通用高可用云技术：COLO 从原型到产品

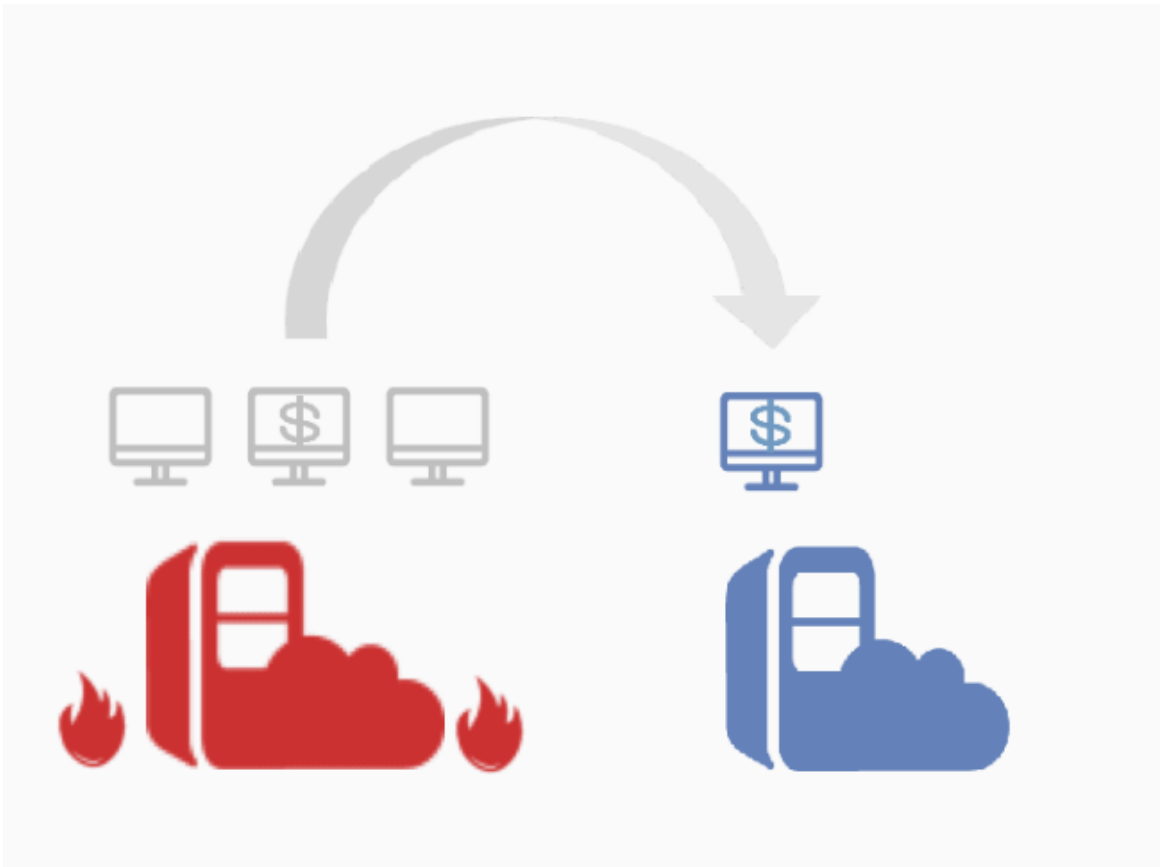
董耀祖，英特尔亚太研发有限公司资深首席工程师

Eddie Dong (eddie.dong@intel.com)

10/25/2020

Agenda

- 高可用技术介绍
- COLO架构
- 通用高可用云
- COLO 产品化



云计算和高可用

- 云计算是各种资源服和务的池化以及拆零贩售
  - 计算资源，存储资源，软件服务，平台服务。。。
- 任何资源都可能失效：计算节点，存储介质, 但是
  - 客户要的是数据永不丢失，服务永远可用
- 计算资源的高可用是基础

计算资源的高可用

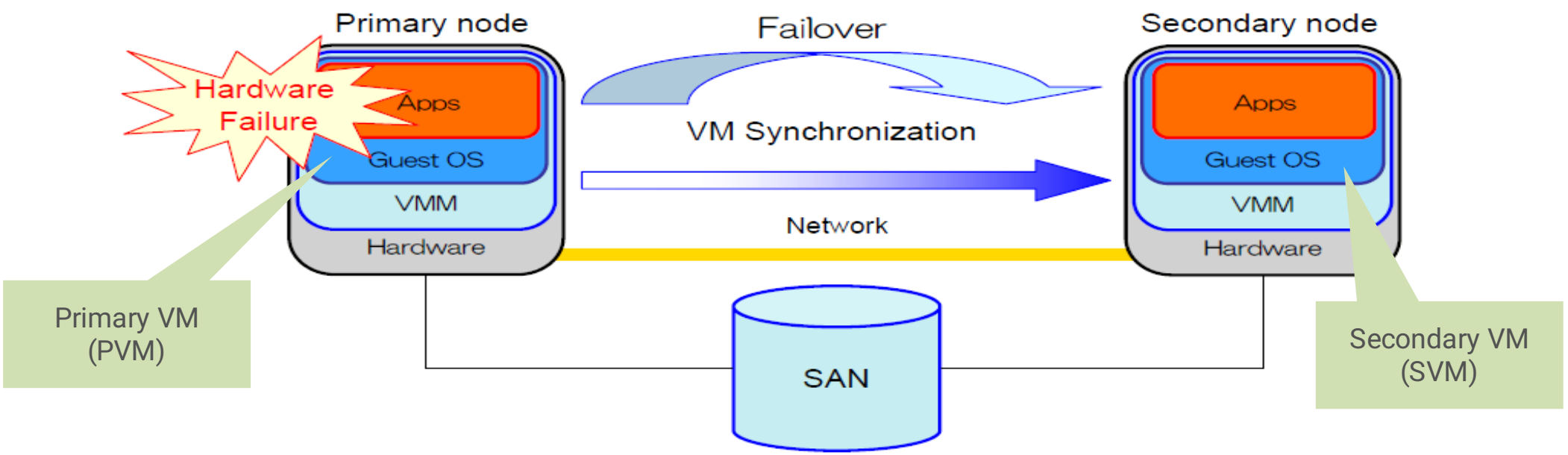
- Hardware Solution
  - Robust Hardware Components
    - Very expensive
    - But still Unpredictable, such as alpha-particles in cosmic ray

- Software Solution
  - Replication/热备份: Construct backup replicas at run time
  - Failover: Backups take over when the primary fails

不同层面的热备份方法

- 应用层面
  - Extensive software customization
  - Impractical to legacy software
- 操作系统层面
  - Large complexity, not commoditized
- 虚拟机层面
  - Application and OS agnostic

虚拟机热备份


Primary node
Failover
Secondary node
Hardware Failure
Apps
Guest OS
VMM
Hardware
VM Synchronization
Network
Apps
Guest OS
VMM
Hardware
SAN
Primary VM (PVM)
Secondary VM (SVM)

传统的虚拟机热备份方法

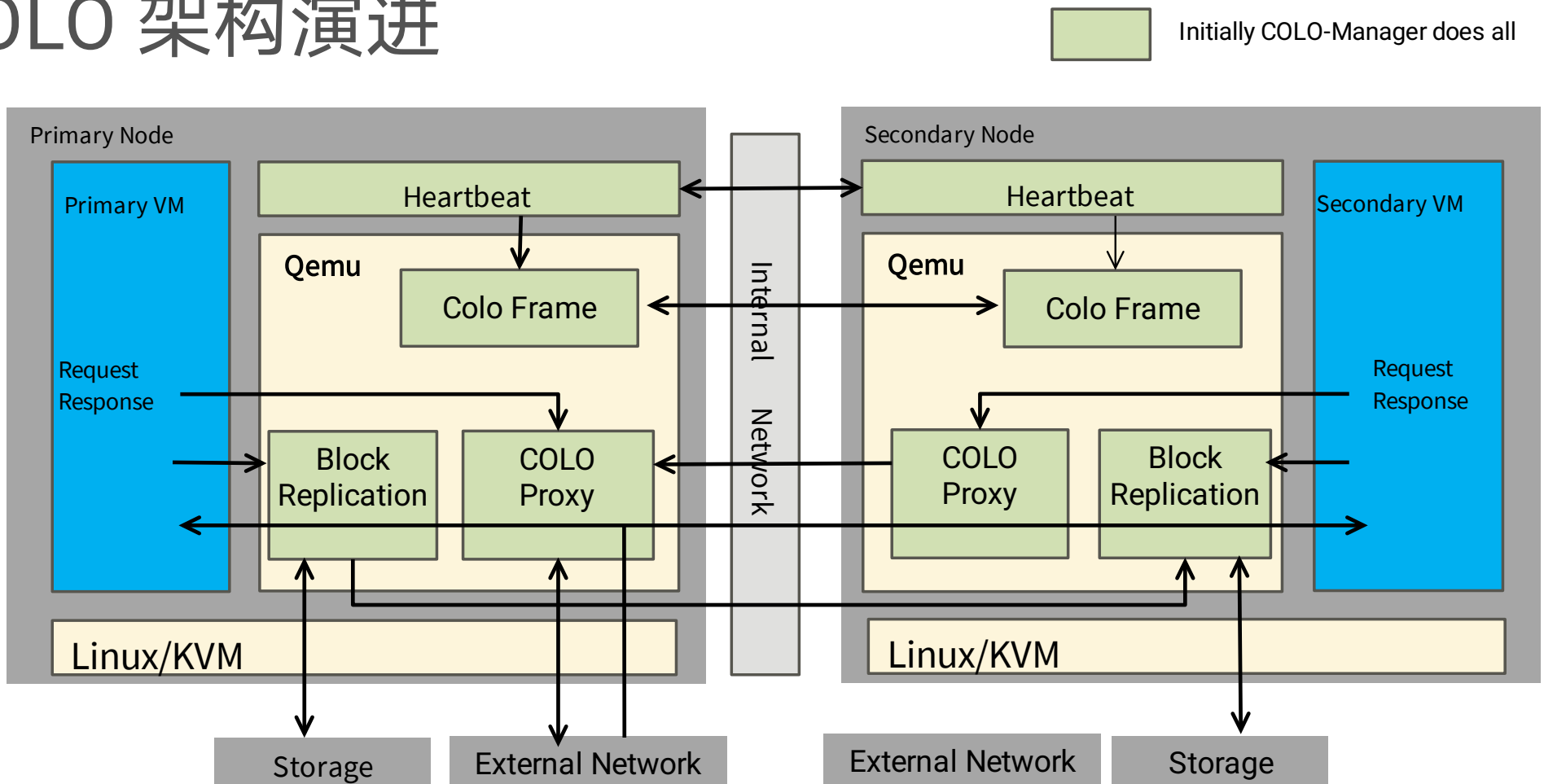
- Lock-stepping/锁步: Replicating per instruction
  - Execute in parallel for deterministic instructions
  - Lock and step for nondeterministic instructions
  - **Exact machine state matching at any instruction boundary**

- Checkpoint/同步: Replicating per epoch
  - Output is buffered within an epoch
  - **Exact machine state matching from external observers**

COLO: COarse-grain LOck-stepping (COLO) VMs for Non-stop Service – 粗粒度锁步虚拟机实现不间断服务

- 当且仅当备份虚拟机的输出和主虚拟机不同的时候，才进行主备虚拟机间的状态同步
  - 减少状态同步的频率和提高虚拟机响应的速度
  - 详见SOCC论文： http://www.socc2013.org/home/program or http://soft.cs.tsinghua.edu.cn/os2atc2014/ppt/keynote/keynote-dongyaozu.pdf

COLO 架构演进


Initially COLO-Manager does all
Primary Node
Primary VM
Heartbeat
Qemu
Colo Frame
Request
Response
Block
Replication
COLO
Proxy
Linux/KVM
Storage
External Network
Internal Network
Secondary Node
Heartbeat
Qemu
Colo Frame
Secondary VM
Request
Response
COLO
Proxy
Block
Replication
Linux/KVM
External Network
Storage

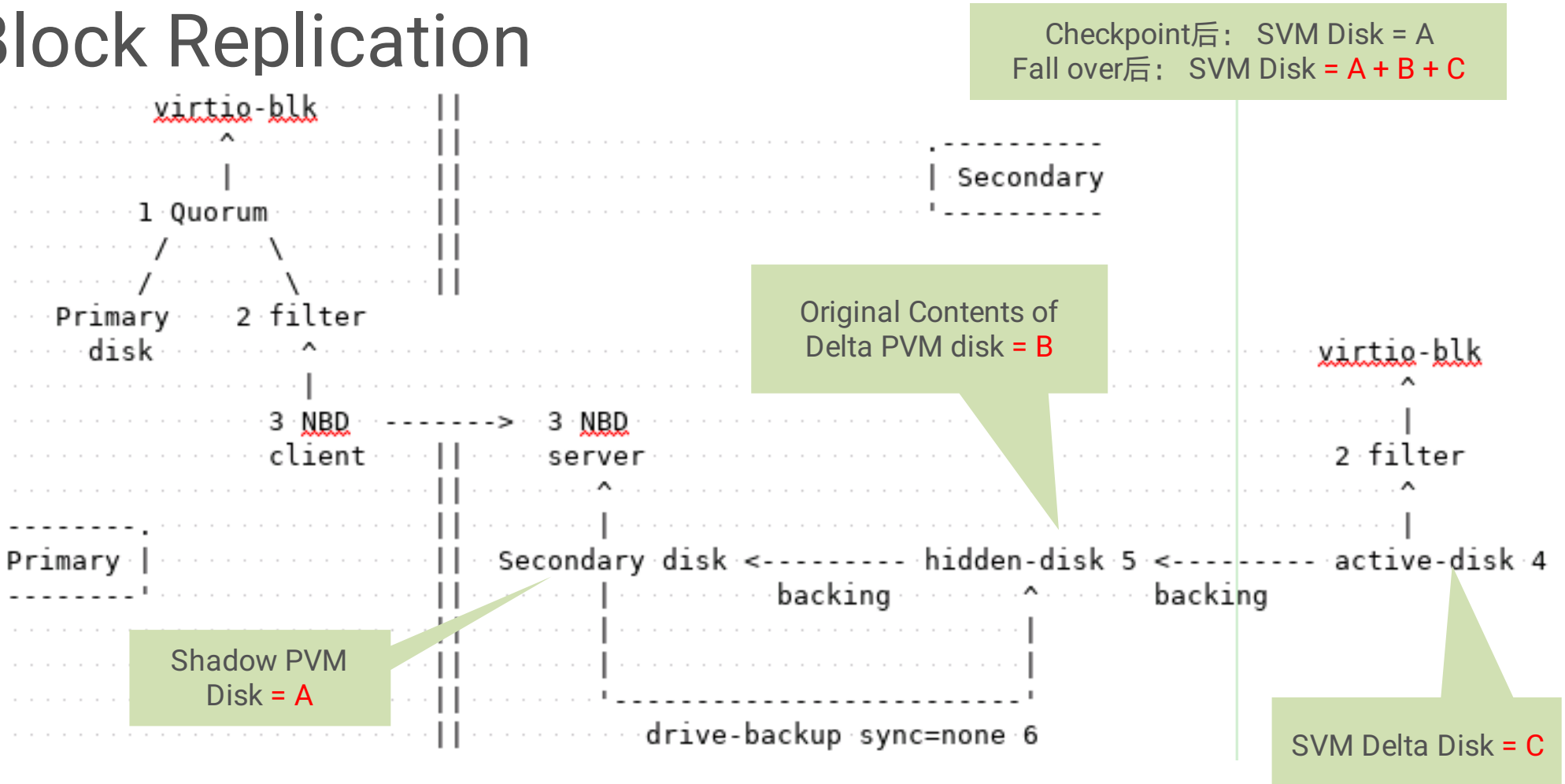
COLO Frame

- On-demand VM checkpoint (PVM & SVM differs)
  - Reuse the Qemu process of live VM migration.
  - Update SVM state with checkpointed PVM state
- Timeout VM checkpoint
  - Throttle the max duration of differentiation
  - To limit the possible delta state
- Additional Optimization
  - May spontaneously transfer dirty pages in advance

COLO Proxy

- Forward/Mirroring/Re-direct（PVM和SVM获得相同的Input）
  - PVM端将输入包Mirroring到SVM端, SVM端将收到的包Injecting SVM
  - SVM端将需要比较的输出包转发到PVM端
- Compare/比较（以确定SVM是否合法备机）
  - 包过滤：如忽略 UDP包
  - 延迟比较：主备机生成的响应可能有一定时间的延迟
- COLO实现了一个通用netfilter的机制，以方便实现各种对网络包的操作

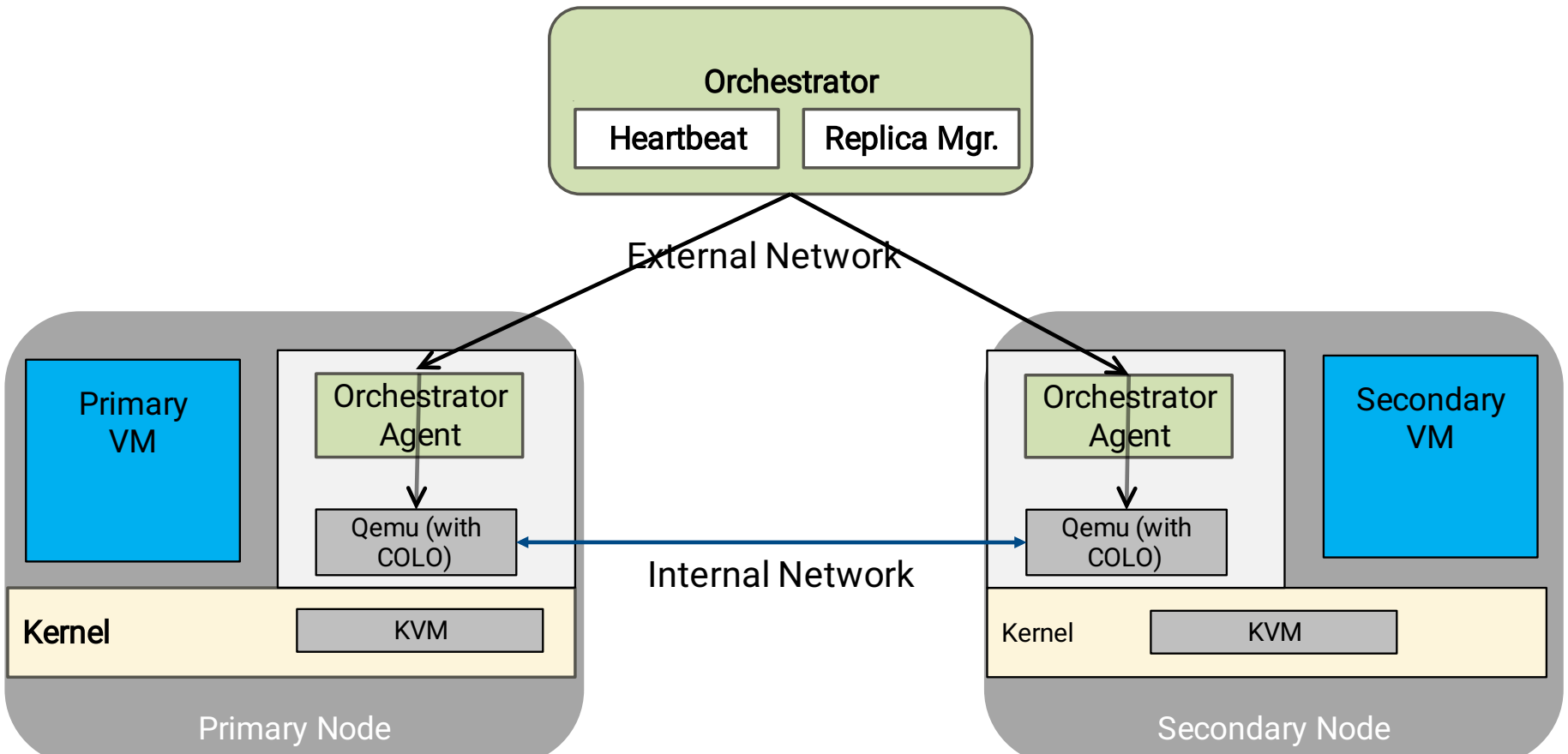
Block Replication



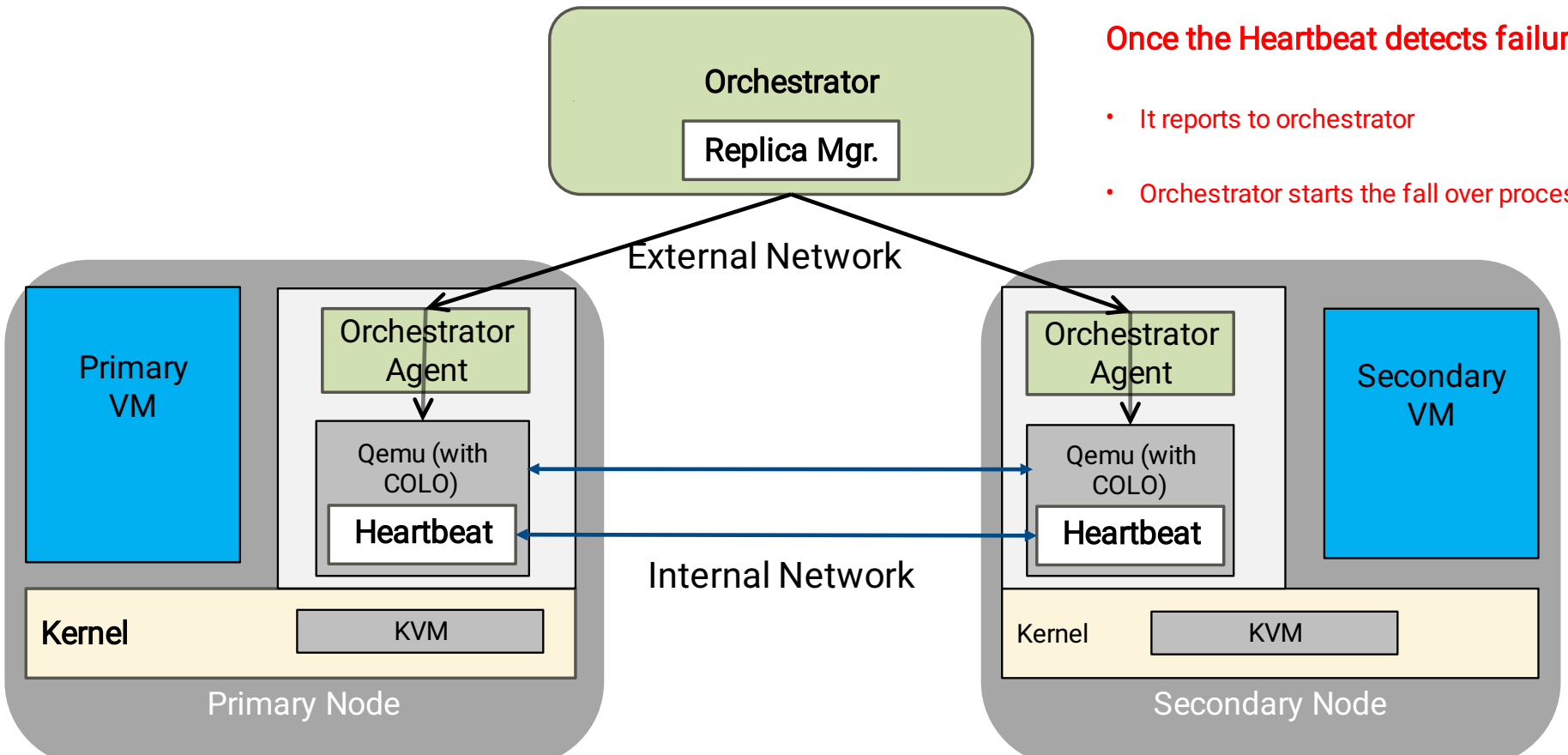
Network QoS

- COLO uses network heavily, and some of the connections are latency sensitive
  - COLO proxy connection is latency sensitive, similar for block replication connection
  - Heartbeat connection may malfunction if the latency exceeds threshold
  - VM Migration and guest use of network may saturate the bandwidth
- Use priority-based TC to grant the high priority connections, and throttle the bandwidth of each connection at same priority

通用高可用云方式 1 – 外部心跳检测



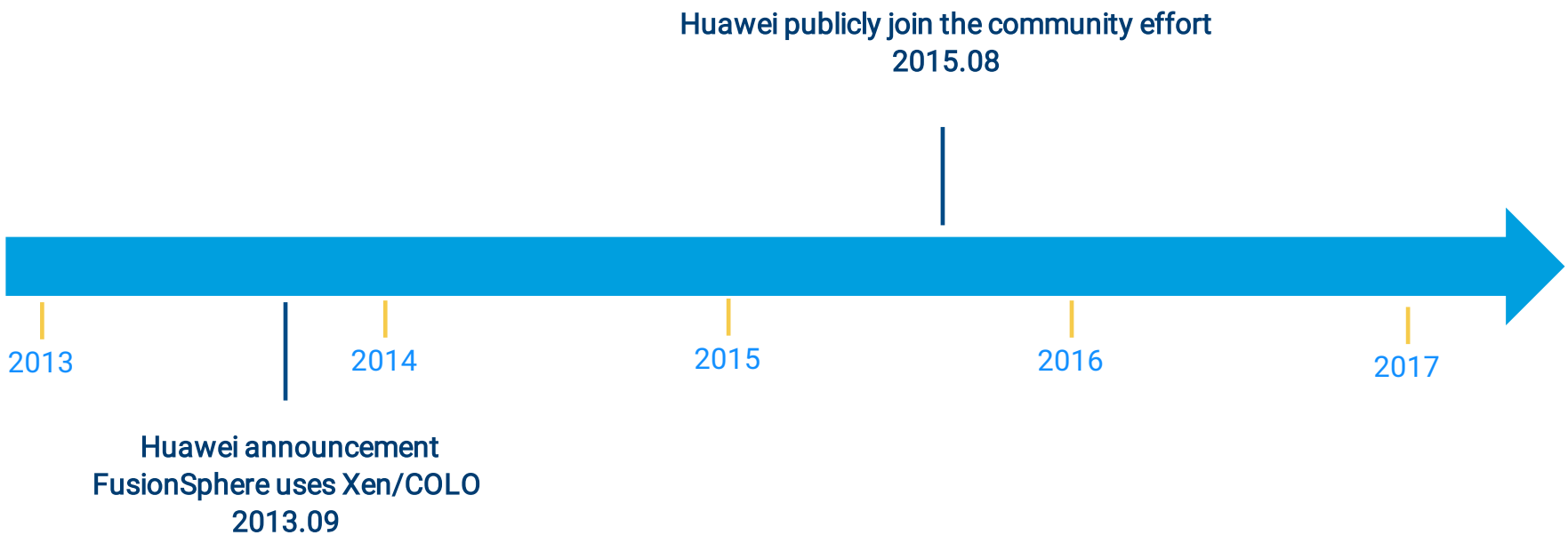
通用高可用云方式2 – 内部心跳检测



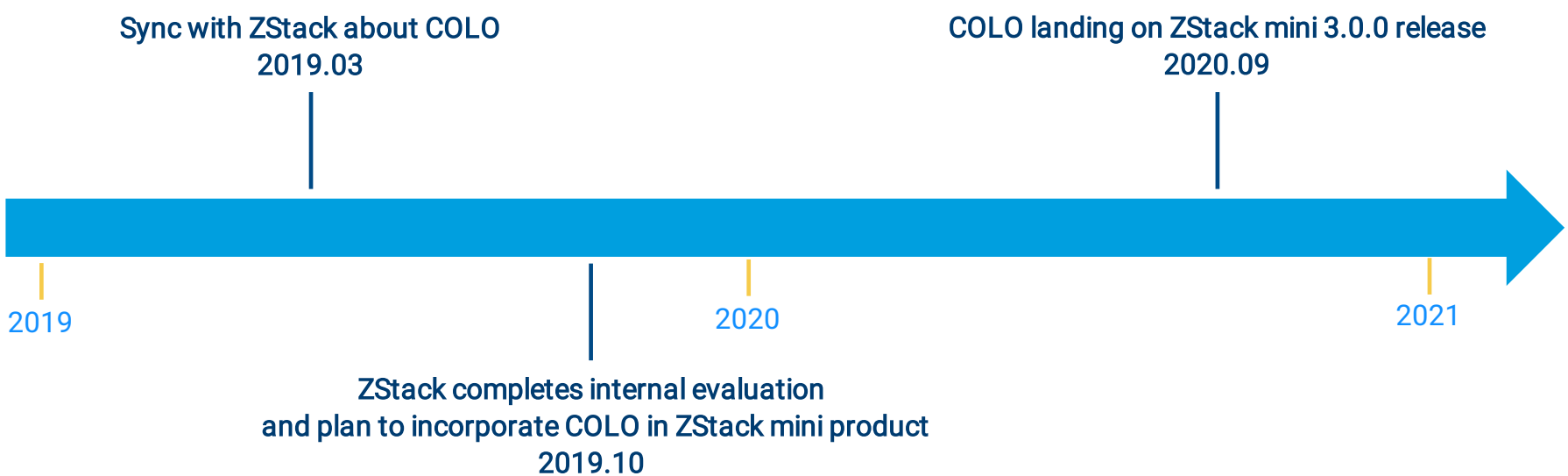
Once the Heartbeat detects failure

- It reports to orchestrator
- Orchestrator starts the fall over process

产品化演进Case1: 华为FusionSphere



产品化演进Case2: 云轴科技ZStack



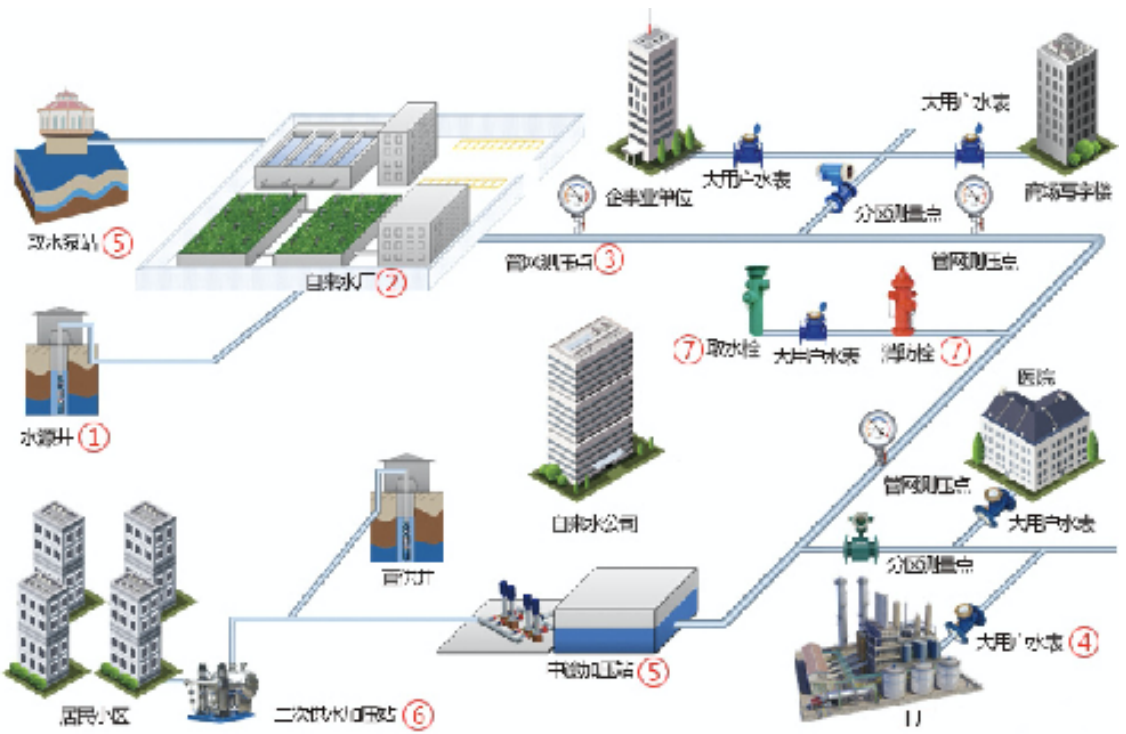
应用场景1 – 公共服务

某省水务集团现场应用系统改造

- 水务系统中的水厂、加压站、管网测压点、分区测压点、自来水公司等链条环节都需要IT应用参与。
- 现场人员的IT运维能力较弱，且现场应用程序往往是单机形式运行，一旦宕机严重将导致区域无法供水。

COLO应用与收益

- 采用界面全可视化操作的双节点进行虚拟化部署，存储通过高速网络互联同步，同时COLO负责两个节点上的虚拟机同步。
- 通过虚拟化FT技术，现场单机应用无需任何改造即可达到连续可用性，大大降低供水中断风险。

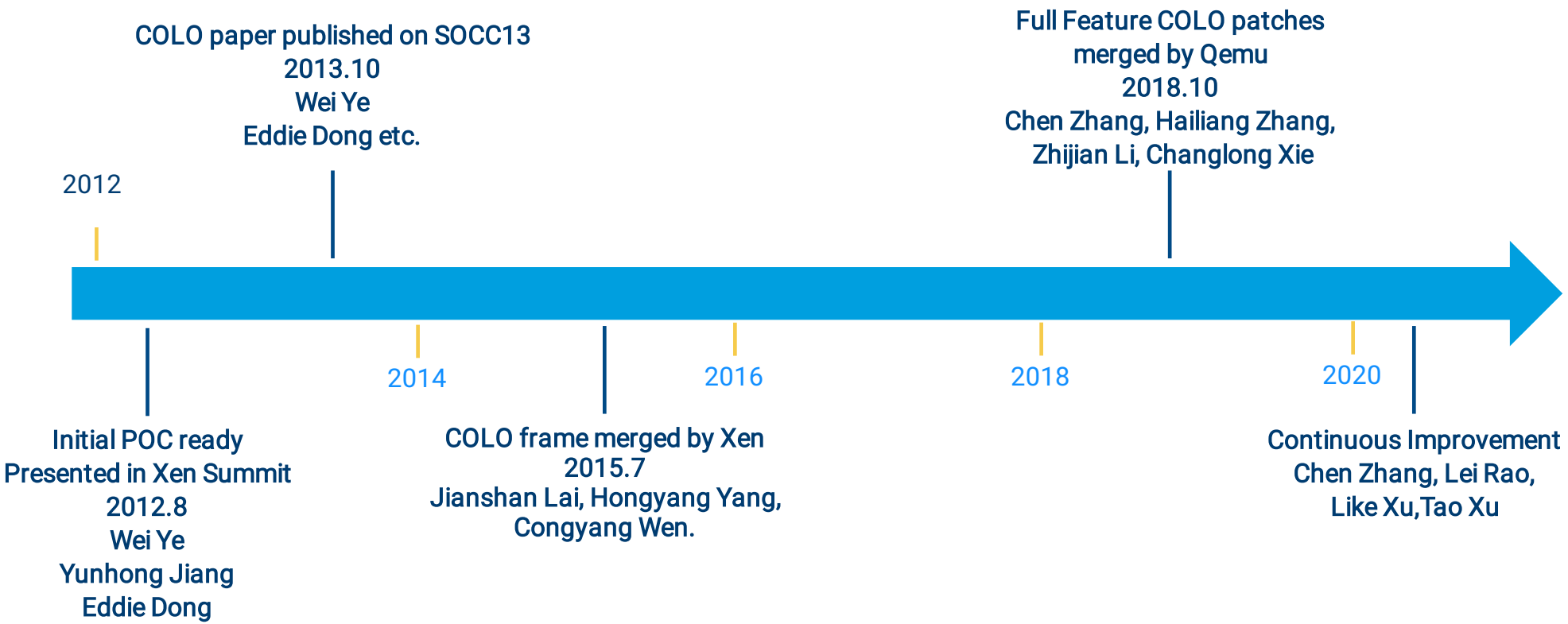


应用场景2 – 生产制造

某大型焦炭、焦化产品、精细化工生产供应商

- 集团总部采用某云平台部署诸如ERP、MES、财务等应用，作业现场部署同品牌边缘虚拟化产品，并采用COLO作为虚拟化容错技术，承载DCS、PLC、SCADA等自动化控制与监测应用，集团总部可进行基础设施的统一管理。
- COLO容错技术因为采用虚拟化形式进行部署，避免了应用操作系统对硬件的兼容适配等问题，同时老旧应用系统可以采用P2V形式进行快速地虚拟化改造，COLO容错技术使得现场控制系统的可靠性大大提高，保证生产安全以及连续性。

COLO的开发历史和开发者



后续研发

- Libvirt support
  - Incorporate with cloud Oses by default
- Combine with RSA capabilities
- Spontaneous memory page check point
  - Reduce the cost of coming VM checkpoint
- Take the advantage of latest technology such as NVDIMM
- Continuous improvement

欢迎大家加入COLO社区：试用 .. 使用 .. 开发