性

基于Container的弹性HPC服务

- 优势效果
  - 灵活的HPC环境提供
    - 不同的运行环境：MPI、OpenMP、Fortran、Java
    - 不用的运行库版本
    - 硬件加速卡
  - 快速部署支持
    - 预先导入镜像，快速启动
  - 弹性规模伸缩
    - 根据作业规模动态调整节点规模
  - 低性能折损
    - 基于TH2的测试数据表明，CPU性能损耗在**1%**以内，高速网络带宽损耗在**1%**左右

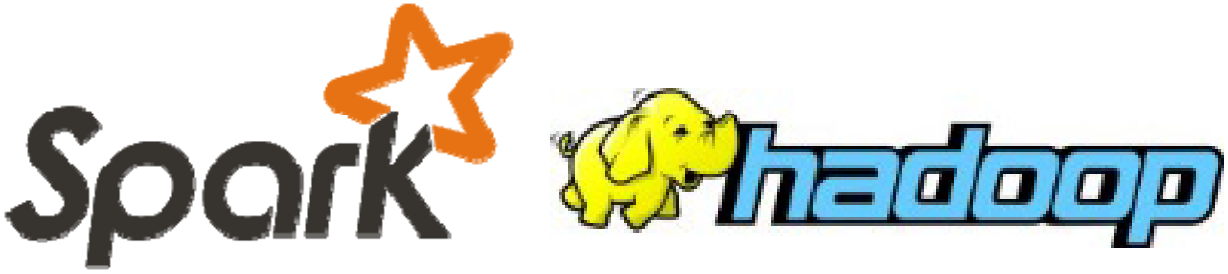






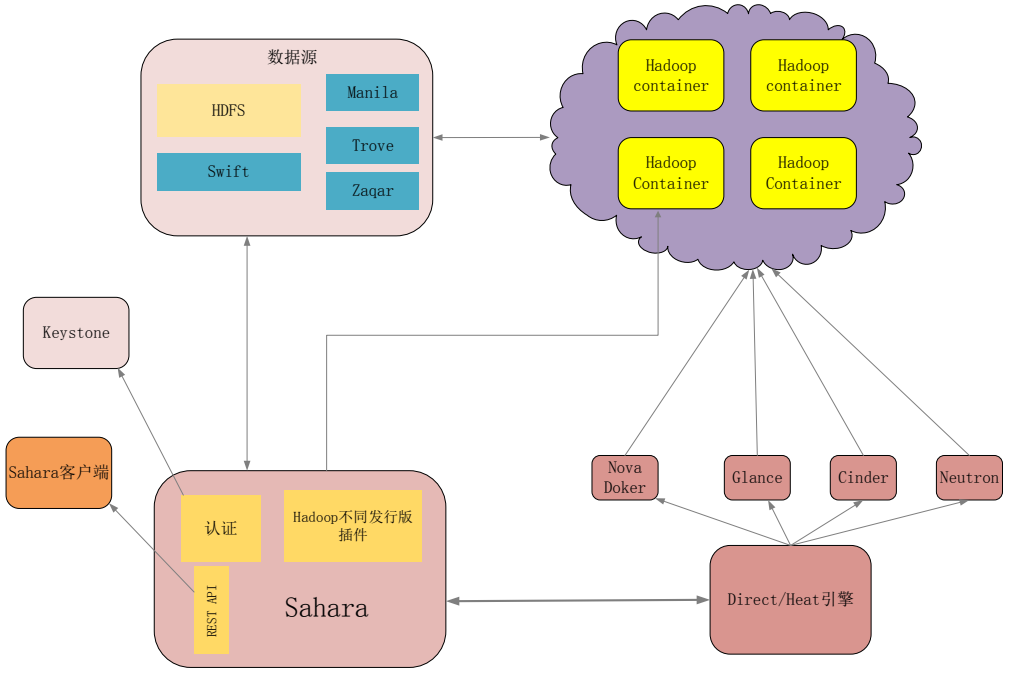

基于Container的大数据处理平台

- 目的
  - 基于云平台的大规模集群能力，结合容器技术，搭建Hadoop大数据处理平台，为用户提供高性能的海量数据处理能力
- Hadoop平台适配
  - 完成Java、Hadoop、Spark等软件栈的适配与移植
  - 实现应用感知的自动伸缩和统一管理



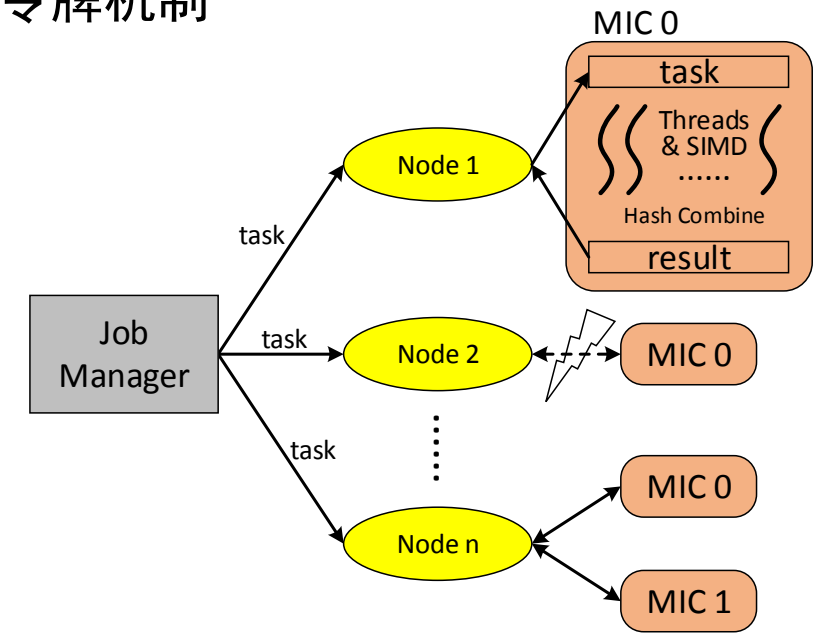
基于Container的大数据处理平台

- 主要措施与优化
  - 基于nova-docker通过sahara实现Hadoop、HDBS、Hive等大数据组件的自动化安装部署，后端引擎支持Direct和Heat-engine机制



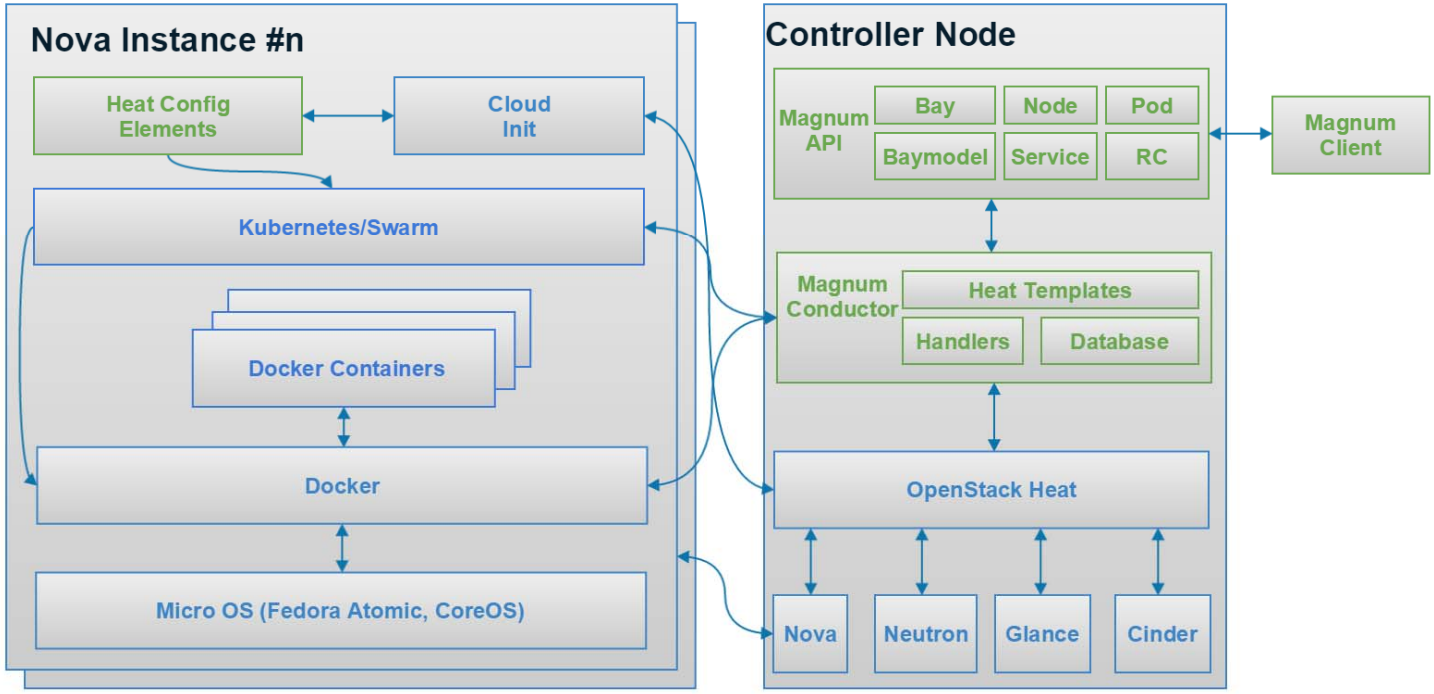
基于Container的大数据处理平台

- 主要措施与优化
  - 异构的Map/Reduce框架支持
    - 支持CPU-MIC异构集群的Map/Reduce框架
    - 动态高效任务调度
    - 支持故障处理与监控的MIC令牌机制
    - vCPUs加速的SIMD映射
    - 多缓冲区的MIC内存管理
  - SIMD哈希算法
    - 消除数据的相关性
  - 异步任务传输
    - 数据传输时间的重叠



KylinCloud + Container

- Magnum结合
  - 通过使用Magnum启动应用容器集群，为用户提供可靠的多租户容器服务



内容

国产飞腾平台支持

- 平台环境
  - 硬件平台：FT1500A
  - 操作系统/内核：Kylin v4
  - 软件平台：

    KylinCloud+Docker+Hadoop/Spark

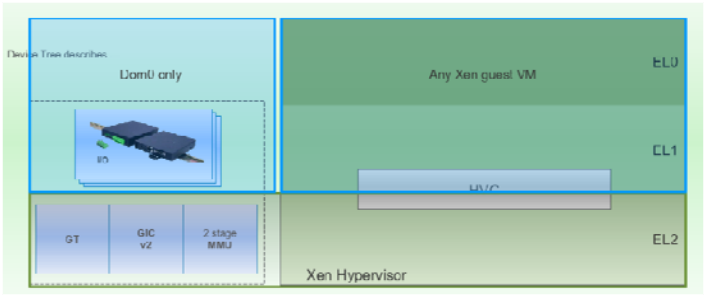
是目前国内唯一的全栈国产化云平台解决方案提供商

国产飞腾平台支持

- 适配工作
  - LXC/Docker@FT适配
    - 操作系统内核适配
    - LXC/Docker适配
    - 定制容器管理工具
    - VNC桌面支持
    - 容器内典型应用、数据库、中间件的适配

国产飞腾平台支持

- 适配工作
  - KVM@FT适配
    - Qemu支持网络类型：User Networking、Socket、Tap、VDE
    - ARM64 UEFI启动
    - KVM加速支持
  - XEN@FT适配
    - 支持硬件加速技术的虚拟机监控器
      - CPU虚拟化、内存虚拟化、I/O虚拟化
    - 物理内存理论最大可达256TB
    - 客户操作系统虚拟处理器数最大可达128
    - 客户操作系统性能折损低于5%
    - 兼容32位模式的客户操作系统



国产飞腾平台应用情况

- 支持计算密集型应用
  - 仿真、模拟等计算开销较大的应用
  - 通过轻量级虚拟化LXC/Docker实现
  - 支持对MIC/GPU/THNI硬件的直接访问，降低开销

国产飞腾平台应用情况

- 支持数据处理类应用
  - 包含了基于Hadoop/Spark的大数据处理平台
  - 底层支持轻量级虚拟化LXC/Docker
  - 支持集群规模的动态伸缩
- 部署案例
  - 某
    - 全FT-1500A系统
    - 资料大数据处理平台
      - 大数据处理框架的适配与优化

国产飞腾平台应用情况

- 支持信息系统类应用
  - 通过XEN/KVM支持现有系统的全虚拟化
  - 支持中间件、办公软件的预先部署
  - 支持信息系统的高可用和负载均衡
- 部署案例
  - 某
    - 支撑事务处理类和数据处理类典型应用
    - FT-1500A节点与X86混合调度
    - Hadoop软件栈

yusong.tan@kylin-cloud.com
li.wang@kylin-cloud.com


CCTC
2016 中国云计算技术大会
Cloud Computing Technology Conference 2016
CSDN
Thanks!