贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

基于RoCE技术的无损以太网
替代SAN存储网络在信创云平台的探索和实践

贵阳农商银行

目录

PART 01

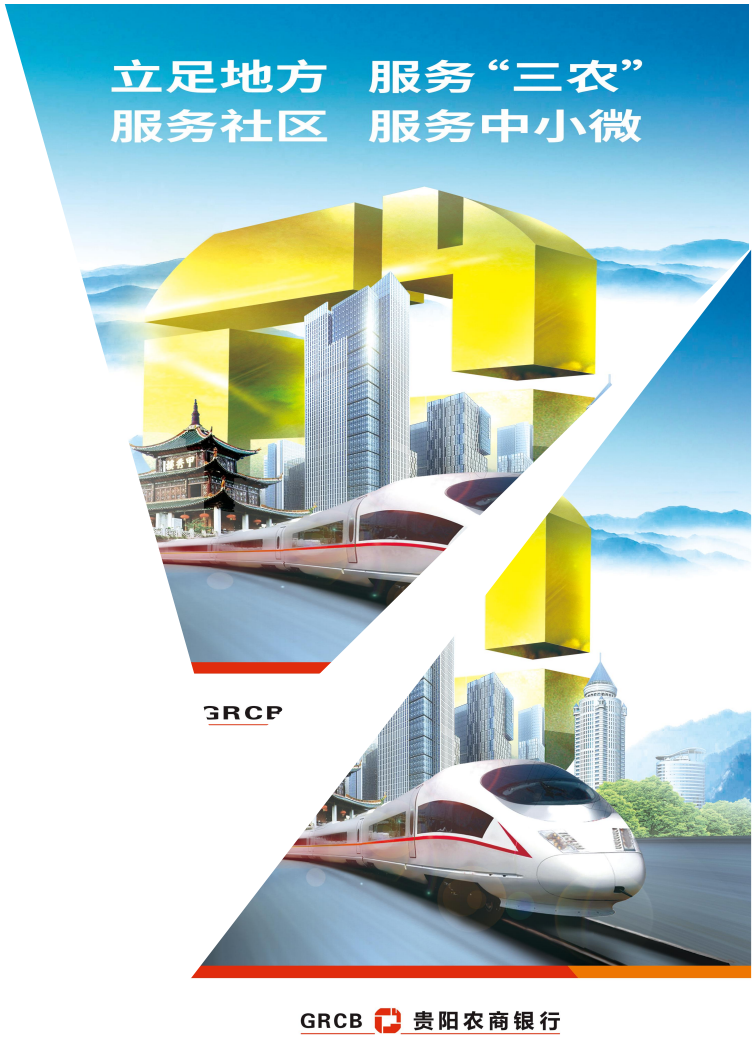
项目背景及目标

项目背景及目标

贵阳农村商业银行股份有限公司（简称贵阳农商银行）成立于2011年12月22日，注册资本38.23亿元人民币，是由原贵阳市云岩、南明、小河、白云四城区农村信用社（合行）改制创建而成的，有着50多年的历史传承和文化底蕴，是贵阳市委、市政府直接领导的地方国有银行、是贵州省第三大地方法人银行、是全省规模最大的地方性农村法人金融机构。

作为贵州省第一家股份制农村商业银行，本行健全“党委核心领导、董事会战略决策、高管层执行落实、监事会依法监督”的治理机制；确立了乡村振兴特色银行战略定位，构建起“一体两翼四轮驱动”的战略发展格局，科学谋划 “十四五”战略规划及执行体系。不断扎实推进改革管理创新与服务创新的有机结合，有效利用大数据、科技金融手段提升特色化金融服务的能力和水平，不断强化风险内控建设、规范内部管理流程、树立优质企业文化，加速形成特色品牌和差异化竞争优势，致力建设治理完善、管理先进、经营稳健、实力雄厚的地方性商业银行。

截至2023年10月末，资产总额1540.01亿元，较2011年末增长7.46倍；存款规模1237.29亿元，较2011年末增长6.95倍；贷款规模935.11亿元，较2011年末增长8.28倍，智能机具业务功能替代率较2011年提升78.6个百分点至83.7%。



项目背景及目标

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

国家战略

2020年中央经济工作会议把强化国家战略科技和增强产业链供应链自主可控能力作为2021年工作重点，从国家规划层面，信息技术应用创新（以下简称“信创”）已经上升为一项国家战略2023年《政府工作报告》提出增强科技创新引领作用，强化国家战略科技力量，实施一批科技创新重大项目，加强关键核心技术攻关。

行业要求

为了实现金融业关键信息基础设施安全可控，一行两会联合中央网信办、工信部推动了金融业关键基础设施国产化试点工作，要求金融行业逐步推进使用各类国产化软、硬件产品。

本行需求

持续完成存储、服务器、网络、虚拟化平台、数据库等软、硬件方面的信创选型，形成我行自有的信创基线，完成本单位硬件基座国产化替代，为业务系统国产化改造提供有力支撑。

项目背景及目标

为充分控制机房能耗和最大程度利用机柜容量，贵阳农商银行除核心系统外，所有信息系统应用节点均部署在VMware虚拟化平台上，生产虚拟机约1000余台，测试环境虚拟机约3000多台。随着信创试点工作覆盖面逐步增大，面临着服务器数量巨增，机房部署压力大的难题，若不考虑用信创虚拟化云环境来替代物理服务器，建设、维护成本会越来越高。

项目背景及目标

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

目标一

基于RoCE 国产无损
以太网络替换SAN网络

目标二

实际业务基于存储
传输无损网络的应用

目标三

降低数据中心能耗
和推动整体解决方案的落地

PART 02

探索与实践

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

2021年8月
POC测试

解决问题：
- 为我行云平台选型做POC测试；
- 存储FC-SAN替换POC测试

2021年12月
基础环境XC技术路线确定

解决问题：
- 确定紫光云作为我行私有云基座；
- 确定RoCE-SAN网络作为FC-SAN替换方案

2022年4月
XC云平台上线

解决问题：
- 完成VMware国产化替代
- 海光、鲲鹏环境国产虚拟化替换部署
- 海光环境RoCE功能上线

2022年7月
FTP业务系统上线

真实业务：
- 内部资金转移定价（FTP）、绩效管理、北塔性能监控共3个应用系统行在RoCE存储网络上，完成真实业务在RoCE存储网络上的业务验证

2022年9月
RDMA无法抓包，无法进行故障流量分析

解决问题：
- 升级到libpcap1.9解决无法抓包问题

2022年10月
通过NVMe-oF对接存储部分厂商存在兼容性问题

解决问题：
- 增加不同厂商存储多路径兼容性配置

2022年12月
NVME-oF性能提升

解决问题：
- 采用智能网卡虚拟交换机子网对接RoCE场景，性能比较低
- 对接RoCE存储协议时，对接宏杉存储时，IO存储性能慢
- 对接RoCE协议存储，主机侧测试性能有丢包；
- dd性能从350M/S提升到425m/S,性能提升约20%

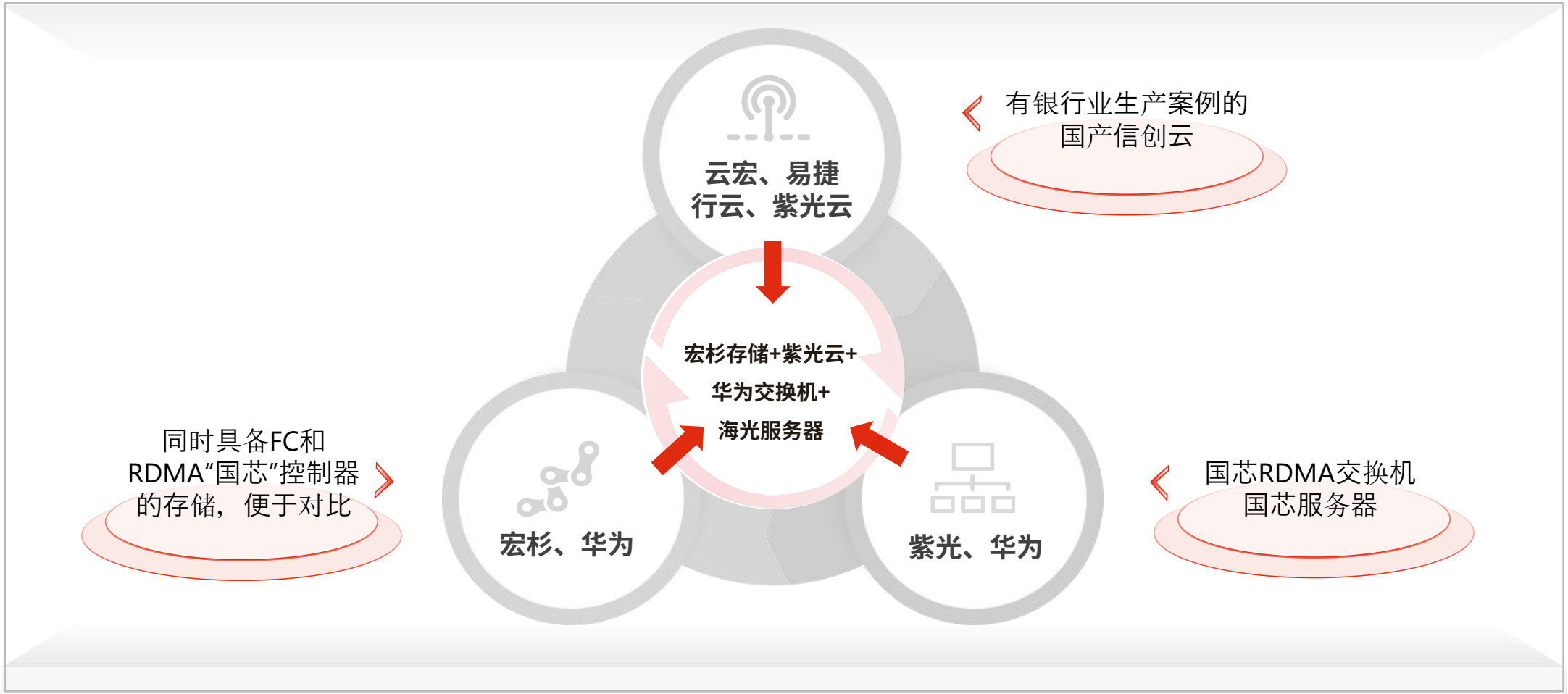
2022年12月
信贷业务系统上线

真实业务：
- 承载了超链融、信贷、电子函证、数据超市、数据铁笼、存款保险、EAST 5.0、数据治理、统信应用商店等19个系统的运行

2023年12 月
柜面、ESB系统上线

真实业务：
- 承载了柜面、ESB、信贷、电子验印、综合理财、数据超市、数据铁笼、存款保险、EAST 5.0、数据治理、统信应用商店等30个生产系统的运行

探索与实践



2021年7-10月，耗费3个多月对上述产品的十多种组合，从兼容性、功能、性能、稳定性等方面进行了详细的测试和验证

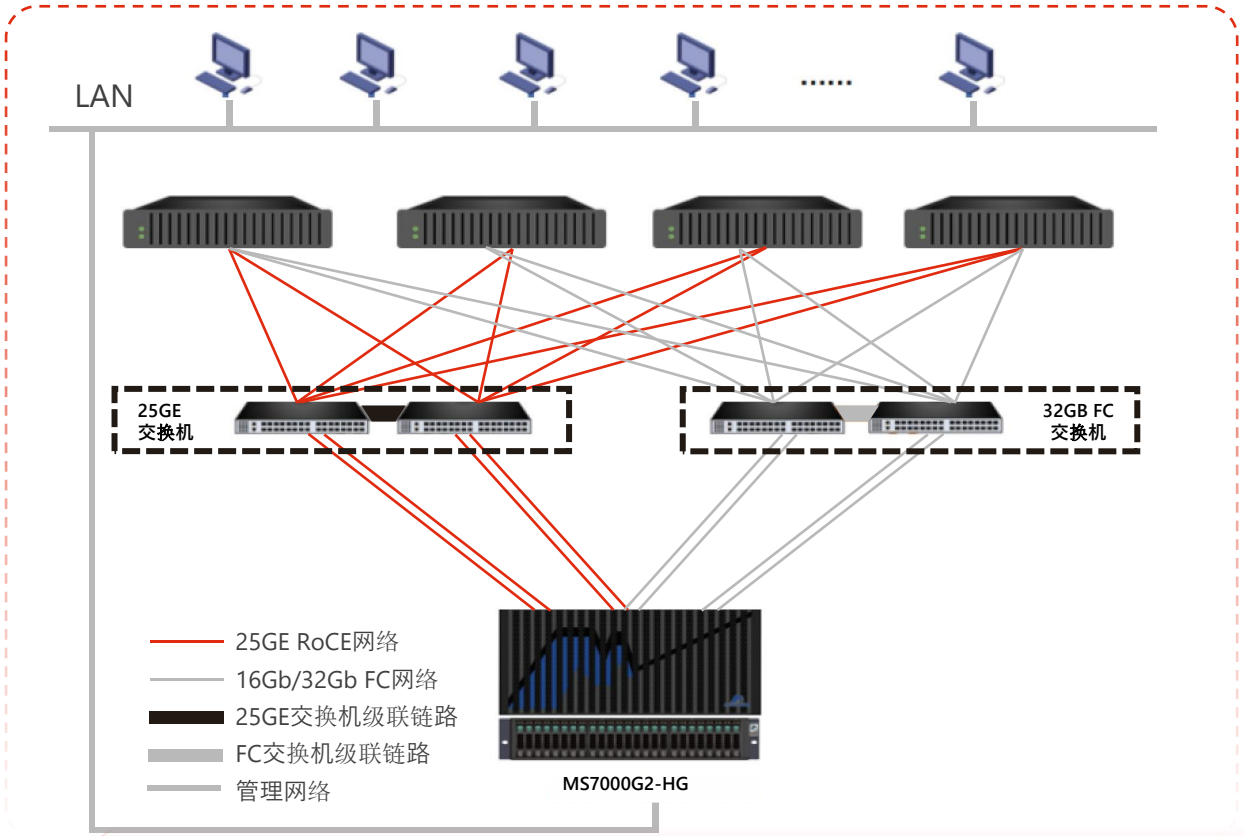
贵阳农商银行选择宏杉存储+紫光服务器（海光）+华为交换机+紫光云，结合RoCE v2协议的组网方式，来实现基础环境国产化

探索与实践

RDMA协议包含InfinibandIB、iWARP和RoCE。iWARP协议栈相比其他两者更为复杂，并且由于TCP的限制，只能支持可靠传输。Infiniband协议本身定义了一套全新的层次架构，从链路层到传输层，都无法与现有的以太网设备兼容。RoCE协议用户从以太网切换到RoCE只需要购买支持RoCE的网卡即可，其他网络设备都是兼容的。**综合考虑，我行选取了性价比最高的RoCE协议作为替代FC-SAN的最终方案。**

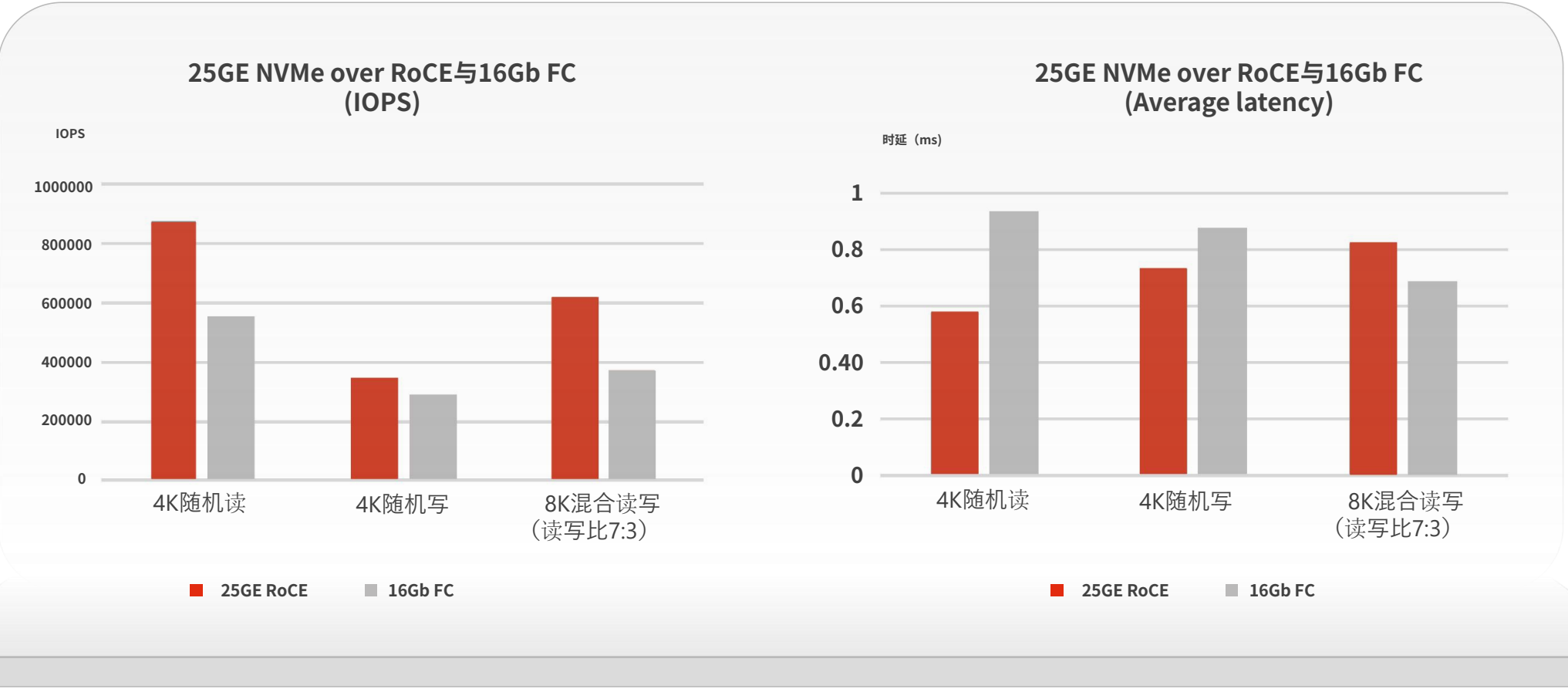
对比项	InfiniBand	iWARP	RoCE
性能	高	中	高
成本	高	中	低
稳定性	高	低	中
交换机	IB交换机	以太网交换机	以太网交换机
技术垄断	完全他国垄断	无	无

探索与实践



设备类型	配置描述	数量
存储设备	双控制器MS7000G2-AFT-HG，支持NVMe over Fabric协议	1台
	4端口25Gb RoCE I/O模块	2张
	4端口16Gb I/O模块	2张
	3.84TB NVMe接口类型硬盘模块及驱动器	21块
服务器	x86 服务器，128G内存	20
	双端口25GE RoCE卡	20
	双端口16Gbps FC HBA卡	20
网络交换机	华为 CE6860 48*25Gbps 网络交换机（含光模块）	2
FC交换机	16Gbps FC交换机	2
光纤	3m LC-LC多模光纤跳线（对）	若干
操作系统	紫光UnisLinux Release 1.1.0 (Core)	/

探索与实践



在25GE NVMe over RoCE和16Gb FC组网环境下，相同物理环境、测试模型场景中进行了性能、功能等指标的测试评估，并经过严格测试证明，基于NVMe over RoCE无损网络解决方案在IOPS、延迟等性能方面均优于16Gb FC网络

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

测试类别	用例名称	测试结果
基本功能测试	兼容性测试	通过
	存储多路径	通过
	NVMe-oF共享文件系统告警	通过
高可靠性测试	HA迁移	通过
	存储控制器故障	通过
	无损交换机故障	通过
NVMe-oF与FC功能对标测试	虚拟机部署	通过，耗时无差异
	虚拟机快照	通过，耗时无差异
	虚拟机迁移	通过，FC迁移速度优于NVMe-oF
压力测试	虚拟机IO加压	通过
	虚拟机CPU加压	通过
	虚拟机内存加压	通过
	虚拟机存储IO加压	通过
NVMe-oF与FC性能对比测试	4K随机读写场景	通过，在IOPS、延迟等性能方面NVMe-oF均优于FC
	8K随机读写场景	通过，在IOPS、延迟等性能方面NVMe-oF均优于FC
	512K读写场景	通过，在IOPS、延迟等性能方面NVMe-oF均优于FC
信贷系统测试	NVMe-oF架构性能测试	通过
	传统架构和NVMe-oF架构 性能对比测试	通过，NVMe-oF架构性能优于传统架构

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

信贷系统信创架构（ROCE存储网络）与传统架构（FC-SAN存储网络）相比，TPS有10%左右提升。

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

2021年8月	2022年1月	2022年4月	2022年8月	2023年3月	持续性工作
技术路线测试	功能测试	生产平台上线/应用上线	配置性能调优	版本升级性能优化	架构优化

测试及应用上线过程中遇到的问题处理

关键兼容性问题

- 数量：10+
- 测试过程中解决国产服务器与云平台兼容性问题10余起，最终选定海光路线。

关键功能问题

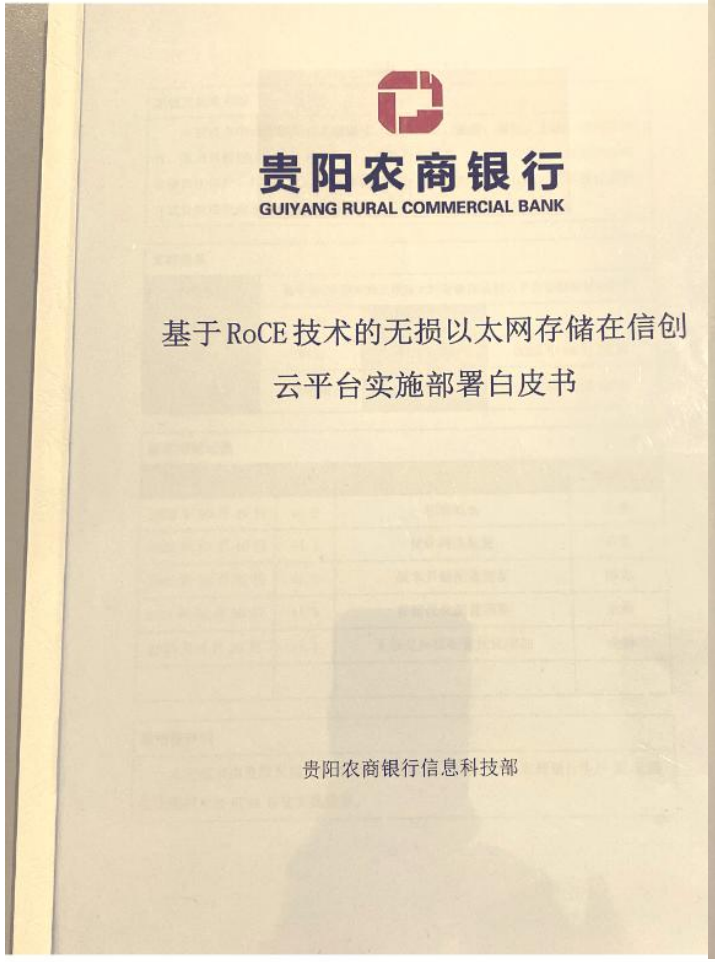
- 数量：5+
- 选定海光路线后，解决存储无法共享、部分ROCE名录不支持，存储无法连接等问题。

关键性能问题

- 数量：10+
- 应用上线后，解决IO慢、格式化存储慢、大IO性能低、有丢包错包等问题10余起。

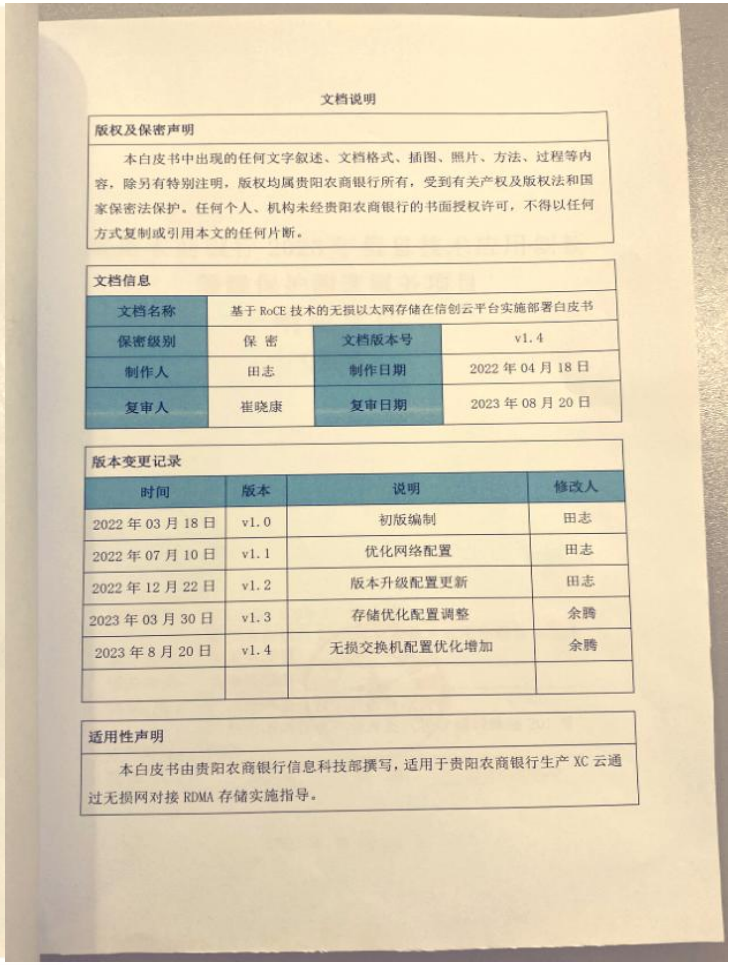
建设过程中遇到的问题详见《基于RoCE技术的无损以太网存储在信创云平台实施部署白皮书》。

探索与实践


贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

基于RoCE技术的无损以太网存储在信创
云平台实施部署白皮书

贵阳农商银行信息科技部


文档说明

版权及保密声明

本白皮书中出现的任何文字叙述、文档格式、插图、照片、方法、过程等内容，除另有特别注明，版权均属贵阳农商银行所有，受到有关产权及版权法和国家保密法保护。任何个人、机构未经贵阳农商银行的书面授权许可，不得以任何方式复制或引用本文的任何片断。

文档信息

文档名称	基于 RoCE 技术的无损以太网存储在信创云平台实施部署白皮书		
保密级别	保 密	文档版本号	v1.4
制作人	田志	制作日期	2022 年 04 月 18 日
复审人	崔晓康	复审日期	2023 年 08 月 20 日

版本变更记录

时间	版本	说明	修改人
2022 年 03 月 18 日	v1.0	初版编制	田志
2022 年 07 月 10 日	v1.1	优化网络配置	田志
2022 年 12 月 22 日	v1.2	版本升级配置更新	田志
2023 年 03 月 30 日	v1.3	存储优化配置调整	余腾
2023 年 8 月 20 日	v1.4	无损交换机配置优化增加	余腾

适用性声明

本白皮书由贵阳农商银行信息科技部撰写，适用于贵阳农商银行生产 XC 云通过无损网对接 RDMA 存储实施指导。

基于RoCE技术的无损以太网存储在信创云平台实施部署白皮书

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

序号	问题描述	造成影响	问题定位分析	解决方案
1	在虚拟机上使用vdbench工具进行IO测试显示异常	RDMA不生效	RDMA环境中主机和存储之间读写数据是通过内存直接访问，不经过操作系统和CPU。虚拟机磁盘模式在直接读写模式下数据是不经过缓存，在虚拟机磁盘处于直接读写模式下缓存被关闭RDMA网络环境中CAS平台虚拟机单端的缓存失效导致RDMA不生效。	将虚拟机磁盘模式调整为一级虚拟机缓存.
2	通过CAS平台创建虚拟机缓慢，通过WEB页面查看主机存储中共享文件系统信息缓慢。	影响虚拟机生命周期相关操作	需要优先检查整体物理网络环境是否正常。逐级排查主机到ROCE交换机物理链路、配置情况。 1、在ROCE交换机上使用disp queue-statistics int tw1/0/24 out查看连接主机和存储端口outbound方向流量通过情况发下有报文被丢弃的情况。	使用命令：qos wred ecn enable。否则会导致流量无法进入到队列中，导致报文被丢弃情况产生。
3	ROCE交换机部分命令不支持	有些ROCE交换机配置无法生效	在ROCE交换机上配置priority-flow-control dot1p 3 ingress-threshold-offset 500提示not support根据查看交换机版本命令相关信息是由于交换机版本过低导致。	需要将交换机版本升级到E8108P10解决。
4	通过存储侧提供的nvme discover命令查看主机发现存储信息。在部署实施阶段通过CVK底层无法查看到连接信息。正常情况下存储上面就能看见服务器网卡的nqn号。但是现在存储上面看不见服务器的nqn号	RDMA协议不可用	该问题是因为没有启用RDMA协议。（备注：当前前使用0760P02版本在测试时如果使用ovs的模式测试性能是不如Linuxbridge的，所以现场使用Linuxbridge模式通过网络层面与宏衫存储进行对接。）	使用Linuxbridge方法在底层启动RDMA功能和配置25G智能网卡IP地址。

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

序号	问题描述	造成影响	问题定位分析	解决方案
5	使用mellanox工具包ofed5.2对接NVMe-oF协议存储，主机侧测试性能有错包	协议错包导致报文drop及重传，影响性能	ofed5.2版本内置驱动兼容性问题。	升级ofed版本为5.4。
6	使用mellanox工具包ofed5.4对接NVMe-oF协议存储，主机侧测试性能有丢包	协议丢包导致报文drop及重传，性能降低10%左右	主机侧RoCE配置不正确。	调整RoCE优先级PFC从5变成3。
7	对接NVMe-oF协议存储时，大块IO（512k以上）读写及存储格式化等业务性能低	大块IO测试性能低50%左右	主机侧网口配置未配置egress优先级。	设置CX6网口属性egress优先级为3。
8	采用智能网卡虚拟交换机子网对接NVMe-oF场景，性能比较低	性能比较低，只有Linux方式的1/3	智能网卡策略中使用网卡vf进行RDMA通信。	调整为linux方式，通过网卡的pf进行nvmeof通信。
9	对接NVMe-oF协议存储时，对接宏杉存储时，IO存储性能慢	IO存储性能慢，降低30%左右	mtu设置为1500，导致报文处理时进行了大量的分包，导致性能降低。	主机侧、交换机侧、存储侧，调整mtu为9000，存储性能恢复正常 。

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

序号	问题描述	造成影响	问题定位分析	解决方案
10	重启主机后，NVMe-oF存储连接断开	NVMe-oF存储不可用	CX6网卡的roce_enable属性变成0，导致NVMe-oF无法建立连接。	重启主机后，启动vdpa服务时，强制设置CX6网卡的roce_enable属性为1，重启主机后，NVMe-oF存储连接正常 。
11	发现RDMA报文无法抓包分析，出故障场景下无法对流量进行分析	无法抓包分析报文，从而确认不了报文drop位置	libpcap1.9之前版本不支持RDMA抓包。	升级到libpcap1.9版本解决。
12	计算主机侧添加NVMe-oF存储失败	NVMe-oF存储不可用	配置NVMe-oF多路径失败，NVMe-oF协议多路径配置文件与其他协议不同，并且需要针对不同厂商进行区分适配。	针对NVMe-oF不同厂商存储多路径配置文件中做兼容性适配，支持NVMe-oF协议下不同厂商。
13	NVMe-oF协议连接存储，存储卷容量显示错误，与实际容量不符	NVMe-oF协议查询容量不准确	NVMe-oF协议查询容量接口返回值单位与其他协议不同。	对NVMe-oF存储查询容量接口做兼容性适配。

探索与实践

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

2022年3月
XC云虚拟化软件实部署
V7.0 D0999

解决问题：
- 完成VMware国产化替代
- 海光、鲲鹏环境国产虚拟化替换部署
- 海光环境RDMA功能替代

2022年6月
版本功能升级
V7.0 E0760P01

解决问题：
- 支持RDMA网卡性能数据上报
- 支持TrunkPort特性
- 任务台进度优化

2022年12月
性能提升升级
V7.0 E0760P02

解决问题：
- 裸金属
- Qemu热升级
- 支持云平台SRM容灾接口

2023年3月
海光版本功能升级
V7.0 E0760P03（海光）

解决问题：
- 虚拟机支持postcopy
- V主机池/集群/主机下虚拟机列表

2023年8月
鲲鹏版本升级
V7.0 E0760P03（鲲鹏）

解决问题：
- 支持华为存储RDMA版本升级

2023年12月
版本演进计划

解决问题：
- 裸金属管理
- 升级内核5.10
- 无损存储方案演进

2024年3月（计划）
版本演进计划

解决问题：
- ARM版本无损支持
- 易用性、管理性、可靠性增强

PART 03

业务功能与技术架构

业务功能与技术架构

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

应用系统

柜面	ESB	信贷	电子验印	超链融
综合理财	电子函证	数据治理	统信应用商店	.....

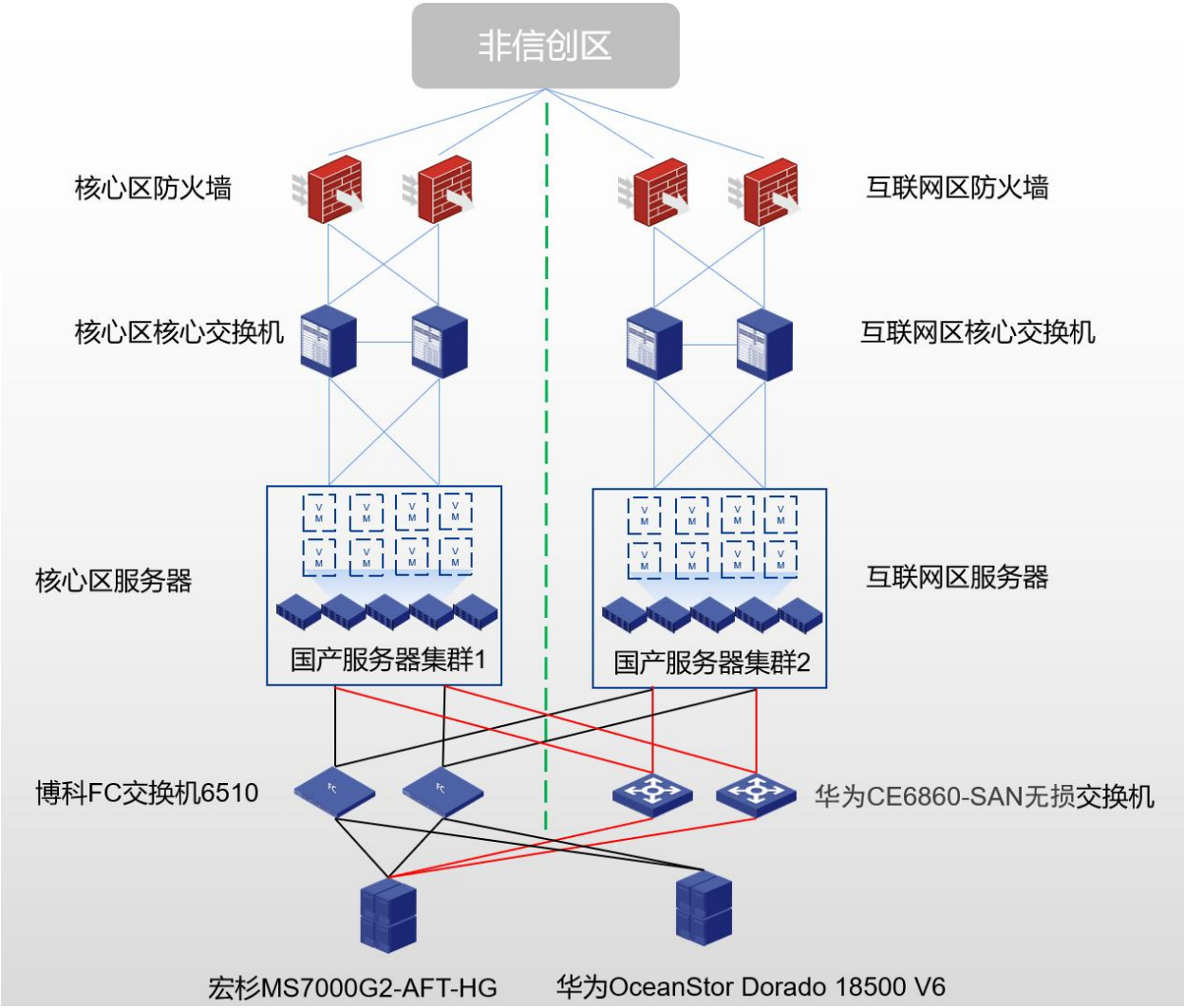
紫光云平台

CVM	云彩虹	监控告警	虚拟化安全管理	备份容灾管理			
CVK自治	CVK-Client	主机高可靠	主机监控	主机配置	虚拟机生命周期管理	虚拟机监控	告警管理
CVK内核	计算虚拟化	存储虚拟化	网络虚拟化				

硬件基础设施

国芯服务器	国芯存储	国芯RDMA交换机

业务功能与技术架构



采用NVMe协议的国产芯片宏杉存储（MS7020G2-AFT-HG）+国产芯片服务器+华为CE6860-SAN无损交换机+紫光云平台，结合RoCE v2协议的组网方式，实现了贵阳农商银行基础环境国产化替代架构。

PART 04

技术与业务创新点

技术与业务创新点

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

高性能、高可靠

同比原有FC-SAN架构，业务整体性能提升了10%，传输效率提升了近40%。另外在网络故障的场景下，交换机可快速感知主机、存储及网络的状态变化，并通报主机进行多路径切换，使端到端的切换时间小于1秒，确保了系统的高可靠。

自主可控、开放生态

采用通用以太网交换机构建，基于IP和RoCE通用存储网络协议运行,能够与业界主流的OS进行协同对接，可基于标准开放的API与更多第三方伙伴共建场景化服务，共同打造生态圈。

灵活上云

实现数据中心内SAN网络与普通业务场景LAN网络的无缝对接及混流运行，在将存储服务器云化的同时，降低了管理的复杂度和运维成本，提升了整个IT基础设施的自动化水平和业务敏捷性。

截止2023年12月，已有30个生产系统运行在该信创云平台上，其中包括ESB、柜面等多套重要信息系统的信创试点并行环境。

技术与业务创新点

通过一系列的测试和验证，形成具有代表性的金融行业基础环境国产化替代方案，并通过产金联动推进厂商之间的产品互认证，逐步丰富适合金融行业要求的产品生态圈。

PART 05

经济与社会效益

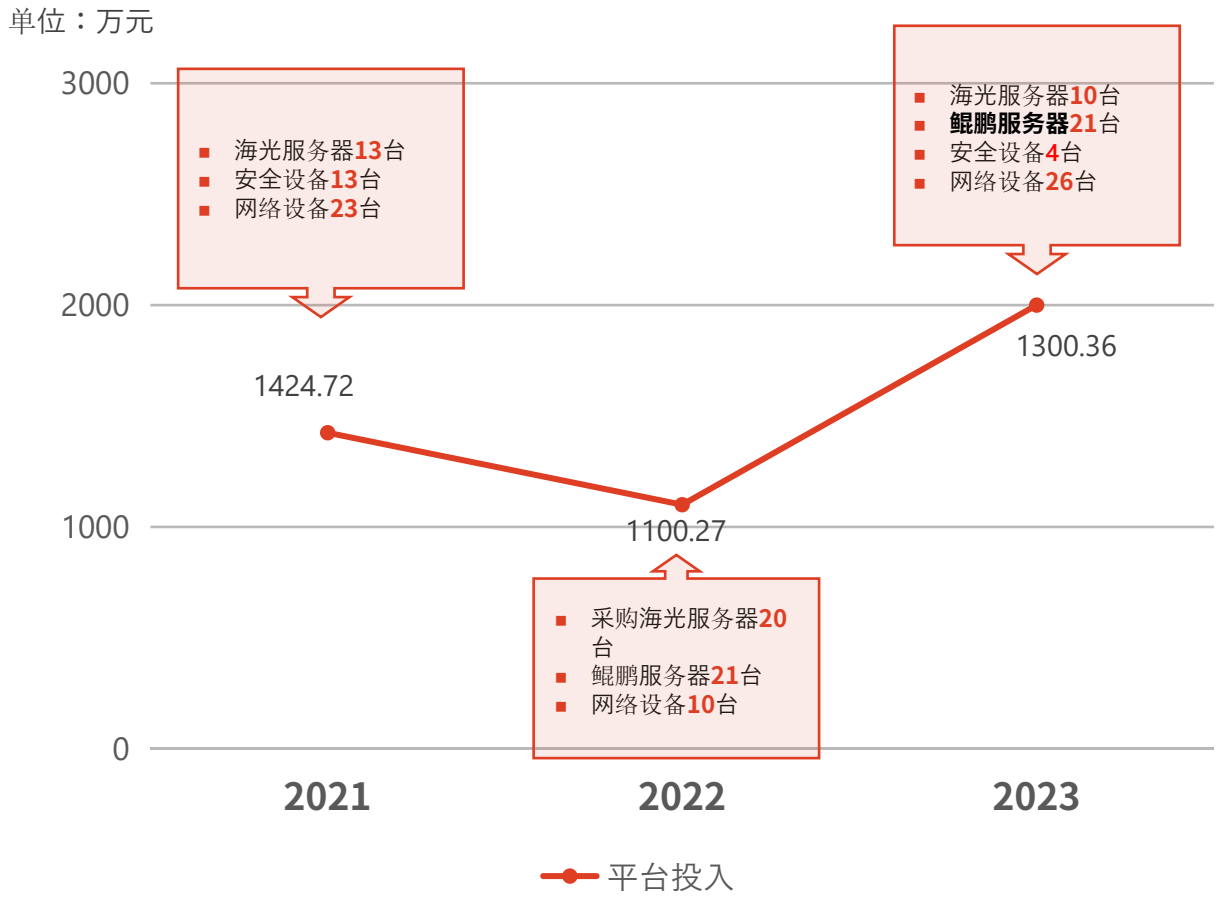
经济与社会效用

平台投入

2021年云平台信创软硬件投入1424.72万元，采购海光服务器13台，安全设备13台，网络设备23台。

2022年云平台信创软硬件投入1100.27万元，采购海光服务器20台，鲲鹏服务器21台，网络设备10台。

2023年云平台信创软硬件投入1300.36万元，采购海光服务器10台，鲲鹏服务器21台，安全设备4台，网络设备26台。

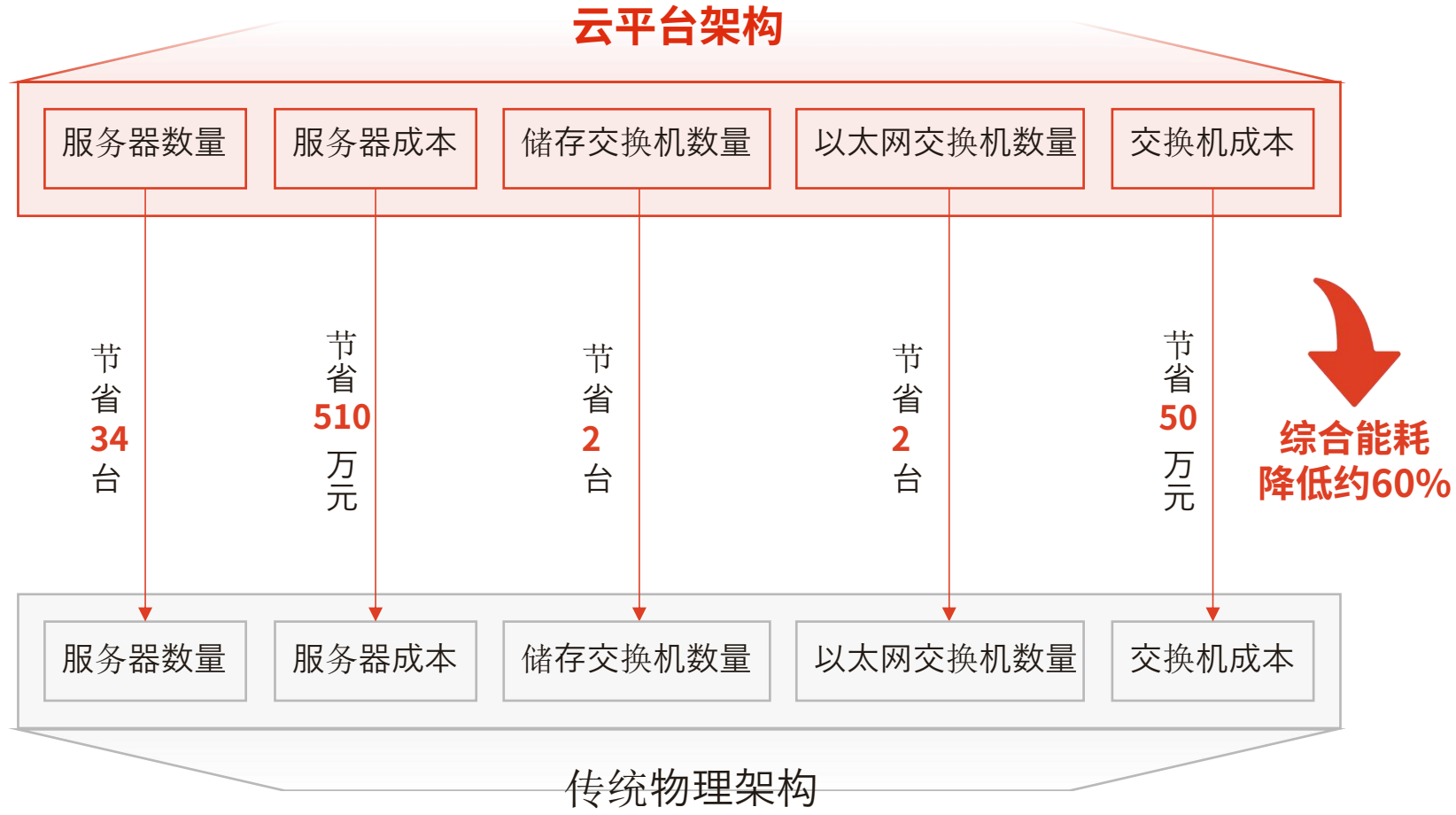


经济与社会效用

若采用传统的物理架构，仅2022年生产环境19套信创新建或信创改造的信息系统所需服务器就高达64余台，采用RoCE技术的无损以太网技术的云平台，节省了服务器34台，节省近60%，且云平台还剩大量计算资源可用。以一台服务器15万元计算，节省了采购成本510余万元。

节省存储交换机2台、以太网交换机2台，节省成本50余万元。

与传统物理架构相比，云平台架构所需物理设备数量大幅减少，综合能耗降低了约60%。



经济与社会效用

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

降低了金融机构的试错成本

通过在真实金融业务场景长期应用，各类设备和云平台功能得以完善，在相同基础设施架构下能让同业少走弯路，降低试错成本。

自主可控

保护金融业务关键数据资产的安全可控，推进关键IT基础设施建设的国产化替代工作，以确保核心技术安全可控，减少国外设备采购，提高国产化比例，实现数据安全与业务发展的动态平衡。

形成可复制的信创替代方案

架构的基础设施全部为“国芯国魂”硬件设备，能适配国产操作系统、中间件，具有经济成本优势，形成了可供同业参考的信息信创改造或建设替代方案。

PART 06

下一步规划

下一步规划

贵阳农商银行
GUIYANG RURAL COMMERCIAL BANK

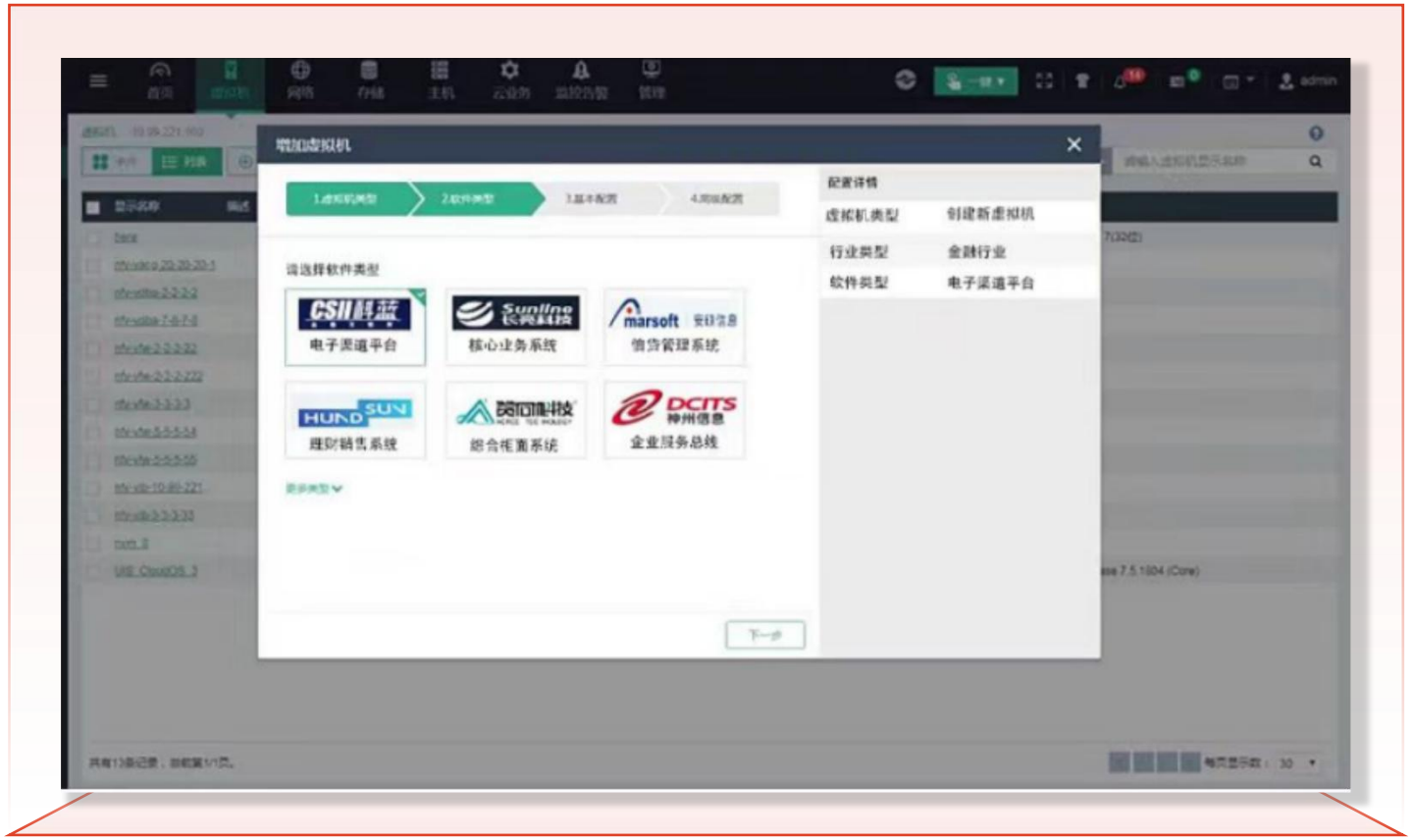
加大迁移力度
不断完善基础设施

近两年完成国产化改造和适配，在该硬件基座上线的重要信息系统有:

- 电子渠道
- 互联网开放平台
- 柜面系统
- ESB
- 数字化集中运营平台
- 理财和资管
- ……

重要信息系统从VMware虚拟化平台迁移至国产信创云平台将达到**70%**

下一步规划



利用云平台优势，形成行业系统快速部署标准和模板

与银行业务系统厂商进行联合，将部署业务系统需要的服务器资源，应用程序、中间件、数据库进行标准化、模板化

下一步规划

关注业务连续性保障、同城灾备场景的研发和实施

备份

我行使用数腾数据备份与恢复管理软件为信创云提供备份恢复服务，实现信创工作对业务连续性和稳定性要求。

灾备

探索基于RoCE技术的无损以太网异地长距离传输、数据在线压缩等功能，构建我行数据信创容灾架构。