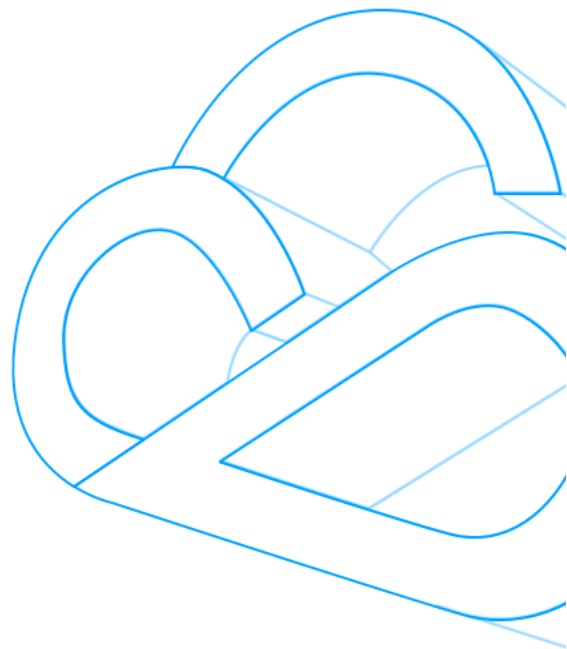




腾讯云数据库 TDSQL MySQL 版 分布式 V10.3.22.8/集中式 V8.0.22.8 部署规划



文档版本:

发布日期:

腾讯云计算（北京）有限责任公司

版权声明

本文档著作权归腾讯云计算（北京）有限责任公司（以下简称“腾讯云”）单独所有，未经腾讯云事先书面许可，任何主体不得以任何方式或理由使用本文档，包括但不限于复制、修改、传播、公开、剽窃全部或部分本文档内容。

本文档及其所含内容均属腾讯云内部资料，并且仅供腾讯云指定的主体查看。如果您非经腾讯云授权而获得本文档的全部或部分内容，敬请予以删除，切勿以复制、披露、传播等任何方式使用本文档或其任何内容，亦请切勿依本文档或其任何内容而采取任何行动。

商标声明



“腾讯”、“腾讯云”及其它腾讯云服务相关的商标、标识等均为腾讯云及其关联公司各自所有。若本文档涉及第三方主体的商标，则应依法由其权利人所有。

免责声明

本文档旨在向客户介绍本文档撰写时，腾讯云相关产品、服务的当时的整体概况，部分产品或服务在后续可能因技术调整或项目设计等任何原因，导致其服务内容、标准等有所调整。因此，本文档仅供参考，腾讯云不对其准确性、适用性或完整性等做任何保证。您所购买、使用的腾讯云产品、服务的种类、内容、服务标准等，应以您和腾讯云之间签署的合同约定为准，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

修订记录

文档版本	发布日期	修订人	修订内容
------	------	-----	------

目录

- 修订记录.....ii
- 目录.....iii
- 前言.....iv
- 1 部署规划 LLD 1
- 2 资源规划..... 2
 - 2.1 简单环境部署（体验环境） 2
 - 2.1.1 概述..... 2
 - 2.1.2 硬件规划..... 2
 - 2.1.3 软件规划..... 3
 - 2.2 正式部署规划（生产环境） 6
 - 2.2.1 硬件规划..... 6
 - 2.2.2 软件规划..... 12
 - 2.3 分布式核心系统数据库资源评估..... 13
 - 2.3.1 分布式实例(GroupShard)资源评估方法 14
 - 2.3.2 非分布式实例(No-Shard)资源评估方法 17
 - 2.3.3 数据库部署方案..... 18
 - 2.3.4 架构模型及服务器资源预估..... 20
- 3 端口规划..... 24
- 4 网络规划..... 27

前言

文档目的

本文档用于帮助用户掌握云产品的操作方法与注意事项。





目标读者

本文档主要适用于如下对象群体：

- 客户
- 交付 PM
- 交付技术架构师
- 交付工程师
- 产品交付架构师
- 研发工程师
- 运维工程师

符号约定

本文档中可能采用的符号约定如下：

符号	说明
 说明：	表示是正文的附加信息，是对正文的强调和补充。
 注意：	表示有低度的潜在风险，主要是用户必读或较关键信息，若用户忽略注意消息，可能会因误操作而带来一定的不良后果或者无法成功操作。
 警告：	表示有中度的潜在风险，例如用户应注意的高危操作，如果忽视这些文本，可能导致设备损坏、数据丢失、设备性能降低或不可预知的结果。
 禁止：	表示有高度潜在危险，例如用户应注意的禁用操作，如果不能避免，会导致系统崩溃、数据丢失且无法修复等严重问题。

1 部署规划 LLD

附件：[腾讯云数据库 TDSQL10.3.22.x 部署计划模板 LLD V3.2.xlsx](#)

2 资源规划

2.1 简单环境部署（体验环境）

2.1.1 概述

如果您部署的目标是用于其他产品的研发、测试基座，产品功能体验，请参考本章节要求部署即可。

由于体验的范围不同，体验环境部署也分为：

- 基础能力体验环境：仅体验数据库最基础的开发、运维、可用性能力。
- 完整能力体验环境：体验本产品全部功能，但不用于生产环境、性能压测、高可用测试等。

警告：

部分用户对数据库的高可用、性能、完整功能并无强制要求，会在生产环境参考该部署方案，请务必谨慎评估。

2.1.2 硬件规划

基础能力体验环境

基础能力体验所需环境相对简单，请参考以下配置分配好设备：

硬件环境要求	物理机、虚拟机均可
设备台数	3 台
CPU 配置	8 核以上
内存配置	16GB 以上
磁盘配置	500GB 以上
网卡配置	不限
操作系统	Centos 7.6、7.8、7.9

机房部署要求	不限
--------	----

说明:

测试环境只做功能验证，可以使用虚拟机，CPU 不低于 8 核，内存不低于 16G，磁盘不低于 500G。

完整能力体验环境

完整能力体验所需环境相对简单，请参考以下配置分配好设备：

硬件环境要求	物理机、虚拟机均可
设备台数	8 台
CPU 配置	8 核以上
内存配置	16GB 以上
磁盘配置	500GB 以上
网卡配置	千兆或以上
操作系统	Centos 7.6、7.8、7.9
机房部署要求	不限

2.1.3 软件规划

基础能力体验部署组件

子系统名称	实际涉及软件模块	功能说明	本次部署是否包含	开源/商用说明	规划设备台数
运营管理系统	Chitu Zookeeper Monitor Scheduler Manager OSS CloudDBA	提供数据运维管理、高可用调度、监控告警等功能。	■是 □否	Zookeeper 可以选用商用版本，或可对应开源版本。	1

数据库核心	DataBase Proxy Agent Oc_agent	提供数据库核心计算、数据存储等能力	■是 □否		2
数据库负载均衡	LVS	提供数据库实例唯一接入 IP、并实现分布式场景下负载均衡能力	□是 ■否	可以选用商用负载均衡、商用 DNS、腾讯商用 CLB、VPC；LVS 可以选用商用版本，或可对应开源版本。	不涉及
备份存储集群	HDFS	长期存储数据库备份、日志等文件；用于数据灾难恢复，从机重建等场景	□是 ■否	可以选用商用 HDFS，商用 NAS，商用对象存储；HDFS 可以选用商用版本，或可对应开源版本。磁带库可后置于 HDFS 后。	不涉及
全局分布式事务调度服务	Metacluster GTS	提供分布式事务实时读一致能力。不安装该组件时均提供分布式事务最终一致性特性。	□是 ■否		不涉及
数据订阅与生产	Kafka Consumer	提供数据生产到消息中间件队列中，订阅、	□是 ■否	Kafka 可以选用商用版本，或可对应开源版本。	不涉及
日志存储与查询	ElasticSearch LogStash Kibana	提供长期的日志存储、查询和日志监控能力	□是 ■否	腾讯不提供相关软件安装包，请您自己部署商用或开源版本。	不涉及

完整能力体验部署组件

子系统名称	实际涉及软件模块	功能说明	本次部署是否包含	开源/商用说明	规划设备台数
运营管理系统	Chitu Zookeeper Monitor Scheduler Manager OSS CloudDB	提供数据运维管理、高可用调度、监控告警等功能。	■是 □否	Zookeeper 可以选用商用版本，或可对应开源版本。	3

	A				
数据库核心	DataBase Proxy Agent Oc_agent	提供数据库核心计算、数据存储等能力	■是 □否		3
数据库负载均衡	LVS	提供数据库实例唯一接入 IP、并实现分布式场景下负载均衡能力	■是 □否	可以选用商用负载均衡、商用 DNS、腾讯商用 CLB、VPC；LVS 可以选用商用版本，或可对应开源版本。	2
备份存储集群	HDFS	长期存储数据库备份、日志等文件；用于数据灾难恢复，从机重建等场景	■是 □否	可以选用商用 HDFS，商用 NAS，商用对象存储；HDFS 可以选用商用版本，或可对应开源版本。磁带库可后置于 HDFS 后。	与数据库负载均衡混合部署。
全局分布式事务调度服务	Metacluster GTS	提供分布式事务实时读一致能力。不安装该组件时均提供分布式事务最终一致性特性。	■是 □否		与运营管理系统混合部署。
数据订阅与生产	Kafka Consumer	提供数据生产到消息中间件队列中，订阅、	■是 □否	Kafka 可以选用商用版本，或可对应开源版本。	与运营管理系统混合部署。
日志存储与查询	ElasticSearch LogStash Kibana	提供长期的日志存储、查询和日志监控能力	□是 ■否	腾讯不提供相关软件安装包，请您自己部署商用或开源版本。	不涉及

其他依赖软件（可选）

为提高产品的可靠性，产品还需依赖如下软件，请按需配置：

- NTP 网络时间服务器（不限版本）。
- 操作系统官方 yum 源。

2.2 正式部署规划（生产环境）

2.2.1 硬件规划

步骤 1

- 业内一线数据库厂商性能均在同一水平，如现有历史系统，其 CPU、内存、磁盘均可以参考现有业务系统类比参考估算。
- 若无现有业务系统参考的，可基于数据库管理员的经验值估计，以下为经验数据，可灵活参考：
 - 客户业务较复杂，即计算型业务：CPU：内存 $\geq 1:4$ ，内存：磁盘 $\geq 1:200$ ，磁盘建议选择 nvme-SSD。
 - 客户业务简单，以大量读写请求为主，即内存型业务：CPU：内存 $\geq 1:12$ ，内存：磁盘 $\geq 1:200$ ，磁盘建议选择 nvme-SSD。
 - 客户业务简单，性能要求不高，但容量大，即存储型业务：CPU：内存 $\geq 1:8$ ，内存：磁盘 $\geq 1:500$ ，磁盘建议选择 SAS-SSD。
 - 存在具体 TPS，QPS 性能需求，但无测试报告的，此时的 TPS，QPS 数据是无意义的，参考上述配置即可。
 - 存在具体 TPS，QPS 性能需求，已有性能报告，根据性能报告需给出测试的机型和客户实际业务需求综合估计估算。
 - 磁盘应该估计到实际存储数据、索引、临时表、日志等实际存储空间，并最好能估计到 3 年内的需求；若仅能估算实际存储数据大小的，基于实际存储数据乘以 3 估计即可。（经验值）
- 示例如下：

计划部署数据库	CPU（物理核数）	内存（GB）	磁盘（GB）	实例个数
业务 A 实例	x	x	x	1
业务 B 实例	x	x	x	20
业务...实例	x	x	x	1
合计总规格	192	960	57600GB	-

步骤 2

- 冗余系数 0.75：是为操作系统和极端情况下预留的资源的差值，根据经验也可以设定为 0.7、0.8 等。
- 物理设备配置的磁盘数据，建议考虑 RAID 后的值（根据当前实际情况，可以考虑 RAID0，RAID5，RAID10 等）。
- 物理设备配置按 CPU、内存、磁盘分别计算后需要取最大值，即应该基于木桶原理考虑设备配置。
- 示例如下：

计划部署数据库	CPU（物理核数）	内存（GB）	磁盘（GB）	备注
合计总规格	192	960	57600	-
物理设备配置	32	192	12000	-
物理设备配置	24	144	9000	-
向上取整 [合计总规格 ÷ （物理设备配置 * 冗余系数 0.75）]	8	实际为 6.7；向上取整为 7	实际为 6.4，向上取整为 7	-

步骤 3

- 示例如下：本次预期部署为同城双中心双活架构，主从节点全部设备数量 = 主节点设备数量 * 合计乘数 = 8 * 4 = 32 台；即 A 机房、B 机房分别 16 台设备

部署模型	A 机房副本数	B 机房副本数	异地机房副本数	合计乘数
同城单中心（无需数据强一致）	2	0	0	2
同城单中心（需要数据强一致）	3	0	0	3
同城双中心（另一中心为灾备机房）	2	1	0	3
同城双中心（另一中心为双活机房）	2	2	0	4
两地三中心（在同城双中心双活基础上增加可切换业务异地机房）	2	2	2	6

步骤 4

- 示例如下：本次需要单独部署 SQL 引擎

部署模型	A 机房节点数	B 机房节点数	异地机房副本数	合计增配 SQL 引擎
同城双中心（另一中心为双活机房）	16	16	0	4

步骤 5

部署模型	A 机房设备数	B 机房设备数	异地机房设备数	合计设备数
同城单中心（管理系统无需高可用）	1	0	0	1
同城单中心（需要数据强一致）	3	0	0	3
同城双中心（另一中心为灾备机房）	2	1	0	3
同城双中心（另一中心为双活机房）：方案 1	2	2（其中 1 台部署但不启用）	0	4
同城双中心（另一中心为双活机房）：方案 2	2	2	1	5
两地三中心（在同城双中心双活基础上增加可切换业务异地机房）-AB 机房集群	2	2	1	5
两地三中心（在同城双中心双活基础上增加可切换业务异地机房）-异地机房集群	2	2	3	7

案例：某行分布式核心两地三中心的部署规划范例：

各组件的服务器配置要求：

No	配置	DB & proxy	ZK	Schedule r/OSS	赤兔/扁鹊	监控 & 采集	LV S	HDFS	KAFKA/ES
1	高性能服务器 SSD 盘，CPU 和内存性能高	Y							
2	普通服务器 SAS 盘，CPU 和内存性能普通		Y	Y	Y	Y	Y		
3	大磁盘容量服务器							Y	

x86机型：监控库DB节点			
项目	描述	数量	RAID及分区登记（包括系统根分区登记）
主机	-	3台	
处理器	Intel Xeon(R) Gold 5220R 2.2Ghz	2	56C
内存	32G DDR4 2933GMhz	6	总计：252G
硬盘1	600G HDD	2	600G*2 (raid1) 系统盘
硬盘2	SATA SSD 960G	6	6*960G (raid10) /data 500g (db软件目录) /data1 1.5T (DB数据) /data1 1.5T (DB日志)
阵列卡	支持raid0,1,5,6,10,50,60, 2G以上及电池保护模块	1	
网卡	万兆接口：双口10Gb Broadcom BCM957412 芯片光口 网卡*2	2	bond, mode=1

x86机型：应用DB节点			
项目	描述	数量	RAID及分区登记（包括系统根分区登记）
主机	-	32台	
处理器	Intel Xeon(R) Gold 5220R 2.2Ghz	2	112C
内存	32G DDR4 2933GMhz	8	总计：512G
硬盘1	600G HDD	2	600G*2 (raid1) 系统盘
硬盘2	SATA SSD 960G	6	6*960G (raid10) /data 500g (db软件目录) /data1 1.5T (DB数据) /data1 1.5T (DB日志)
阵列卡	支持raid0,1,5,6,10,50,60, 2G以上及电池保护模块	1	
网卡	万兆接口：双口10Gb Broadcom BCM957412 芯片光口 网卡*2	2	bond, mode=1

x86机型：赤兔PROXY节点			
项目	描述	数量	RAID及分区登记（包括系统根分区登记）
主机	-	3台	
处理器	Intel Xeon(R) Gold 5220R 2.2Ghz	2	56C
内存	32G DDR4 2933GMhz	6	总计：252G
硬盘1	600G HDD	2	600G*2 (raid1) 系统盘
硬盘2	SATA SSD 960G	6	6*960G (raid10) /data 500g (PROXY程序+PROXY日志+管控程序) /data1
阵列卡	支持raid0,1,5,6,10,50,60, 2G以上及电池保护模块	1	
网卡	万兆接口：双口10Gb Broadcom BCM957412 芯片光口 网卡*2	2	bond, mode=1

x86机型：应用PROXY节点			
项目	描述	数量	RAID及分区登记（包括系统根分区登记）
主机	-	8台	
处理器	Intel Xeon(R) Gold 5220R 2.2Ghz	2	112C
内存	32G DDR4 2933GMhz	8	总计：512G
硬盘1	600G HDD	2	600G*2 (raid1) 系统盘
硬盘2	SATA SSD 960G	6	6*960G (raid10) /data 500g (PROXY程序+PROXY日志+管 控程序) /data1
阵列卡	支持raid0,1,5,6,10,50,60, 2G以上及电池保护模块	1	
网卡	万兆接口：双口10Gb Broadcom BCM957412 芯片光口 网卡*2	2	bond, mode=1

x86机型：管控节点			
项目	描述	数量	RAID及分区登记（包括系统根分区登记）
主机	-	5台	
处理器	Intel Xeon(R) Gold 5220R 2.2Ghz	2	56C
内存	32G DDR4 2933GMhz	6	总计：189G
硬盘1	600G HDD	2	600G*2 (raid1) 系统盘
硬盘2	SATA SSD 960G	6	6*960G (raid10) /data 500g /data1 1.5T
阵列卡	支持raid0,1,5,6,10,50,60, 2G以上及电池保护模块	1	
网卡	万兆接口：双口10Gb Broadcom BCM957412 芯片光口 网卡*2	2	bond, mode=1

步骤 6

- 其中 LVS 建议选择物理机，基本配置应选择为具有较高转发能力物理集群。
- HDFS 考虑可维护性，推荐使用商用的 NAS、COS 等存储。
- 备份存储集群对磁盘的要求为，对设备 CPU、内存、磁盘类型、网卡类型无明显要求：

子系统名称	实际涉及软件模块	功能说明	本次部署是否包含	开源/商用说明	规划设备台数
数据库负载均衡	LVS	提供数据库实例唯一接入 IP、并实现分布式场景下负载均衡能力	<input checked="" type="checkbox"/> 是 <input type="checkbox"/> 否	可以选用商用负载均衡、商用 DNS、腾讯商用 CLB、VPC；LVS 可以选用商用版本，或可对应开源版本。	2

备份存储集群	HDFS	长期存储数据库备份、日志等文件；用于数据灾难恢复，从机重建等场景	■是 □否	可以选用商用 HDFS，商用 NAS，商用对象存储；HDFS 可以选用商用版本，或可对应开源版本。磁带库可后置于 HDFS 后。	3 台起。
全局分布式事务调度服务	Metacluster GTS	提供分布式事务实时读一致能力。不安装该组件时均提供分布式事务最终一致性特性。	■是 □否		1 台或 3 台
数据订阅与生产	Kafka Consumer	提供数据生产到消息中间件队列中，订阅、	■是 □否	Kafka 可以选用商用版本，或可对应开源版本。	3 台起
日志存储与查询	ElasticSearch LogStash Kibana	提供长期的日志存储、查询和日志监控能力	□是 ■否	腾讯不提供相关软件安装包，请您自己部署商用或开源版本。	不涉及

步骤 7

子系统名称	规划设备配置	虚拟机/物理机	规划设备台数	其他要求
运营管理系统	-	虚拟机	5	同机房内跨机架部署
数据库核心	-	物理机	36	同机房内跨机架部署
数据库负载均衡	-	物理机	4	同机房内跨机架部署
备份存储集群	-	物理机	3	-
全局分布式事务调度服务	-	虚拟机	1	-
数据订阅与生产	-	虚拟机	3	-
日志存储与查询	不涉及	不涉及	不涉及	不涉及

2.2.2 软件规划

软件规划表

子系统名称	服务器 IP	本地防火墙端口 (选填)	异地防火墙端口 (选填)	备注
运营管理系统				
数据库核心				
数据库负载均衡				
备份存储集群				
全局分布式事务调度服务				
数据订阅与生产				
日志存储与查询	不涉及	不涉及	不涉及	
子系统名称	服务器 IP	本地防火墙端口 (选填)	异地防火墙端口 (选填)	备注
Chitu				
Zookeeper				
Monitor				
Scheduler/Manager				
OSS				
CloudDBA				
DATABASE				
DataBase (含 Agent/ocagent)				
Proxy (含 Agent/ocagent)				
LVS				
HDFS (选填)				
Metacluster (含 GTS)				
Kafka				

其他依赖软件（必选）

为提高产品的可靠性，产品还需依赖如下软件，请按需配置。

- NTP 网络时间服务器（不限版本）。
- 操作系统官方 yum 源。

2.3 分布式核心系统数据库资源评估

为了快速评估应用系统所需分布式数据库资源，特制定本手册。本手册中的评估方法为基于经验值和简单测试，每个应用系统的实际情况可能有较大差异，最优资源评估方案应以每个系统的性能测试结果为准。如下评估资源以 SSD 为基础，评估资源时，应考虑应用系统未来 5 年的业务量增长情况。

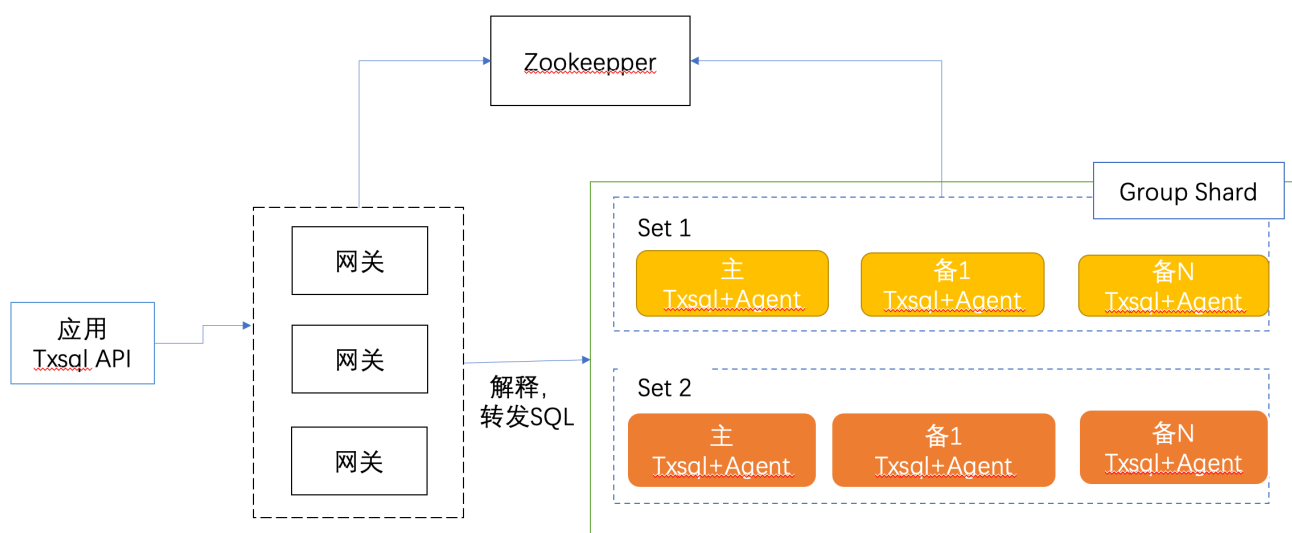
分布式数据库分为 No-Shard（非分布式模式）和 GroupShard（分布式模式），对比如下：

特性	No-shard	Shard
分库分表	应用层实现	数据库提供
分布式事务		
水平扩展能力		
存储过程、函数、触发器	YES	NO
传统 MYSQL 应用改造	兼容	需改造
关联查询	单一分库内支持	ShardKey 下最优

以下原则均为经验估算值，最终应以系统实测数据为准。

2.3.1 分布式实例(GroupShard)资源评估方法

分布式实例架构图



Group Shard 数据分片(Set)数量

此处的分片为物理分片，一个物理分片对应 1 个 Set，为一组一主多从复制关系的 TXSQL 实例，主库负责读写。一个 Group Shard 实例包含多个物理分片。

- 单个数据库实例管理最大物理分片数不超过 32。如果测算后分片数超过 32，建议进行数据库实例拆分改造。
- 单个分片数据容量（数据+索引）分界为 1T，TPS（数据库每秒事务数）最大分界为 3000-4000，QPS（数据库每秒 SQL 数）最大分界为 40000-50000，任一指标超出分界需增加分片。单 Set 平稳运行 QPS（数据库每秒 SQL 数）为 30000（读写比约 3:2），TPS（数据库每秒事务数，按每事务 20 条 SQL 估算）为 1500。
- 当 TPS 远低于 QPS 值时，例如对公业务，单个事务的 SQL 数为 50-90 条（QPS:TPS = 50:1-90:1），应参考 QPS 指标，但测算出的分片数不能超过 16。

注：

- 对于无数据依赖的应用，建议数据库采用 No-Shard 方式进行分库，避免使用分布式特性。
- 可尝试通过调整热点数据分布等方式来避免单分片 TPS 过高。
- 为确保分片性能，需参考《分布式数据库开发手册》进行适配开发，同时建议分布式事务占比不超过 15%。
- 若 TPS 要求较低，例如历史库，可以适当增加单分片数据容量至不超过 2T。

Proxy（网关节点）数量

Proxy 是无状态服务，多个 Proxy 机器可以同中心或者跨中心组成一个网关组，为同一数据库实例服务。考虑到 Proxy 的高可用性：

- 灾备四级系统，建议主中心同一网关组部署 Proxy 机器数量最少为 2，一般单中心的 Proxy 实例数为数据分片数/2，即单个 Proxy 可承接两个数据分片的业务；每个网关组最多部署 2 个 Proxy 实例。
- 灾备五级系统，建议单中心同一网关组部署 Proxy 机器数量最少为 2，若两中心部署，需要的 Proxy 机器数量最少为 4；一般单中心的 Proxy 实例数为数据分片数/2，即单个 Proxy 可承接两个数据分片的业务；每个网关组仅部署 1 个 Proxy 实例。

后续根据实际运行情况决定是否扩容。建议单个网关组 Proxy 机器数量不超过 16，单中心网关组 Proxy 机器数量不超过 8。

集群数量

一个集群共用一套管理节点，可以纳管多个数据库实例（数据分片及其对应的 Proxy）。

- 灾备四级系统建议单个集群纳管的数据分片总数不超过 64，超过需新建集群。
- 灾备五级系统建议单个集群纳管的数据分片总数不超过 64，超过需新建集群。5+级系统一个业务系统建议 1 套集群。

注：如果集群管理节点故障，该集群中所有纳管分片都将受影响，集群的监控告警、数据库自动切换等功能将不可用，进而可能影响业务，因此共用管理节点前请充分评估故障场景下的影响，建议重要系统单独部署集群。不建议跨网络分区部署集群。

分片设计相关案例

下述系统的模拟分片方案供参考。

说明：容量以每年 15%增长进行估算。TPS/QPS 以每年 30%增长进行估算。以 3 个维度的最大值进行最终分片建议估算。（具体增长情况由应用具体评估，以下为测算样例）。

案例 1：系统 1

	基线	增量（2 年增长）	总量	模型分片	分片建议
容量(T)	1	0.3	1.3	2 分片	4 分片
TPS	6000	3600	9600	4 分片	
QPS	25000	15000	40000	1 分片	

案例 2：系统 2

	基线	增量（2 年增长）	总量	模型分片	分片建议

容量(T)	1.5	0.45	1.95	2 分片	2 分片
TPS	500	300	800	1 分片	
QPS	5000	3000	8000	1 分片	

案例 3：系统 3（模拟使用 GroupShard 场景，建议保留现在分库方式）

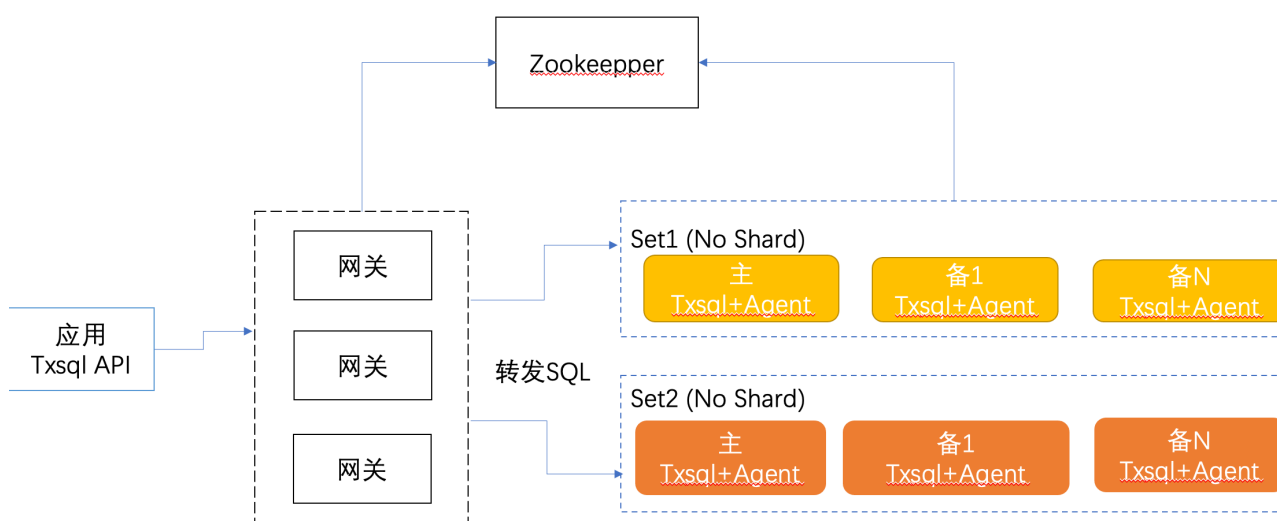
	基线	增量（2 年增长）	总量	模型分片	分片建议
容量(T)	10	3	13	13 分片	16 分片
TPS	40000	24000	64000	16 分片	
QPS	280000	168000	448000	12 分片	

案例 4：系统 4

	基线	增量（2 年增长）	总量	模型分片	分片建议
容量	3	0.9	3.9	4 分片	4 分片
TPS	6600	3960	10560	4 分片	
QPS	125000	75000	200000	4 分片	

2.3.2 非分布式实例(No-Shard)资源评估方法

非分布式实例架构图



No-Shard 数据库实例数量

每个 No-Shard 数据库实例仅包含 1 个 Set, 为一组一主多从复制的 TXSQL 实例，主库负责读写。

- 单个分片数据容量（数据+索引）分界为 1T, TPS（数据库每秒事务数）最大分界为 3000-4000, QPS（数据库每秒 SQL 数）最大分界为 40000-50000，任一指标超出分界需增加分片。单 Set 平稳运行 QPS（数据库每秒 SQL 数）为 30000（读写比约 3:2），TPS（数据库每秒事务数，按每事务 20 条 SQL 估算）为 1500。
- 若 TPS 要求较低，例如历史库，可以适当增加单分片数据容量至不超过 2T。

Proxy（网关节点）数量

No-Shard 模式下，Proxy 的主要功能为转发收到的 SQL，资源占用较少，一般不会成为性能瓶颈。考虑到 Proxy 的高可用性：

- 灾备四级系统，建议主中心同一网关组部署 Proxy 机器数量最少为 2，每个网关组最多部署 2 个 Proxy 实例。
- 灾备五级系统，建议单中心同一网关组部署 Proxy 机器数量最少为 2，若两中心部署，需要的 Proxy 机器数量最少为 4；每个网关组仅部署 1 个 Proxy 实例。

i 说明:

若服务器配置较高，例如服务器 CPU 核数较多等情况，可依据以上原则进行评估。建议单个网关组 Proxy 机器数量不超过 16，单中心网关组 Proxy 机器数量不超过 8。

集群数量

一个集群共用一套管理节点，可以纳管多个数据库实例（数据分片及其对应的 Proxy）。

- 灾备四级系统建议单个集群纳管的数据分片总数不超过 64，超过需新建集群。
- 灾备五级系统建议单个集群纳管的数据分片总数不超过 64，超过需新建集群。

i 说明:

- NVME 建议数据分片总数不超过 128。
- 如果集群管理节点故障，该集群中所有纳管分片都将受影响，集群的监控告警、数据库自动切换等功能将不可用，进而可能影响业务，因此共用管理节点前请充分评估故障场景下的影响，建议重要系统单独部署集群。

2.3.3 数据库部署方案

当前服务器配置基线

ARM 服务器

灾备四级和五级系统的管理节点、Proxy 和 DB 参考的服务器配置为:

- CPU: 鲲鹏 920（ARM 架构）2*48C
- 内存: 256G
- 存储: 系统盘 2*600G RAID1，SSD 数据盘 8*800G RAID10
- 网卡: 双万兆

X86 服务器

灾备四级和五级系统的管理节点和 Proxy 参考的服务器配置为:

- CPU: Intel x86 2*24C
- 内存: 256G
- 存储: 系统盘 2*600G RAID1，SATA SSD 数据盘 960G*4 RAID10
- 网卡: 双万兆

灾备四级系统 DB 参考的服务器配置为:

- CPU: Intel x86 2*24C
- 内存: 512G
- 存储: 系统盘 2*600G RAID1, SATA SSD 数据盘 960G*12 RAID10
- 网卡: 双万兆

灾备五级系统 DB 参考的服务器配置为:

- CPU: Intel x86 2*24C
- 内存: 512G
- 存储: 系统盘 2*600G RAID1, SATA SSD 数据盘 960G*6 RAID10
- 网卡: 双万兆

说明:

- 根据运维要求,一般日常情况下 CPU 使用率峰值不超过 30%, 联机高峰期不超过 50%, 数据盘使用率不超过 60%。
 - 本手册中容量评估均基于以上服务器配置进行。如果服务器配置有所变化,可根据实际情况进行调整。
-

数据分片部署规则

ARM 服务器

灾备四级系统建议单台 DB 服务器最多用于部署 2 个数据分片, 不建议两个分片的主库都部署在同一台服务器。根据当前服务器配置, 单分片可使用的资源为:

- CPU 为 24C, 计算方法为: $2*48C$ (总 CPU 数) $*0.5$ (联机高峰期不超过 50%) $/2$ (复用实例数 2 个)。
- 可用内存为 96G, 计算方法为: $256G$ (总内存数) $*0.75$ (预留 25%) $/2$ (复用实例数 2 个)。
- 可存储数据 (数据 70%+索引 30%) 约 1T, 可存储日志约 500G。数据+日志总空间约 1.5T (其中数据空间占总空间的 75%, 日志空间占总空间的 25%), 计算方法为: $(800G$ (盘大小) $*12$ (盘数) $/2$ (RAID10) $-500G$ (/data 使用)) $*0.7$ (预留 30%) $/2$ (复用实例数 2 个)。当前现有配置 8*800G 情况下由于硬盘大小和空闲率要求, 最多只能放置 1 个片。

灾备五级系统建议单台 DB 服务器仅用于部署 1 个数据分片。根据当前服务器配置, 该分片可使用的资源约为:

- CPU 为 48C, 计算方法为: $2*48C$ (总 CPU 数) $*0.5$ (联机高峰期不超过 50%)。
- 可用内存为 192G, 计算方法为: $256G$ (总内存数) $*0.75$ (预留 25%)。
- 可存储数据 (数据 70%+索引 30%) 约 1T, 可存储日志约 500G。数据+日志总空间约 1.5T, 计算方法为: $(800G$ (盘大小) $*8$ (盘数) $/2$ (RAID10) $-500G$ (/data 使用)) $*0.6$ (预留 40%), 其中数据空间占总空间的 75%, 日志空间占总空间的 25%。

- 若服务器配置较高，例如服务器安装的硬盘数量较多等情况，可依据以上原则进行评估。

X86 服务器

灾备四级系统建议单台 DB 服务器最多用于部署 2 个数据分片，不建议两个分片的主库都部署在同一台服务器。根据当前服务器配置，单分片可使用的资源为：

- CPU 为 24C，计算方法为： $2 \times 24C \times 2$ （CPU 开启超线程） $\times 0.5$ （联机高峰期不超过 50%） $\div 2$ （复用实例数 2 个）。
- 可用内存为 192G，计算方法为： $512G$ （总内存数） $\times 0.75$ （预留 25%） $\div 2$ （复用实例数 2 个）。
- 可存储数据（数据 70%+索引 30%）约 1T, 可存储日志约 500G。数据+日志总空间约 1.5T，计算方法为： $(960G$ （盘大小） $\times 12$ （盘数） $\div 2$ （RAID10） $- 500G$ （/data 使用）） $\times 0.6$ （预留 40%） $\div 2$ （复用实例数 2 个），其中数据空间占总空间的 75%，日志空间占总空间的 25%。

灾备五级系统建议单台 DB 服务器仅用于部署 1 个数据分片。根据当前服务器配置，该分片可使用的资源约为：

- CPU 为 48C，计算方法为： $2 \times 24C \times 2$ （CPU 开启超线程） $\times 0.5$ （联机高峰期不超过 50%）。
- 可用内存为 384G，计算方法为： $512G$ （总内存数） $\times 0.75$ （预留 25%）。
- 可存储数据（数据 70%+索引 30%）约 1T, 可存储日志约 500G。数据+日志总空间约 1.5T，计算方法为： $(960G$ （盘大小） $\times 6$ （盘数） $\div 2$ （RAID10） $- 500G$ （/data 使用）） $\times 0.75$ （预留 25%），其中数据空间占总空间的 75%，日志空间占总空间的 25%。

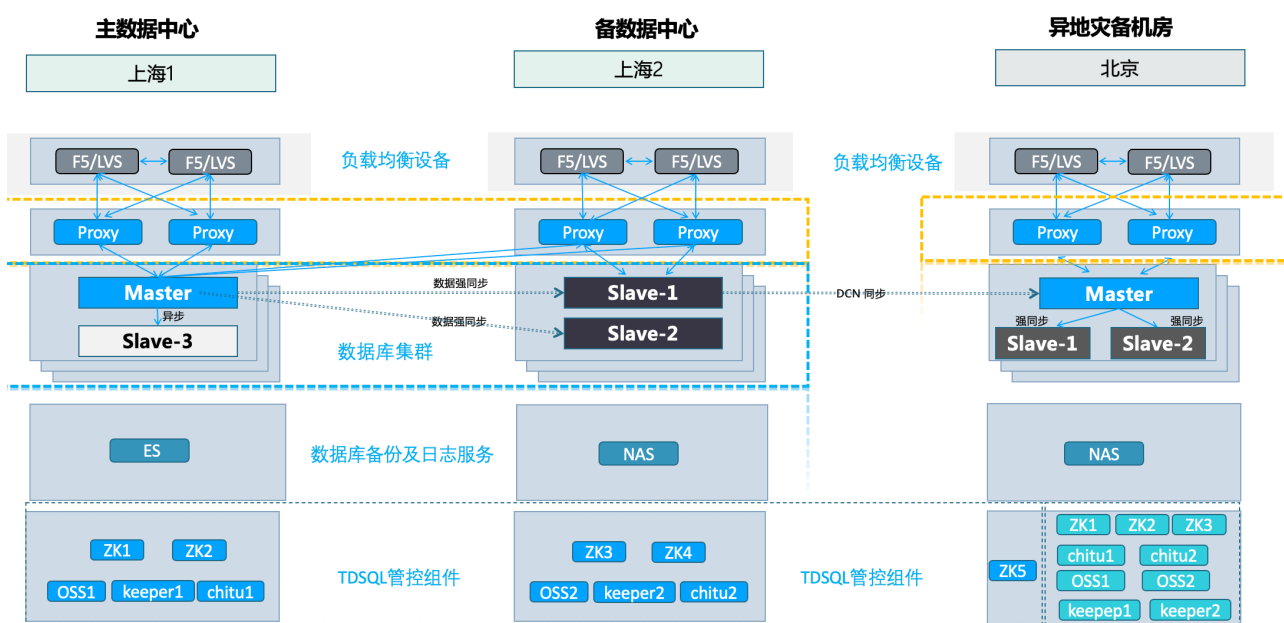
若服务器配置较高，例如服务器安装的硬盘数量较多等情况，可依据以上原则进行评估。

2.3.4 架构模型及服务器资源预估

集群部署架构模型

灾备等级 5 级

异地灾备部署：两地三中心

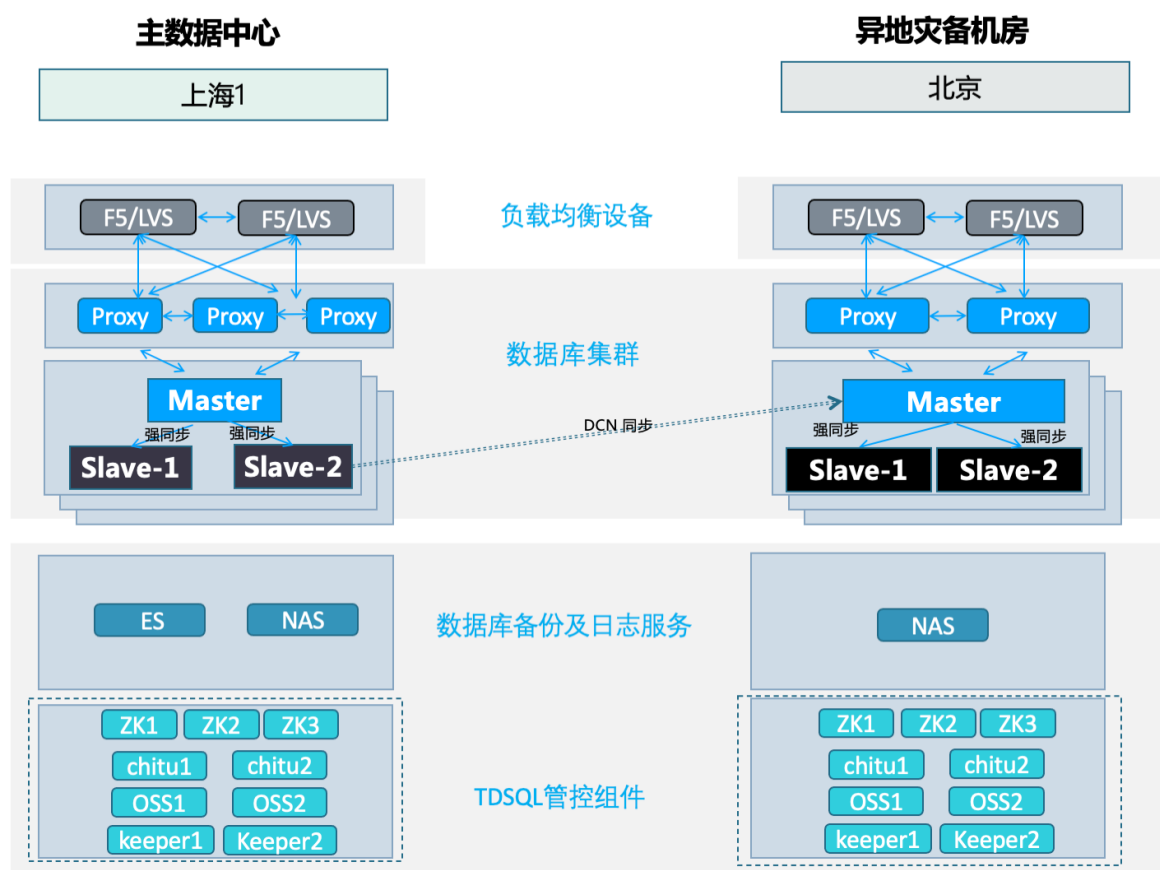


说明:

考虑 RPO 优先为前提，上海和北京集群生产用途数据库保持强同步不可退化模式，监控库保持强同步可退化模式。

灾备等级 4 级

异地灾备部署：两地两中心



① 说明:

- 备注：北京机房不需要 ES。
- 考虑 RPO 优先为前提，上海和北京集群生产用途数据库保持强同步不可退化模式，监控库保持强同步可退化模式。

数据分片服务器数量

首先确定数据分片数、副本数（1主几从），再结合数据分片部署方案确定集群所需服务器数量。

异地灾备集群服务器数量减半。

Proxy 服务器数量

根据数据分片数确定 Proxy 实例数量，再结合 Proxy 部署方案确定主集群所需服务器数量。

异地灾备集群服务器数量减半。

管理节点服务器数量

单个集群里的所有数据库实例共用一套管理节点。根据不同的部署架构，管理节点需要的服务器数量也不相同。

主要有以下三种情况：

1. 本地高可用（两地两中心）：单个集群各部署管理节点 3 台服务器。
2. 同城高可用 + 两地三中心部署（一地两中心）：单个集群管理节点 5 台服务器。
3. 同城高可用 + 两地三中心部署 + 异地灾备（两地三中心）：单个集群主集群管理节点 5 台服务器，灾备集群管理节点 3 台服务器。

① 说明：

- 单个集群需要一套监控库，独占 1 个 Set，用来存放监控数据以及应急使用，监控库与业务库不能使用相同物理机。（5 级系统 1 主 3 从，4 级系统 1 主 2 从）。
 - 单个集群审计节点 1 台服务器，用于 SQL 审计。
-

其它

- 负载均衡设备 F5 各中心按需申请。
- NAS、带库、备份服务器按需申请。
- DMC 服务器按需申请。

3 端口规划

如果您在数据库集群之内、数据库集群之间或数据库与客户端之间存在端口限制，请根据实际情况放通如下端口。

TCP 端口

端口用途	端口范围	说明	部署地址
SSH 登录	22、36000	1. 系统服务，可配置，按实际情况配置即可；常用 22,36000 2. 目前只支持统一配置，要求每台机器的 ssh 端口必须一致	
zookeeper 服务	2118、2181、2338、2558、9033	下面三个端口支持自定义，按实际情况配置即可，默认值： tdsql_zk_clientport: 2118 tdsql_zk_serverport1: 2338 tdsql_zk_serverport2: 2558 adminserver: 9033 # 默认不启动	/data/application/zookeeper
Keeper	8978		/data/application/scheduler
OSS	8080	OSS 接口服务监听端口，供赤兔等各类服务访问调用	/data/application/oss
OC-Agent	8966	服务监听端口，跨机器的通信操作强依赖，包括跨集群 DCN 操作	/data/oc_agent
tdsqlpcload/nginx	80、443	老赤兔，用的 php+nginx，http 用的 80 端口，https 用的 443 端口	/data/website/tdsqlpcload
tdsqlchitu	80、443	新赤兔，http 用的 80 端口，https 用的 443 端口 新赤兔后续逐步会替换老赤兔	/data/tdsqlchitu
PROXY	15001-30000	根据实例生成，默认配置是 15001-30000，在 manager 配置文件中设置。一般不修改，建议维持默认配置。	
DB	TCP: 4001-4900 14001-14900	1. DB 端口范围私有云默认配置为 4001-4900，在 manager 配置文件中设置； 2. 14001-14900 为 DB 的管理端口，默	

		认是 DB 服务端口+10000 3. 这些配置一般不修改，建议维持默认配置	
OPS	8081	自动化部署（10.3.21 引入）&升级服务（10.3.22 引入）	/data/tdsql_ops
exporter	9101、9201	秒级监控服务，10.3.22 版本引入	/data/application/tdsql_exporter
monitor	8082	秒级监控服务，10.3.22 版本引入	/data/application/tdsql_monitor
CloudDBA	9011	扁鹊，诊断服务	/data/application/clouddba
MC	12379、12380、12381	全局一致性读组件服务	/data/application/TD SQL_MetaCluster
Kafka 服务	9092	可选组件，KAFKA 服务，用于多源同步	/data/application/kafka/
HDFS 服务	8480、8485、9864、9866、9002、9867、8019、50070	可选组件，如果不用 HDFS 则可以忽略	/data/home/tdsql/hadoop

UDP 端口

端口用途	端口范围	说明
业务 DB 服务	20000~22000	agent 用来收发 alldump 的
业务 DB 服务	4001-4999	用于强同步返回 ACK 用

跨集群之间通讯

警告：

跨集群之间通讯端口可能会因为复用组件不同而有一定差异，下列表格仅列出最必要的端口，请根据实际请完整梳理。

端口用途	端口范围	协议
跨集群互相通讯	22、36000、8966	TCP
纳管异地集群 OSS 服务端口	8080	TCP

跨集群 zookeeper 服务	2118、2181、2338、2558、 2888、3888	TCP
跨集群 DCN 同步	4001-4900、14001-14900	TCP
跨集群 DCN 同步	20000-22000	UDP

4 网络规划

TDSQL 网络质量要求

- 对于数据安全等级较高的业务，一般采用两地三中心的部署，同城之间的网络延迟正常不超过 3ms（采用强同步，延迟过大会导致写入性能瓶颈）。
- 异地之间，如果采用异步复制方式，网络延迟以业务容忍度为准，网络延迟越大，数据同步的速度越慢，异地主从之间的数据同步时延越大,网络延迟不超过 30ms。
- 网络 ping 丢包率 0%。
- 网络 ping 的 min/avg/max/mdev 结果应该在同一个数量级，不存在延迟跳包。
- 网络带宽，服务器 10Gbit 网卡，同城专线 10Gbit 以上带宽，不允许走外网搭建主从。

异地网络流量评估

设计文档流量预估

1. 当前异地网络情况：异地带宽 500Mbps，平均延时 35ms。
2. 假设上海数据中心到深圳数据中心目前带宽只有 500Mbps，根据凌晨批量导入 7800 万数据的业务场景推算，单条记录的大小为 0.3kb，预计 2 分钟完成数据导入，凌晨的带宽将会达到 1520Mbps（ $78000000 \times 0.3 \times 8 / 120 / 1024$ ），因此凌晨会出现 3~5 分钟业务异地传输网络带宽打满的情况。
3. 假设当前部署架构中，上海数据中心 IDC1 有 16 台 DB 服务器、上海数据中心 IDC2 有 16 台 DB 服务器。数据库 TDSQL-DCN 异步复制流量主要是从上海数据中心 IDC2 数据中心到深圳数据中心。核算引擎系统在线库实例是 32 分片，分布在上海数据中心 IDC1 和上海数据中心 IDC2DB 服务器上，32 分片的主都在上海数据中心 IDC1 服务器。单台上海数据中心 IDC1 服务器有 2 个分片的 Master 和 2 个分片的异步 Slave。单台上海数据中心 IDC2 服务器有 2 个分片的 2 个强同步 Slave。DCN 异步复制流量为平时流量的一半。

生产实际情况及异地流量估算

当前生产流量统计及异步复制流量预测如下：

类别	日常带宽平均值	批量带宽峰值	备注
单台 DB 带宽（最大）	16Mbps	540Mbps	选取上海数据中心 IDC2 最大的一台
单台 DB 带宽（平均）	10Mbps	300Mbps	
单台 DB 的 DCN 带宽	8Mbps	270Mbps	DCN 流量为平时流量的一半

(最大)			
单台 DB 的 DCN 带宽 (平均)	5Mbps	150Mbps	
异地带宽 (最大)	128Mbps	4320Mbps	选最大一台, 16 台, 峰值持续 5 分钟左右
异地带宽 (平均)	80Mbps	2400Mbps	选平均一台, 16 台, 峰值持续 5 分钟左右

- 如果按目前上海到深圳的异地网络带宽上限 500Mbps 进行传输, 那么按最大和平均估算, 会持续 40 分钟或 25 分钟打满异地带宽上限。
- 建议按异地带宽 (平均) 值 2400Mbps 估算进行扩容, 上海到深圳的异地带宽扩容到 3000Mbps。

案例参考

某国有大行异地网络: 异地带宽 5000Mbps, 平均延时小于 30ms;

双中心同城流量估算 (通用系统流量估算)

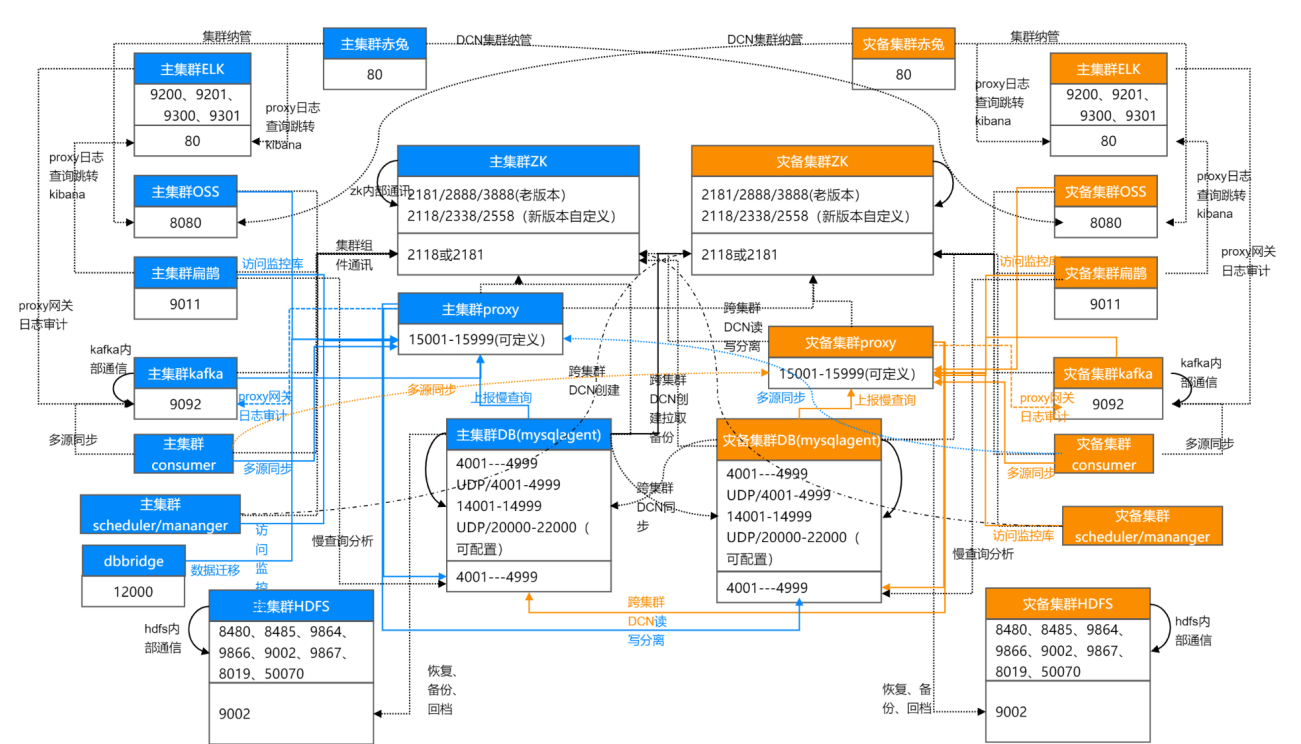
功能项	日常流量 (估算)	批量场景流量 (30 倍)	备注
峰值 QPS	10000	300000	峰值 QPS 预估
单行记录大小 (Byte)	1024	1024	极限情况下一行记录为 1024
南北向流量 (Mbps)	78.125	2343.75	负载均衡流量
万兆网络带宽 (Mbps)	10240	10240	10G 网络
带宽占比	0.76%	22.89%	负载均衡流量占比
东西向流量 1+1 (同城 Mbps)	39.0625	1171.875	一主一从
东西向流量 2+2 (同城 Mbps)	78.125	2343.75	一主三从
东西向流量 3+3 (同城 Mbps)	117.1875	3515.625	一主五从
1+1 带宽占比	0.38%	11.44%	跨 AZ 流量占比
2+2 带宽占比	0.76%	22.89%	跨 AZ 流量占比
3+3 带宽占比	1.14%	34.33%	跨 AZ 流量占比

① 说明:

- 按照 10000 QPS（每秒 SQL 数）峰值计算，根据业务模型，读/写比例为 4:6（极端情况 5:5），单个 SQL 修改或插入的行数为 1 行，按 10000 行来计算每秒需要传输的 binlog 大小。
- 1 Byte = 8 Bits => 10000*0.5*1*1K（极端情况一行 1K 的情况）= 4.88MB/s

TDSQL 架构和防火墙策略

TDSQL 的各个组件通信端口和关系如下图



① 说明:

所有机器需要开通 22、36000、8966 端口用于登录和管理。

具体开通策略列表

源地址组件	目标地址组件	服务端口	备注
集群内所有机器	集群内所有机器	ssh 端口选一个：22、36000 ocagent: 8966	公共端口；音

			是、管理站口
集群内所有机器	本集群 scheduler	8978	o c a g e n t 上 k e e p e r 自通信
主集群赤兔	主集群 OSS	8080	本互网管集群
灾备集群赤兔			跨城集群管理
跨城 intercity cluster 模块			跨城 d

			跨城 d c n 部署)
主集群赤兔	跨城总控 OSS	8080	注丹集君至跨城柜上跨城 d c n 主各柜控，跨城 d c
灾备集群赤兔			
跨城 intercity cluster 模块			

			n 打 客)
主集群 zookeeper	主集群 zookeeper	2181/2888/3888 或 2118/2338/2558	z k p 音 通 计 站 口 , 目 版 本 使 用 黑 计 站 口 , 亲 版 本 站 口 与 西 置 集 群 内 组 件 通
主集群 OSS		2118 或 2181	
主集群 scheduler/manager			
主集群 proxy			
主集群 mysqlagent(db)			
主集群 consumer			
主集群 kafka			

			集群 dcn 该至 个 高 巧 育
灾备集群 scheduler/manager			跨 集 群 dcn 仓 灾 黑 该 跨 块 柜 块 共 月 三 集 群 自 zk 集 群
跨城 OSS/inttercity cluster 模块			
灾备集群 zookeeper	灾备集群 zookeeper	2181/2888/3888 或 2118/2338/25	zk 尸 音

		59	通 计 站 口 ， 旧 版 本 使 用 黑 计 站 口 ， 亲 版 本 ， 站 口 与 可 置
灾备集群 OSS		2118 或 2182	集
灾备集群 scheduler/manager			群
灾备集群 proxy			内
灾备集群 mysqlagent(db)			组
灾备集群 consumer			作
灾备集群 kafka			通 计 ， 旧 版 本 使

			月黑云暗，亲见不靖口，可酉置扁昔」打距集君 d c n 仓亥打耳各份距集君 d c n 这
灾备集群扁鹊			
主集群 mysqlagent(db)			
主集群 proxy			

			主 分 区 自 主 集 群 d c n 仓 延
主集群 scheduler/manager			
主集群 mysqlagent(db)	主集群 mysqlagent(db)	4001---4999 UDP/4001-4999 14001-14999 UDP/20000-22000（可配置）	
灾备集群 mysqlagent(db)			主 分 区 自 主 集 群 d c n 同 步
主集群 proxy		4001---4999	
主集群扁鹊			慢 查 询 分 析 机
灾备集群 proxy			主 分 区 自 主 集 群 d c n 读 主 分 区

主集群 scheduler			心跳 「打走时」 二次迁移
灾备集群 mysqlagent(db)	灾备集群 mysqlagent(db)	4001---4999 UDP4001-4999 14001-14999 UDP/20000-22000（可配置）	跨集群 d c n 同步
主集群 mysqlagent(db)			
灾备集群 proxy		4001---4999	慢查 询 久 树 跨集群 d c n 读写 久 高 心跳 「
灾备集群扁鹊			
主集群 proxy			
灾备集群 Scheduler			

			机 走 时 二 次 抄 源
主集群接入层	主集群 proxy	15001~15999	
主集群 proxy			k i l l 会 话 工 具
			后 昔 该 作 出 控 屏
主集群扁鹊			多 源 同 步
主集群 consumer			C S S 接 入 出 控 屏
灾备集群 consumer			
主集群 OSS			
主集群 scheduler/manager			s c h

			e d u l e r 接 作 入 出 控 屏
主集群 mysqlagent(db)			n y s q l a g e n t 」 排 程 查 询
灾备集群接入层	灾备集群 proxy	15001~15999	
灾备集群集群 proxy			k i l l 会 话 巧 育 屏 昔 该 闻
灾备集群扁鹊			

			出
			控
			屏
主集群 consumer			多
灾备集群 consumer			源
			同
			步
灾备集群 OSS			C
			S
			S
			接
			作
			入
			出
			控
			屏
灾备集群 scheduler/manager			s
			c
			h
			e
			d
			u
			l
			e
			r
			接
			作
			入
			出
			控
			屏
灾备集群 mysqlagent(db)			n
			y
			s
			q
			l
			a
			g
			e
			n

			t 上 机 情 查 询
主集群 hdfs	主集群 hdfs	8480、8485、 9864、9866、 9002、9867、 8019、50070	h d f s 内 音 通 话 端 口
主集群 mysqlagent(db)		9002	恢 复 、 备 份 ， 回 档
灾备集群 hdfs	灾备集群 hdfs	8480、8485、 9864、9866、 9002、9867、 8019、50070	h d f s 内 音 通 话 端 口
灾备集群 mysqlagent(db)		9002	恢 复 、 备 份

			份， ， 国 本
主集群 kafka	主集群 kafka	9092	k a f k a p 音 通 计
主集群 consumer			多 源 同 步
主集群 proxy			p r o x y 网 端
主集群 ELK			E L K 自 主 计 划 自 主 计 划
灾备集群 kafka	灾备集群 kafka	9092	k a f k a p 音 通 计

灾备集群 consumer			多源同步
灾备集群 proxy			
主集群赤兔	主集群 ELK	80	proxy 网络日志审计项目
主集群扁鹊			
主集群赤兔	灾备集群 ELK	80	proxy 日志审计项目
灾备集群赤兔			

			日志查询跨年至kibana
主集群 OSS	主集群扁鹊	9011	扁鹊
灾备集群 OSS	灾备集群扁鹊		备份
主集群 MC 内部	主集群 MC	12380	
主集群 proxy		12381	
主集群 collector		12379	
灾备集群 MC 内部	灾备集群 MC	12380	
灾备集群 proxy		12381	
灾备集群 collector		12379	