GreatSQL

GreatSQL & NVIDIA InfiniBand NVMe SSD 性能测试实战

叶金荣

关于我

- 叶金荣
- 万里数据库开源生态负责人

背景信息 01

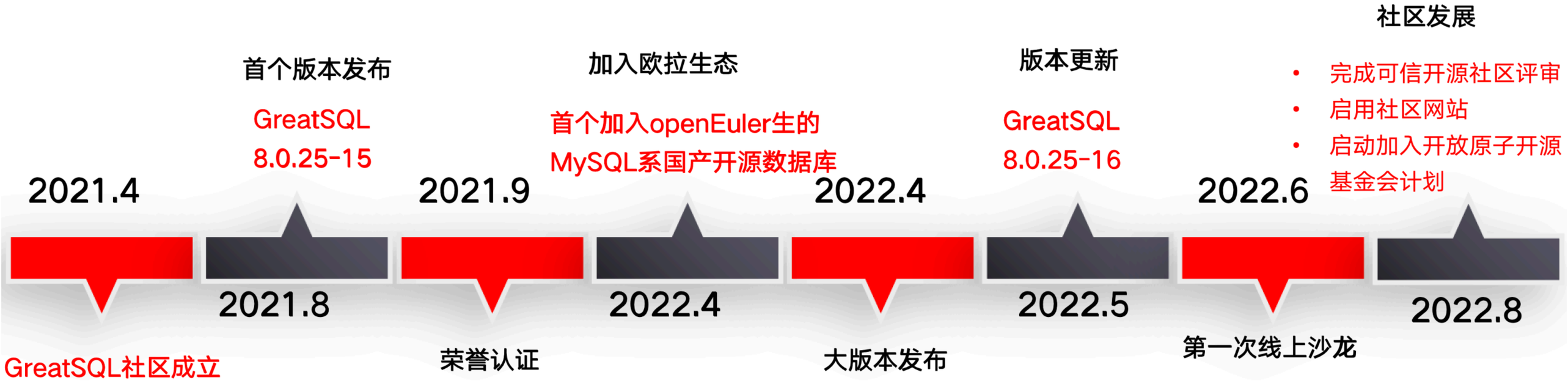
测试背景信息

- NVIDIA InfiniBand是一种被广泛使用的网络互联技术，基于IBTA(InfiniBand Trade Association)而定义的高带宽、低延时、低CPU占用率、大规模易扩展的通信技术，是世界领先的超级计算机的互连首选，为高性能计算、人工智能、云计算、存储等众多数据密集型应用提供了强大的网络性能支撑
- 通过高速的InfiniBand技术，将业务负载由单机运行转化为基于多机协作的高性能计算集群，并使高性能集群的性能得以进一步释放与优化

测试背景信息

- GreatSQL开源数据库**是适用于金融级应用的国内自主MySQL版本**，专注于提升MGR可靠性及性能，支持InnoDB并行查询等特性，可以作为MySQL或Percona Server的可选替换，用于线上生产环境，且完全免费并兼容MySQL或Percona Server
- GreatSQL开源数据库适用于金融级应用场景，具备以下几点优势
  - 地理标签功能，可提升金融级应用场景下多机房部署架构的数据可靠性
  - 仲裁节点功能，可用更低的服务器成本实现更高可用保障
  - 单主模式功能，当在MGR架构用采用单主模式时，尤其是在多机房部署架构中，可进一步提升事务性能
  - 更完善的选主机制，可自定义选主策略，降低服务不可用时长，提升RTO目标
  - InnoDB并行查询特性，满足金融行业周期性统计汇总需求，提升账单计算速度
  - 不断完善优化MGR底层机制，例如流控算法、节点加入&踢出等场景等平稳流畅，MGR运行更流畅不再频繁抖动

About GreatSQL



About GreatSQL

- 官网：https://greatsql.cn
  - FAQ：https://greatsql.cn/doc/#!&v=51_19_0
  - 手册：https://greatsql.cn/doc/
  - 视频：https://greatsql.cn/smx_course-lesson.html?op=video
- 相关资源
  - QQ群：533341697
  - 微信群：GreatSQL/MGR交流（1-3）群
- 社区用户
  - 杭州恒生芸擎网络
  - 福州靠谱云
  - 福建福富
  - 深圳华润
  - 作业帮
  - 其他...

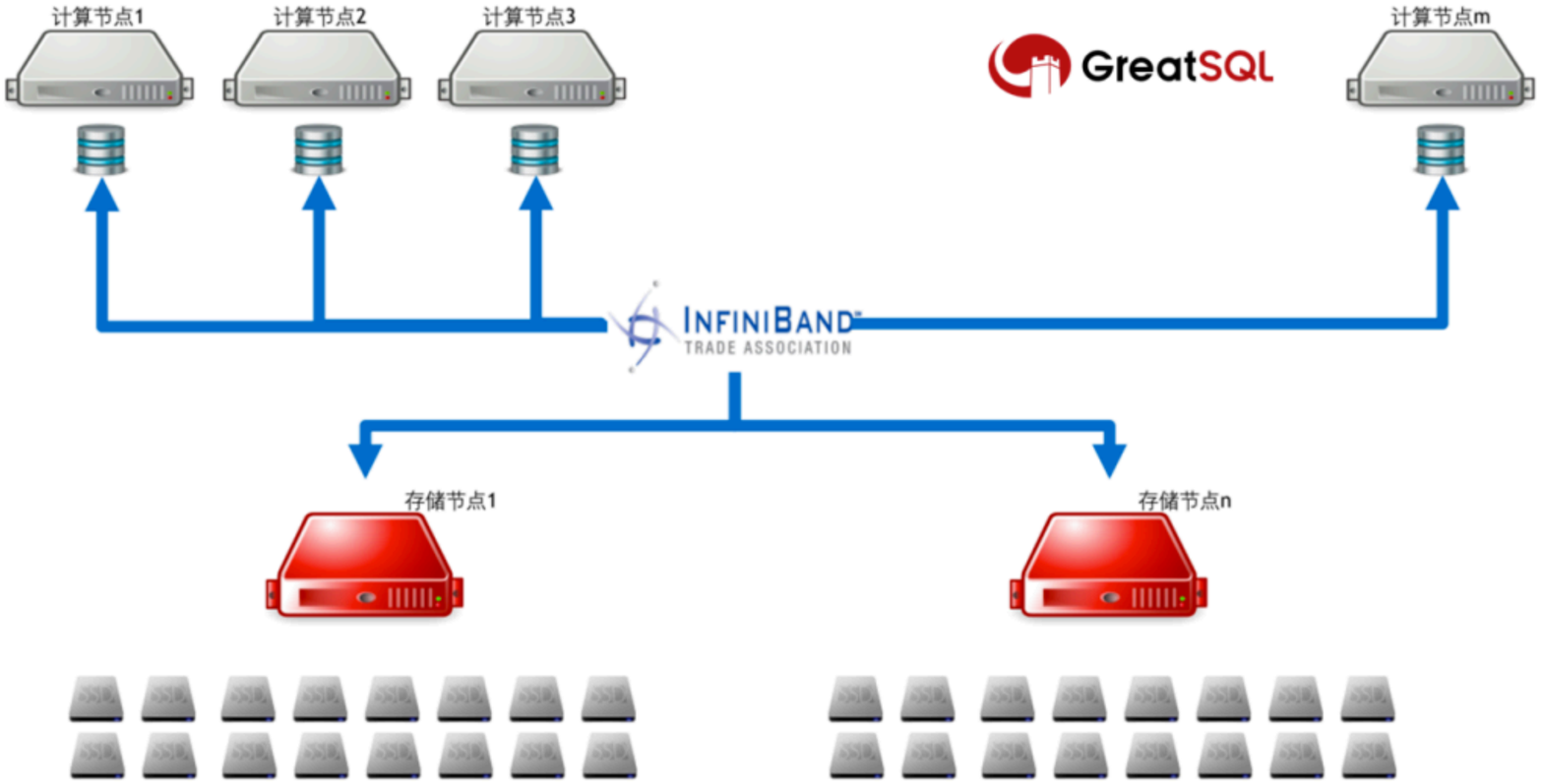
About GreatSQL

- 技术支持与服务
  - 免费技术支持
  - 在线技术交流群
  - 提供Docker镜像
  - 提供Ansible一键安装包
  - 相关文档、视频

- 相关资源
  - 代码：https://gitee.com/GreatSQL
  - 文档：https://gitee.com/GreatSQL/GreatSQL-Doc
  - 社区：微信群、QQ群、微信公众号
  - openEuler生态 https://gitee.com/src-openeuler/greatsql

测试背景信息

- 此次通过对比测试基于InfiniBand 的 NVMe SSD池化方案及本地NVMe SSD的传统方案的性能表现，评估使用基于InfiniBand的存算分离架构对分布式数据库性能的提升程度及扩展性



测试背景信息

- sysbench是一个模块化的、跨平台、多线程基准测试工具
- 主要用于评估测试各种不同系统参数下的数据库负载情况
- 支持 MySQL，PG，Oracle等数据库
- 主要有几种测试模式
  - cpu性能
  - 磁盘io性能
  - 调度程序性能
  - 内存分配及传输速度
  - POSIX线程性能
  - 数据库性能(OLTP基准测试)

性能测试 02

性能测试

- 目的：测试在InfiniBand + NVMe SSD设备上运行GreatSQL的性能表现
- 对比方案
  - InfiniBand + NVMe SSD设备
  - 本机挂NVMe SSD设备

性能测试

- 压测环境
  - CentOS 8
  - XFS（noatime,nodiratime）
  - 内存512G
  - CPU 128核

性能测试

- 压测模式
  - oltp_read_write
  - 每轮压测时长：900秒
  - 每轮压测休眠间隔：180秒
  - 共64个表
  - 每个表12500000条记录
  - 整个测试库大小约186G
  - 采用InnoDB引擎
  - 并发线程数变化： 8、16、32、64、128
  - ibp变化：47G、93G、140G、186G（约为物理数据的25%、50%、75%、100%）

性能测试

- 基准测试的几个误区
  - 压测模式脱离原定目标
  - 受到其他因素干扰
  - 不能保证测试结果的可重复性
  - 只在本地加压
  - 压测数据量小
  - 压测时间过短
  - 压测模式太少
  - 压力负载过大或过小

性能优化 03

性能优化

- 几个优化关键点
  - 采用XFS文件系统
  - 采用jemalloc代替glibc malloc
- 数据库参数选项调优
  - innodb_io_capacity
  - innodb_thread_concurrency
  - innodb_log_file_size
  - innodb_doublewrite

性能优化

- 当ibp不足以覆盖全部物理数据时
  1) GreatSQL 8.x性能远高于GreatSQL 5.7（含MySQL 5.7）
  3) ibp越大，GreatSQL 5.7比MySQL 5.7性能提升越多
  4) ibp越大，GreatSQL 5.7和GreatSQL 8.0性能越接近
- 当ibp基本可以覆盖全部物理数据时
  1) GreatSQL 5.7的性能看起来更好，GreatSQL 8.0紧随其后
  2) GreatSQL性能总是要比MySQL更好

性能优化

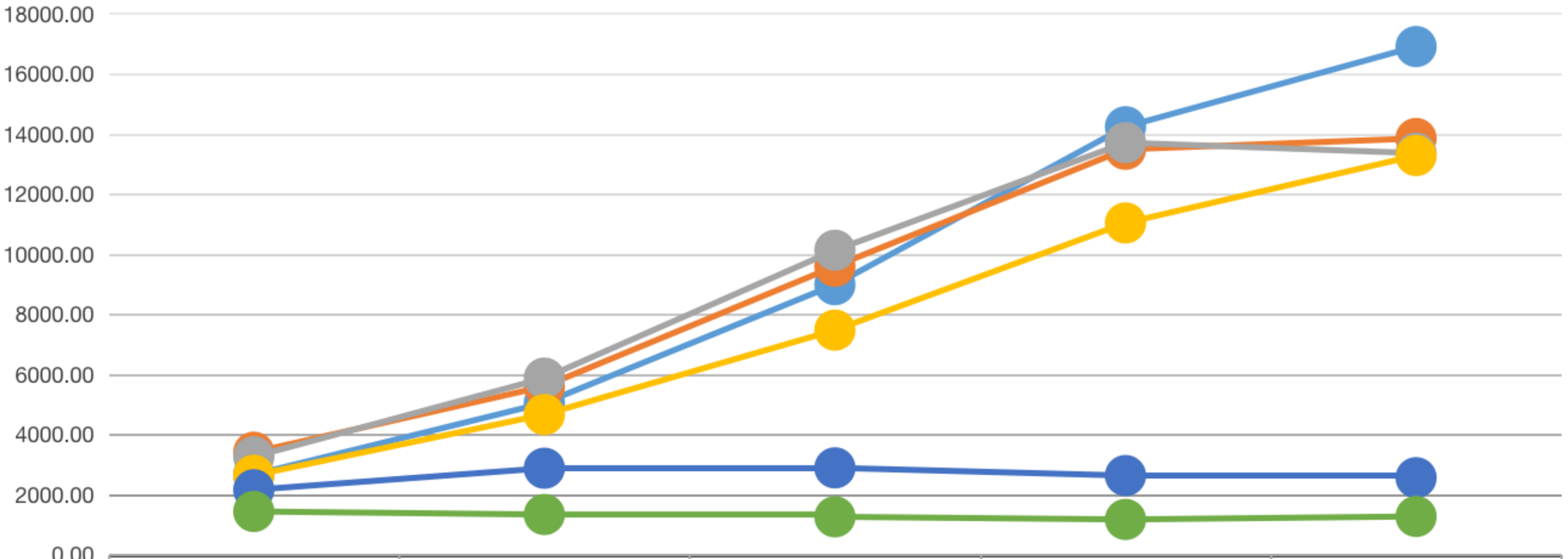
- 总结

1) IB NVMe SSD 相比 本地NVMe SSD的性能要更好，也更稳定一些

2) 当物理内存不足以覆盖业务数据时（生成环境中这种情况很常见），如果单靠增加物理内存以提升数据库性能可能从性价比角度看并不划算

3) 换个思路，提升本地物理I/O设备性能，现在NVMe SSD的性能可以跑到很高

性能优化



	8th	16th	32th	64th	128th
IB NVMe(GreatSQL 57)	2664.68	5046.12	8983.41	14238.59	16906.96
NVMe SSD(GreatSQL 8)	3407.54	5597.82	9594.34	13488.46	13847.41
IB NVMe(GreatSQL 8)	3258.74	5874.58	10128.21	13708.89	13360.97
NVMe SSD(GreatSQL 57)	2640.51	4661.94	7470.80	11037.46	13279.83
IB NVMe(MySQL 8)	2157.62	2875.01	2887.27	2631.05	2558.75
NVMe SSD(MySQL 57)	1435.69	1335.11	1258.61	1170.51	1270.08

结语

- 采用了InfiniBand池化方案数据库性能在不同场景中性能都有不同程度的明显提升
- 尤其在高并发场景下，表现突出
- 万里数据库将联合NVIDIA在万里数据库GreatDB集中式及分布式数据库产品中，探索更多基于InfiniBand在数据库中的结合点和创新点，基于NVIDIA InfiniBand打造数据库+网络软硬一体化联合解决方案，为用户创造更多价值

GreatSQL，更流畅

GreatSQL

成为中国广受欢迎的
开源数据库