PostgreSQL中文社区

2021 PostgreSQL China Conference

主办：PostgreSQL 中文社区

第 11 届 PostgreSQL 中国技术大会

开源论道 × 数据驱动 × 共建数字化未来

腾讯云PostgreSQL生态的技术架构演进

刘少蓉
(shaorongliu@tencent.com)

腾讯云数据库

目录

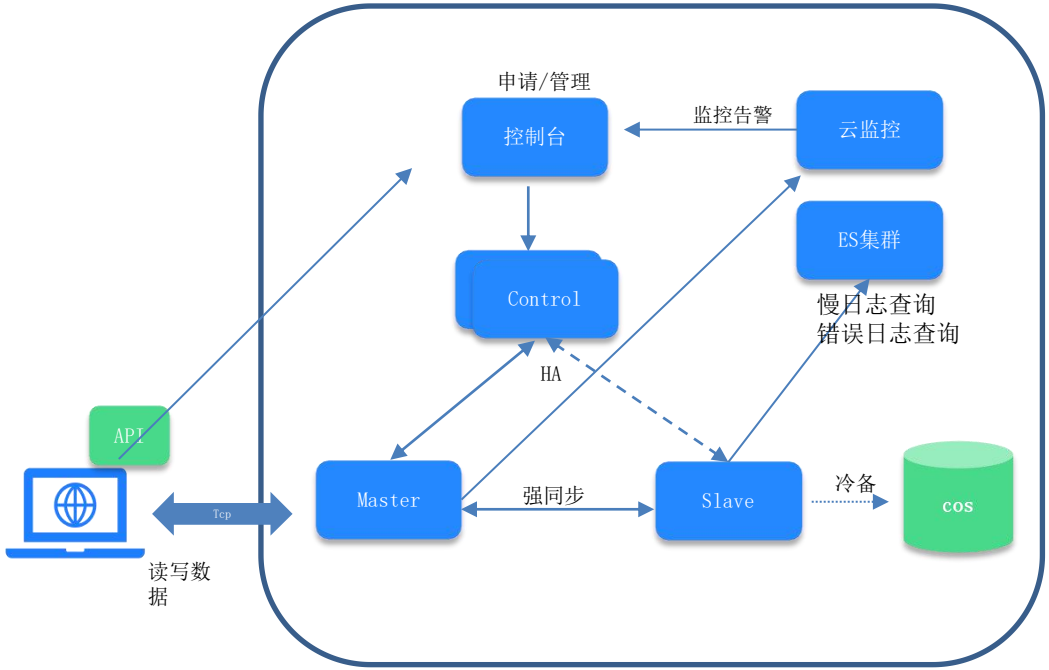
PART 01 - Overview

腾讯云数据库PostgreSQL生态产品

云数据库	分布式HTAP	分布式分析型	云原生
			
TencentDB for PG	TDSQL PG	TDSQL-A	TDSQL-C PG

PART 02 - Tencent’s Journey in PG - from Cloud DB to Cloud-native DB

TencentDB for PostgreSQL：开箱即用的云端数据库服务



易于部署和管理

- 一键创建，丰富可选的软硬件配置
- 轻松管理，完善的自助运维管理工具

高可用

- 服务高可用：一主一从强同步架构
- 备份高可用：cos备份

开放与服务集成

- 多种云产品集成联通
- 控制台、SDK、API等丰富的接入方式

多种插件支持

- 全面集成高级商业特性：安全、告警、SQL、机器学习等

TencentDB for PostgreSQL的优势

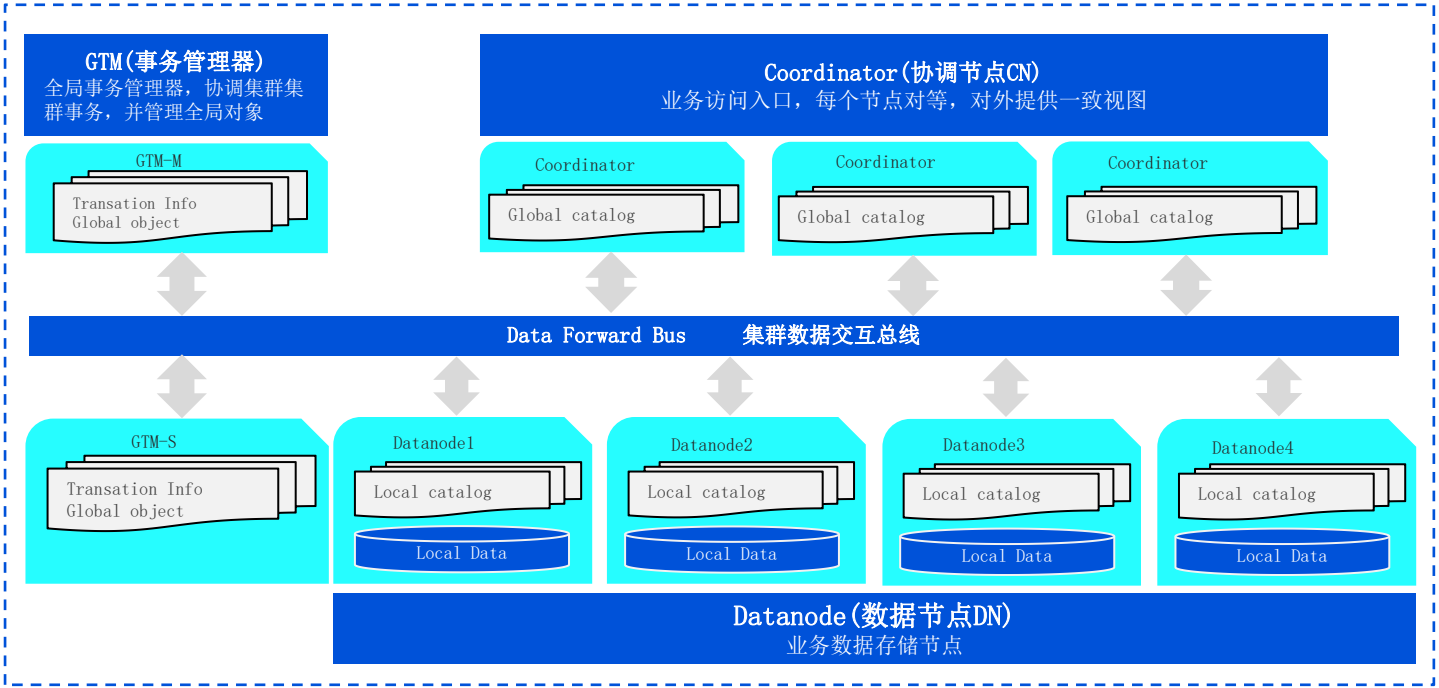
- 场景化的内核优化
  - e.g.，通过异步DDL方式实现主备延迟优化
- 透明数据加密 （TDE）
  - 借助云上 KMS 的能力，实现了 TDE 的功能，提高了安全性
- 丰富的插件，e.g.，
  - timescaledb，时序数据库
  - pipelinedb，流式计算
  - rdkit，针对化学类场景
  - zhparser，中文分词

从单机数据库到分布式数据库

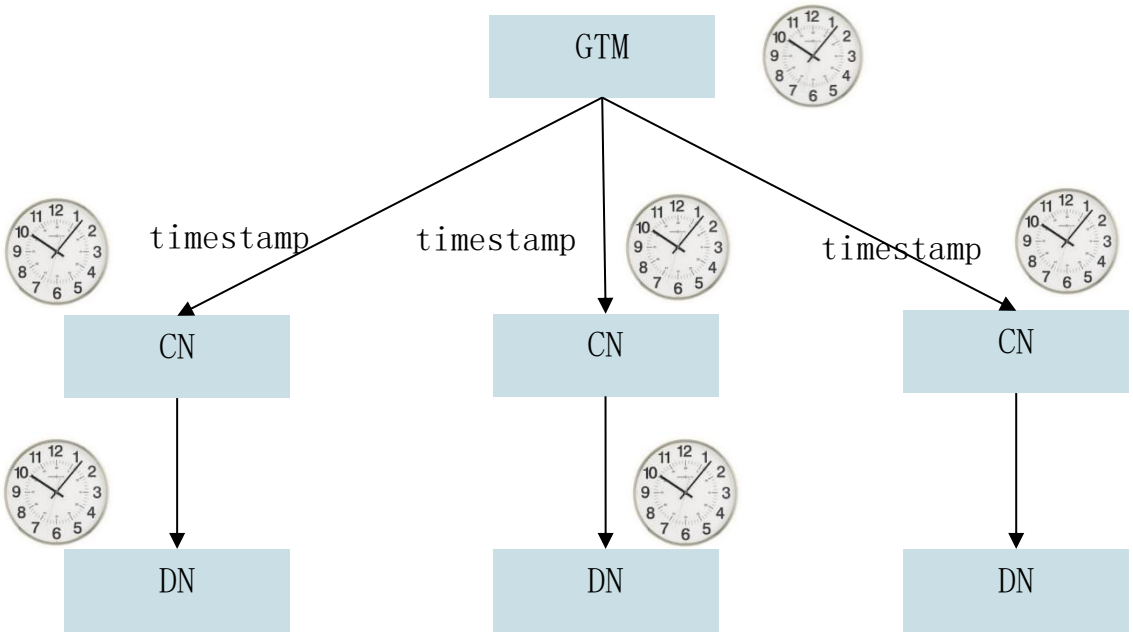
- TencentDB for PG 给一定数量下的业务提供了All In One的数据解决方案
- 当业务不断扩大， 数据量超过单机的limit时， 我们有什么解决方案呢？
  - 传统方法 -分库分表
    - 把一张逻辑表拆分很多物理表
    - 业务需要实现复杂的分布式逻辑 （e.g., 事务， 跨表查询）
  - 分布式MPP架构
    - 水平扩展 (scale out)
    - 把复杂的分布式逻辑留给数据库解决
    - 业务逻辑简单

TDSQL PG 架构

- 指标监控
- 运维管理
- 实时告警
- 安全审计
- 数据治理



TDSQL PG 高效可扩展的分布式事务设计



01 全局时钟同步

GTM提供全局统一时间戳，进行全局事务一致性同步

02 CN，DN向GTM请求时间戳

CN，DN的事务向GTM请求全局时间戳作为事务的版本标识

03 GTM单点可靠性问题

多个GTM节点构成集群，主节点对外提供服务；主备之间通过日志同步时间戳状态，保证GTM核心服务可靠性。

04 高性能可扩展分布式事务

通过专利技术提供高性能可扩展的分布式事务能力

TDSQL PG（代号TBase）开源

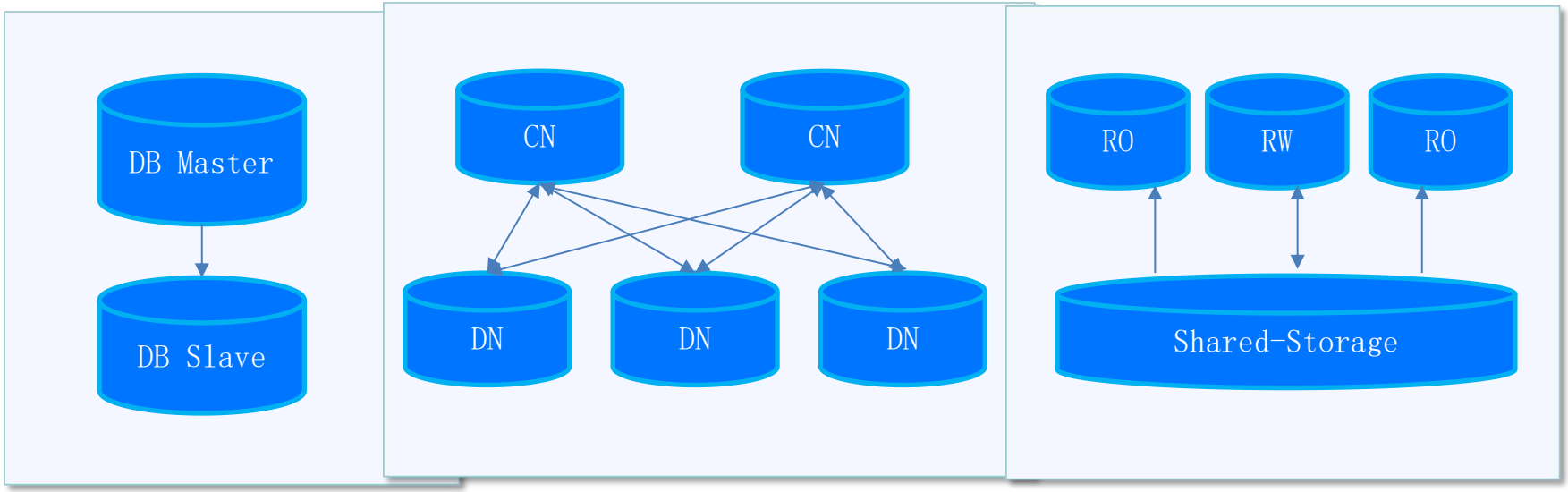
2019.11.07腾讯正式宣布TDSQL-PG开源

- 开源地址：
  - https://github.com/Tencent/TBase
  - https://github.com/Tencent/TBase/wiki
- 版本升级
  - 2020/7 V2.1.0版本 - 多活能力更上一层楼
  - 2021/7 V2.2.0版本 - 性能再提升白倍
  - 2022/1 V2.3.0版本
    - 分区表能力增强
    - 易用性重磅升级

TDSQL-A - 分布式分析型数据库

- TDSQL-A是腾讯自研的一款分布式分析型数据库
  - 基于TDSQL PG的架构上演进， 从架构上全面优化分析性能
  - 100% PostgreSQL 兼容，高度兼容Oracle
- 核心技术
  - 行列混存
  - 列式数据的多级压缩 （透明，轻量）
  - 向量化执行引擎
  - 多种并行执行策略 （MPP， SMP + SIMD）
  - 基于代价的优化引擎(CBO)
  - 分布式延迟物化的优化 - 减少不必要的网络开销

从云数据库到云原生数据库

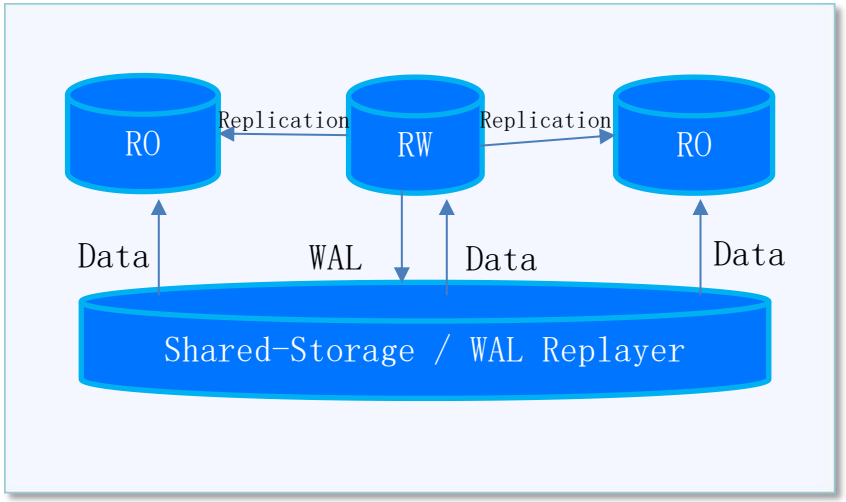


单机数据库

无共享MPP
DN - compute + storage

云原生
计算存储分离（Aurora SIGMOD 2017），按需弹性扩展
资源池化，降本增效

云原生数据库:日志即数据库



日志即数据库 - 减少网络开销， 提高写性能

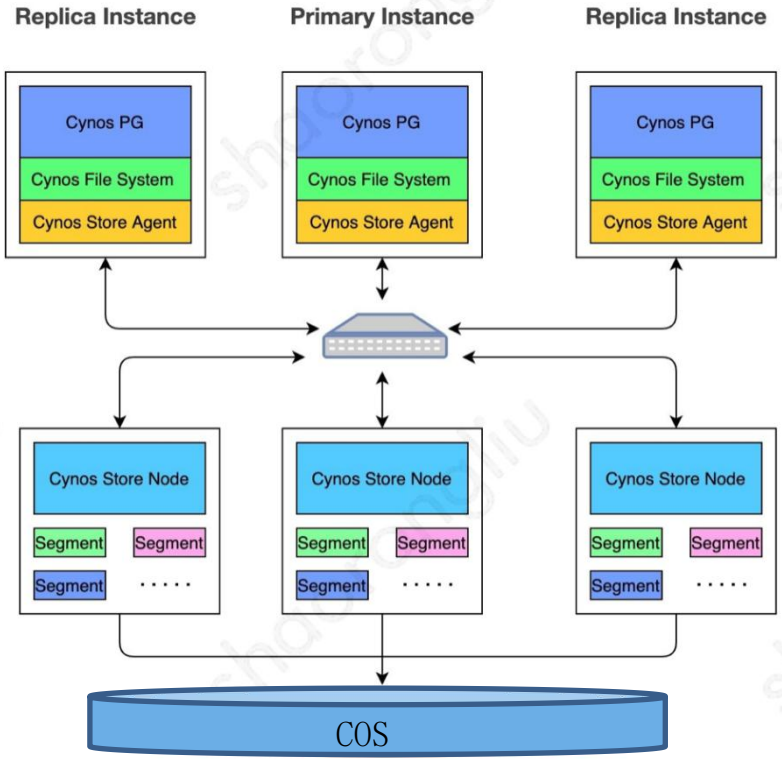
设计思想

- RW 和 RO 基于一份数据，放在共享存储
- RW仅将WAL写入共享存储、不写Page
- 共享存储通过重放WAL，实现存储节点上 Page 页的修改
- 存储层以 Page 为单位维护数据
- RO 从RW接收 WAL，并在缓存中重放，保持缓存中 Page 持续更新

TDSQL-C PG - 腾讯云原生分布式数据库

- 完全自研的云原生数据库
- 基于计算存储分离，日志即数据库的设计思想
- 融合传统数据库，云计算，新硬件技术的优势
- 高可用，高可靠,高性能，极致弹性
- 快速恢复:支持基于快照的秒级备份和回档
- 无锁化设计:减少内核切换
- 100% PG兼容， 高度兼容Oracle

TDSQL-C PG架构



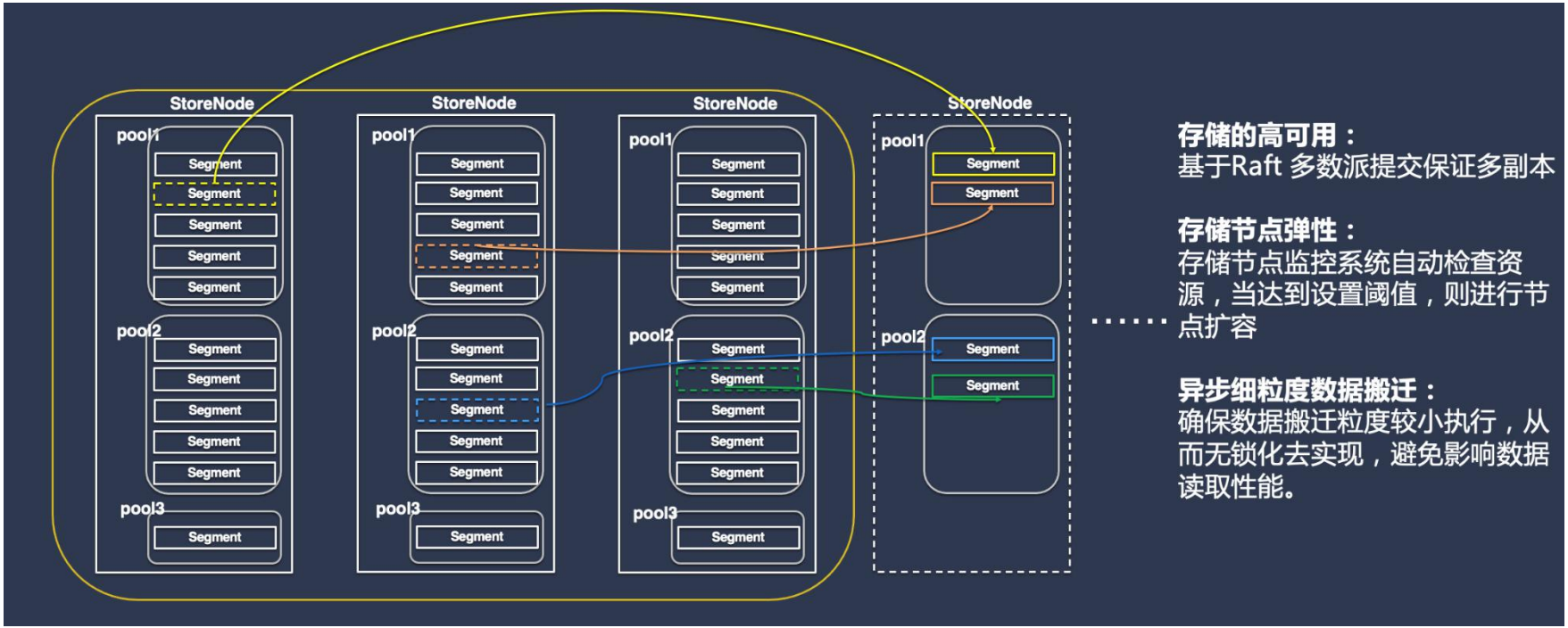
计算层

- CynosPG – 基于PG研发的计算引擎
  - 查询处理器， 事务管理，缓存实现，锁，MVCC
  - ~~Full page write~~
  - ~~脏页刷盘~~
- Cynos File System
  - 用户态文件系统， 主要提供分布式文件管理
- Cynos Store Agent
  - 计算存储之间的读写交互
  - 主备之间的日志流同步

存储层

- WAL日志记录， 日志回放，持久存储，
- 备份/恢复， CRC

TDSQL-C PG - 分布式存储

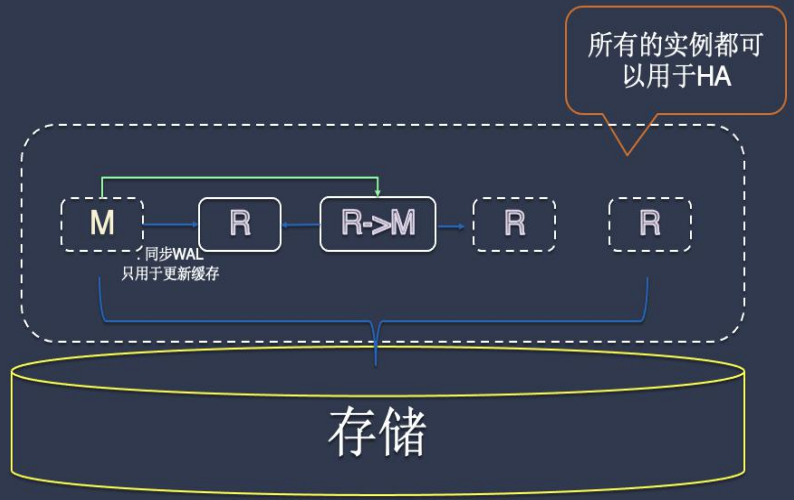


存储的高可用：
基于Raft 多数派提交保证多副本

存储节点弹性：
存储节点监控系统自动检查资源，当达到设置阈值，则进行节点扩容

异步细粒度数据搬迁：
确保数据搬迁粒度较小执行，从而无锁化去实现，避免影响数据读取性能。

TDSQL-C PG - 可调度的计算节点



高可用：

- 得益于存算分离后，无需搬迁数据，而计算节点的无状态。或者（为了可用性）而舍弃的状态保证在秒级拉起计算节点
- HA切换速度上，计算实例间仅同步少量缓存信息，所以可通过快速的备提主以及重建备机来达到更高的可用性

弹性:
容器化般的计算节点 可自由添加不同规格的计算实例，来回切换也可轻松实现。

PART 03 - What’s next?

未来展望

完善生态

软硬一体

架构演进

更智能化

THANKS

谢谢观看