# 中国科学技术大学-OPPO智能计算联合实验室

**面向CubeFS的元数据管理优化**

**合作项目任务书**

|  |
| --- |
| **甲方：OPPO广东移动通信有限公司** |
| **乙方：中国科学技术大学** |

**签订地点：中国广东省东莞市长安镇**

甲方： OPPO广东移动通信有限公司

住所：广东省东莞市长安镇乌沙海滨路18号

邮政编码：523860

法定代表人：刘波

项目联系人： 何小春

联系电话：13480134511

电子邮箱: hexiaochun@oppo.com

乙方：中国科学技术大学

地址：安徽省合肥市黄山路443号中国科学技术大学

邮政编码：230022

法定代表人：包信和

项目联系人： 许胤龙

联系电话：13696504559

电子邮箱: ylxu@ustc.edu.cn

根据《中华人民共和国民法典》及其他有关法律法规，甲乙双方在平等自愿、充分地表达各自意愿的基础上，经充分友好协商，在“中国科大-OPPO 智能计算联合实验室”合作框架协议下，就共同开展 面向CubeFS的元数据管理优化 项目（下称“本项目”）达成如下协议。

# 第一部分 合作开发内容条款

1. **定义**

“关联方”:是指任何一方现在或将来直接或间接控制、受其控制或与其共同控制的任何法人团体、公司、事务所或实体。控制，是指有权决定一方财务和经营政策，并能据以从该方的经营活动中获取利益。共同控制，是指按照协议约定对某项经济活动所共有的控制，仅在与该项经济活动相关的重要财务和经营决策需要分享控制权的投资方一致同意时存在。重大影响，是指对一方的财务和经营政策有参与决策的权力，但并不能够控制或者与其他方一起共同控制这些政策的制定。

“开发成果”是指乙方根据本协议约定负责交付给甲方的全部成果，包括但不限于技术方案和技术样品。

“技术方案”是指乙方根据本协议的约定研发的技术设计。

“技术样品”是指乙方为实际实现其研发的技术方案而开发的样品，可以是软件、硬件或其结合体。

“背景知识产权”是指本协议一方及其关联方在本协议生效之前就已拥有或控制的知识产权或独立于本协议之外拥有或控制的知识产权。

“后续成果”是指甲乙双方于本项目结束后三年内，在开发成果的基础上完成的技术成果。

“工作任务书”是指包含对开发成果、技术方案以及技术样品等交付物和服务进行描述的任何文件，包括对交付物和服务的要求、技术规格或时间表等。

1. **合作内容**

乙方应按照工作任务书的约定展开研发工作，主要包括以下方面。

研发内容：见附件1

研发成果交付

乙方应按照本协议和附件1工作任务书中约定的条件，包括但不限于交付时间、内容、形式等，将研发成果交付给甲方。

专利申请

双方应积极配合对方，共同做好专利申请工作。包括但不限于：

1）乙方应就开发成果的技术方案撰写技术交底书或者技术报告并交付给甲方；

2）乙方应协助甲方就技术方案完成专利申请及专利授权工作，如提供发明人联系方式、签署发明人声明、发明人转让协议等相关法律文件；因乙方未及时履行本条约定的协助义务而产生的额外费用以及导致的所有损失均应由乙方承担，甲方有权要求乙方另行支付或在协议总金额未支付部分中扣除前述费用和损失；

3）对于甲方（包括甲方委托的专利代理人）就技术方案提出的有关问题，乙方应当在十个工作日内答复，可通过电话、传真和电邮等形式进行。甲方（包括甲方委托的专利代理人）认为有必要进行现场技术沟通的，乙方应予以配合。

4）乙方应当对本项目的所有研究资料、数据、材料等信息进行全面和准确的记录和管理，及时响应甲方的信息需求，并在本项目结束后将所有数据移交甲方。

5）专利申请由甲乙双方共同署名，按照贡献占比确定申请人、发明人排序。

1. **费用**

甲方将按照附件1的规定向乙方支付相应费用。

# 第二部分 知识产权条款

1. **知识产权归属**

双方的背景知识产权归各自所有，如本协议研究成果的实施中不可避免必须要使用到一方的背景知识产权，双方另行协商一方具有优先受让的权利。

本协议下产生的全部研究成果及其知识产权，包括但不限于申请专利的权利、专利申请权、专利权、专有技术使用权、商标权、著作权、商业秘密等，均归双方共同所有。

双方及其关联方均有权实施、许可研究成果及其知识产权，并且一方及其关联方自行实施该知识产权所获得的收益无需分配给对方。乙方在协议有效期及协议终止后2年内，不得将该非公开的研究成果及其知识产权转让、许可给与甲方有竞争关系的第三方及其关联方，且不得与该第三方及其关联方合作实施该技术成果。与甲方存在竞争关系的第三方包括但不限于以下品牌的通信技术企业：Qualcomm、Intel、Samsung、LG、Apple、Motorola、华为、中兴、爱立信、诺基亚、VIVO、小米等。

双方理解并同意，乙方有权基于学术或科研目的就研究成果发表文章，所发表的文章须注明受OPPO研究基金资助。

一方对研究成果以任何形式公开前，应当主动提交给另一方审核并获得同意。审核的内容包括但不限于：

1)该公开是否会破坏相关专利申请的授权前景；

2)该公开是否会侵害一方或其关联方的合法权益。

甲方基于本协议产生的后续成果，知识产权归甲方享有；乙方基于本协议产生的后续成果，知识产权归乙方享有。乙方同意当其转让、许可此等后续成果包含的知识产权时，应提前至少6个月告知甲方，甲方及其关联方具有同等条件下优先受让和获得许可的权利。

对于研究成果的保护方式由双方协商决定，对于将研究成果申请专利的，乙方及项目参与成员有义务提供申请专利过程中的必要协助，包括但不限于撰写交底书、文件的签署（例如美国专利申请的assignment、declaration，印度专利申请的assignment，优先权转让等）、申请及官方审查过程中对技术内容的澄清、与甲方委托的专利代理机构及其代理人的沟通等。

乙方保证其在本协议项下的行为及其所提供的研究成果没有侵犯任何第三方的知识产权（包括但不限于著作权、专利权等）及其他合法权益，没有违反中国及合作地域内的任何法律、法规或政策。如乙方向甲方提供的研究成果中包含第三方知识产权的，乙方应确保已获得相关权利人的合法授权及许可，并保证甲方及甲方关联方依据本协议持有和使用项目成果不会侵犯任何第三方的知识产权或其他合法权益。乙方进一步承诺如因乙方在本协议项下的行为及/或其提供给甲方的研究成果引起的任何纠纷，由双方共同协商处理，如确由乙方原因导致则由乙方负责处理并承担所有责任，如给甲方造成损失（包括但不限于诉讼费、律师费），乙方应予以赔偿。

乙方保证向甲方提供的研究成果无病毒、木马，隐形计费或其它后门程序，如发现乙方在程序中故意加插和软件功能无关的程序、木马、吸费等后门程序或预留一些危害软件安全的漏洞，甲方有权要求乙方完善软件，并赔偿由此给甲方和用户造成的直接损失，同时甲方有权直接终止合作。

本知识产权条款在协议期限届满后仍然长期有效。本协议或协议项下其它条款无效、解除、提前终止的，不影响本条款的有效性。

# 第三部分 保密与保证条款

1. **保密**

双方确定因履行本合同应遵守的保密义务如下：

本合同所称的“保密信息”，是指与本合同项目有关的、与合同双方有关的或与第三方有关的，以书面、口头、图形、电子或其他任何形式表现的技术信息、商业信息及其他信息，其中：

1. 技术信息包括但不限于：技术方案、设计要求、服务内容、实现方法、运作流程、技术指标、软件系统、数据、草案、运行环境、作业平台、测试结果、图纸、模型、样品、使用手册、技术文档、仪器设备、涉及技术秘密的业务函电等。
2. 商业信息包括但不限于：客户名称、客户地址及联系方式、需求信息、营销计划、采购资料、定价政策、价格信息、进货渠道、产销策略、招投标中的标底及标书内容、项目组人员构成、费用预算、利润情况及不公开的财务资料等。
3. 其他信息是指甲、乙双方依照法律的规定和有关协议的约定、要求双方承担保密义务的其他事项。

甲乙双方仅能为完成本协议项目所定目标而使用保密信息。一方不经另一方的同意，不得直接或间接的将保密信息泄露、披露或传授给任何与本协议项目无关的人员或第三方。甲、乙双方应采取适当的保密措施，妥善保管保密信息，以保障保密信息不被泄漏，具体保密措施包括但不限于制定保密制度、与有关人员签订保密合同等。

保密期限为本协议生效之日起至项目验收合格后五年时间内。保密期满后，如双方认为有继续保密的必要，可签订新的保密合同。其中，技术信息的保密期限一般为5年，特殊的例外。

下列信息不视为保密信息，双方另有约定的除外：

1. 非因一方的过错，另一方披露时已可公开获取的信息；
2. 通过有权第三方披露而为接收方所合法接触到的信息。

保密信息接收方根据有效的司法、仲裁机构、政府机关及其分支机构或其他监管机构的书面要求而披露信息的，应在不违反法律规定的前提下尽力降低对保密信息披露方的影响，包括但不限于事先通知保密信息披露方、协助保密信息披露方采取相应的措施。

本协议部分无效、变更或终止的，不影响保密条款的效力。一方违反保密义务、泄露保密信息的，甲乙双方按《共建“中国科学技术大学-OPPO智能计算联合实验室”合作框架协议》执行。

1. **诚信保证**

双方陈述并保证：（a）严格遵守向对方做出的承诺、双方签署的合同、协议和备忘录，按时、保质、保量提供开发成果及/或服务；（b）自觉遵守有关法律法规、规章制度中关于公平交易、廉洁自律、反对腐败的相关规定。

# 第四部分 协议终止与违约条款

1. **期限和终止**

本协议自各方授权代表签字盖章当日起生效，至双方履行完毕在本协议（含附件相关工作任务书）下的义务之日起终止。

若发生以下任一情形，各方均有权向对方发出书面通知以提前终止本协议：（一）对方实质性违反本协议的约定，且自本方发出指出其违约的书面通知起三十日内亦未弥补此违约行为；或（二）对方破产，或已进入破产或其他类似性质的程序；或（三）对方决定解散或清算；或（四）对方实质性股权或资产权属变更，包括被接管或与其他单位合并。

双方理解并同意，甲方有权以书面通知的形式终止本协议。在该情形下，已通过验收的开发成果，甲方应按照本协议约定的付款条件履行付款义务；未通过验收的部分，相关费用甲方不再支付。

甲方延迟付款或不提供相应支撑超过30日时，乙方有权单方提出终止合作或解除协议，但已付款项不予返还，上述相应支撑以任务书中规定内容为限。

1. **违约责任**

一方违反本协议约定的，应及时采取合理措施纠正其违约行为。乙方违反本协议约定，甲方有权要求乙方退回其已支付的所有费用；乙方违约给甲方造成损失的，应赔偿甲方的全部损失。

乙方在根据协议约定向甲方支付赔偿款、违约金或滞纳金时，其总额最高不超过乙方依据协议已收取的金额。

1. **争议的解决办法：**

争议的解决办法，甲乙双方约定按《共建“中国科学技术大学 - OPPO智能计算联合实验室”合作框架协议》执行。

1. **不可抗力**

如果发生以下不可抗力事件造成协议无法继续履行或延误，甲方或乙方不对另一方负有赔偿责任。

* 1. 任何政府行为或法规；
  2. 战争、国内暴乱、水灾、火灾、台风和地震或者其他灾难或类似事件；
  3. 双方认可的其他不可抗拒事件。

但是，该方应采取所有可能的行动减少由于上述原因所造成的损失，在5个工作日内通知对方，并提供有关机构出具的证明。双方协商并书面确定是否继续履行协议或解除协议。

1. **其他一般性条款**

各方均在此声明，各方拥有足够的资质和能力签署和履行本协议。

各方在履行本协议的过程中，应遵从所有适用的法律。

各方发出的关于本协议的任何通知应以书面形式做出，递交形式包括：

* 1. 航空挂号信，邮寄日起5日视为收到；
  2. 电报、传真、电子邮件，且航空挂号信确认，确认信邮寄日视为收到；
  3. 专人递交，递交日视为收到。

工作任务书中制定双方项目负责人，负责本协议履行过程中双方的联络与协调。

本协议所载任何内容不应被解释为在甲乙双方间创设合资、合伙、代理或任何其它本协议目的以外的关系。

所有附件均构成本协议的有效组成部分。本协议反映了双方对本协议所述主题的全部协定，并代替所有之前关于本协议所述主题的任何协议及以往惯例。

未经双方书面同意，本协议不得修改。

未经双方事先书面同意，任一方不得转让其在本协议下的任何权利或义务。

若本协议中部分条款因任何原因而被认定无效，此种无效条款并不影响其他条款的有效性，且此种无效条款应自始视为不存在。

一方未能或延迟行使其在本协议下的权利，不能解释为其弃权。

本协议中所有标题仅为方便阅读所设置，并不影响本协议任何部分的含义或解释。

本协议一式 肆 份，自双方签字盖章之日起生效，甲乙双方各执 贰 份，均具有同等法律效力。

双方授权代表签署如下。

**甲方：OPPO广东移动通信有限公司 乙方：中国科学技术大学**

合作项目代表签字： 合作老师签字：

甲方授权代表签字： 乙方授权代表签字：

日期： 日期：

# 附件一：合作项目任务书

## 项目介绍

### 项目名称

面向CubeFS的元数据管理优化

### 项目背景

CubeFS是一个提供了文件及对象存储能力的云原生分布式存储，CubeFS由元数据子系统，数据子系统和资源管理节点以及对象网关组成，可以通过POSIX/HDFS/S3接口访问存储的数据。其中元数据子系统由多个Meta Node节点组成，多个元数据分片（Meta Partition）和Raft实例（基于Multi-Raft复制协议）组成，每个元数据分片存储一个Inode范围的元数据，其中包含两个内存B-Tree：inode B-Tree与dentry B-Tree。

CubeFS当前的元数据管理方式为磁盘持久化+内存全量缓存的模式，内存开销很大。目前分布式文件系统元数据管理的主流方式是使用分布式键值存储，这种方式可以实现较好的负载均衡和可扩展性，且内存开销较小，可以用分布式键值存储来优化CubeFS当前的元数据管理系统，从而减少元数据的内存开销。

### 研究内容

本项目拟研究面向CubeFS的元数据管理优化，主要研究内容分为三个部分，分别是元数据的键值存储结构研究、基于元数据键值存储结构的键值分离优化研究以及基于元数据键值存储结构的冷热感知缓存研究。最终为CubeFS构建低内存开销且高性能的元数据管理结构。

**研究内容1：元数据的键值存储结构研究**

在CubeFS中，元数据分为存储文件信息的inode索引节点与存储目录项信息的dentry节点，它们都以B树的结构全量存储在内存中。为了优化存储结构，研究以键值（Key-Value）方式存储索引节点与目录项节点的存储结构，分别为键和值设计对应的具体结构，使其可以持久化在SSD/HDD中，减少元数据的内存使用开销。研究键值化元数据在不同机器间的分布方式，结合Multi-Raft协议，保证元数据在多个服务器同步。

**研究内容2：基于元数据键值存储结构的键值分离优化研究**

为减少CubeFS中元数据管理的大量内存开销，研究基于键值结构的键值分离优化，通过将值存放在持久化设备中，减少存储元数据所需LSM-Tree的大小，从而可以将LSM-Tree结构缓存在内存中，在减小内存开销的同时达到更高的查询性能。

**研究内容3：基于元数据键值存储结构的冷热感知缓存研究**

基于键值分离的元数据键值存储结构，研究元数据信息的冷热识别策略，从而实时感知元数据写入与读取的访问热点。根据读写访问热点，研究元数据键值存储结构的冷热感知缓存方案，可以进一步探索将热点访问对应的value缓存到内存中，设计缓存结构、缓存策略，从而有效利用内存设备的读取速度优势，提升键值化后元数据子系统的查询性能。

### 技术发展趋势

分布式文件系统在很多大规模数据处理系统中是十分重要的基础设施，在数据量越来越大的趋势下，分布式文件系统的元数据管理成为了可扩展性瓶颈 [1]。早期的分布式文件系统采用单个元数据节点的设计，如GFS[2]、HDFS[3]，当并发访问量增加，超过单个元数据节点的性能上限时，元数据节点就会成为系统性能瓶颈。通常一个分布式文件系统会服务大量用户，且用户量和并发访问量呈越来越高的趋势，因此单个元数据节点的设计逐渐被抛弃。

为提高可扩展性，现在的分布式文件系统通常使用一个集群来管理所有元数据，这种设计下，如何将元数据分配到集群中的不同节点上成为了新的问题。有一些系统采用静态划分的方式，如NFS[4]、Sprite[5]、MapR[6]，每个元数据节点上拥有的文件范围都是固定的，这种划分方式会导致元数据集群负载不均衡，某些特定的节点可能会收到过多的访问从而成为性能瓶颈。Ceph[7]提出了一种动态划分元数据的策略，Ceph会统计一段时间内目录树中每棵子树被访问的次数，据此来判断数据的冷热，把热的子树拆分成更小的子树分散到不同的元数据节点上，从而实现负载均衡。但这种方式有较大的数据迁移开销，当数据访问模式改变较为频繁时，数据冷热频繁变化，数据就会不断迁移，导致系统整体性能下降。

使用分布式键值存储管理元数据是当前的发展趋势，也是目前业界的主流做法[8][9]。Giraffa[10]用动态的分布式元数据管理代替了HDFS原先的单个元数据服务器，将元数据存储在分布式键值存储HBase中，利用HBase来做元数据的划分与节点间的负载均衡。IndexFS[11]使用一个集群管理元数据，集群中每个节点有一个LevelDB实例用于存储元数据，其中LevelDB是一个经典的键值存储系统。为了保证负载均衡，IndexFS以目录为粒度划分元数据，对于包含很多子目录和文件的较大目录，IndexFS会将其动态拆分，分散到不同的节点上。CalvinFS[12]专注于优化地理上分布式的文件系统，它使用一个支持事务的分布式键值存储来管理元数据。CalvinFS使用完整文件路径的哈希值来划分元数据，这种方式与范围划分和子树划分相比能提供更好的负载均衡性，同时，为保证一致性，跨多个文件或目录的元数据操作会被转化为分布式事务。

Facebook Tectonic[8]使用分布式键值存储ZippyDB管理元数据，ZippyDB的存储底座为多个RocksDB实例。Tectonic对元数据进行了细粒度的划分，分成了Name、File、Block三层，分别存储文件名、文件和数据块的信息，并使用哈希分别对每层进行划分以实现负载均衡。InfiniFS[14]使用键值存储系统RocksDB在每个元数据节点上存储元数据，根据目录信息的哈希值进行节点间元数据的分配以实现负载均衡。InfiniFS将目录元数据信息做了解耦，将其分成了目录本身的元数据和目录内容的元数据，将目录内容的元数据和其子目录本身的元数据打包在一起存储，这样可以保证元数据访问的局部性。

### 项目价值

探索使用键值存储来进行CubeFS的元数据管理，可以解决当前CubeFS元数据内存开销较大的问题，降低CubeFS的运行成本。使用键值存储作为分布式文件系统的主流元数据管理方式，近年来有许多工作在优化相关系统性能，CubeFS使用键值存储管理元数据后，可以应用这些前沿工作的优化方案来提升性能。

使用键值存储管理元数据后，大量元数据存储在外存中，若不进行优化，相比原先全内存的管理方式，元数据访问性能必然会下降。我们将探索利于做缓存的键值存储实现方式和高效的缓存机制，优化存储在外存中的元数据的读写性能，从而保障新的元数据管理系统的整体性能。

### 研究目标

在CubeFS中实现基于键值存储的元数据管理系统，将元数据放到外存中进行管理，降低元数据的内存开销。实现缓存友好的键值分离存储结构，将关键元数据缓存在内存中，减少访问所需的IO次数，探索高效的缓存机制，对元数据进行冷热识别，对热数据做缓存，最终实现使用较少内存达到较好读写性能的目标。

### 技术路线

**（1）元数据的键值存储结构研究**

在CubeFS中，存储文件元数据的inode节点与存储目录项的dentry节点均以B树的形式全量存储在内存中，引入了大量的内存开销。为了减少索引的大量内存占用，元数据采用键值存储可以有效将索引数据扁平化，从而有效支持索引数据落盘。

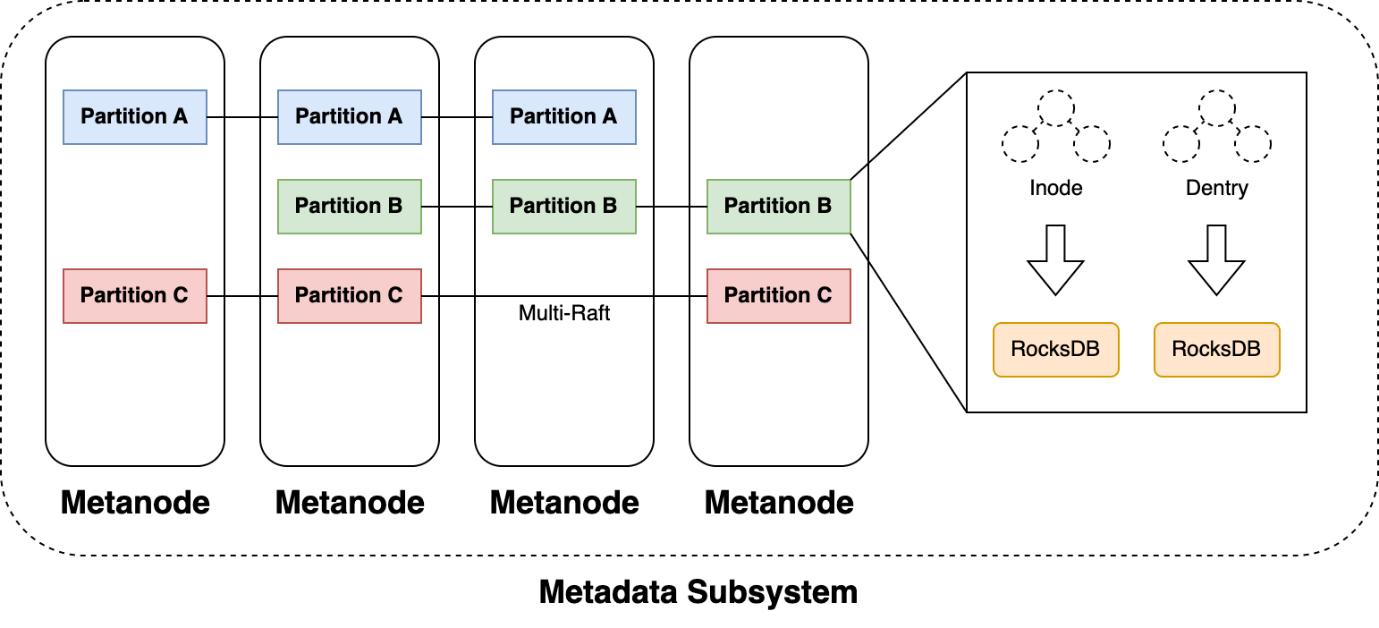


图1:元数据的键值存储结构图

元数据键值化存储的数据结构如图1所示。在CubeFS中，对于inode tree，它的键是Inode ID，值为类型、属性信息、存储位置等具体文件数据；而dentry tree中的键是parent inode ID与文件名的组合，值为该文件的inode ID与文件类型。键值化的元数据保留原先的数据结构，为inode tree与dentry tree分别使用一个RocksDB实例进行存储。RocksDB采用磁盘设备友好的SSTable结构进行下层存储。因此，采用RocksDB对元数据进行存储可以在保留相对高效的索引读写性能的同时，大大减少存储索引的内存占用。

对于元数据分片，CubeFS采用Multi-raft协议在各个服务器间达成共识，保证节点间的强一致性，并通过多组元数据分片实现相对负载均衡。采用RocksDB替换树形结构存储索引信息后，为保证一致性，需要为元数据节点的各个操作对接Raft协议对应的共识操作。同时，结合工作负载，需要结合RocksDB的性能与内存开销特点，调整各分片的大小从而达到更好的读写性能。

**（2）****基于元数据键值存储结构的键值分离优化**

所有元数据都存储在LSM-Tree中时，LSM-Tree的空间占用会很大，无法全部缓存到内存中，若访问到没有缓存的数据则可能需要多次硬盘IO来完成一次访问，因此我们需要一种利于缓存的键值存储结构。同时基于LSM-Tree的键值存储系统存在读写放大的问题，为了保证数据的有序性，LSM-Tree需要定期对数据做归并排序（compaction），在这过程中会不断重复写入已有的键和值，通常键值存储系统中的值占用的空间会比较大，在写放大中占比较高。

键值分离可以解决上述问题。键值分离架构图如图2所示，将键和值分开存储，然后将键和指向值的指针作为新的键值对存储在LSM-Tree中，值存储在一个单独的日志中。对于占用空间较大的值来说，键和指针占用的空间通常比值小得多， compaction时只需要写入键和占用空间较小的指针，compaction的开销和写放大都降低了。同时键值分离使LSM-Tree的总空间大大降低，使得我们可以把全部的键和指针都缓存在内存中，这样每次访问只需要一次磁盘IO来取对应的值即可，减少了磁盘IO的次数，从而实现更高的读写性能。

值存储在值日志中，新的值以追加的方式写入值日志，有的键值对会被写入多次，有的会被删除，如果不对值日志做垃圾回收，则会有很大的空间浪费，但做垃圾回收可能会占用较多硬盘带宽，因此一个轻量化的值日志的垃圾回收算法也是很重要的。我们可以记录下值日志的头和尾指针，新插入的值存到尾指针之后；做垃圾回收时，从头指针开始读取一段键值对，依次到LSM-Tree中查找，若该键值对已被删除则丢弃，仍然存在的则重新写入到尾指针后，最后更新LSM-Tree中对应的值的指针。在初始化值日志时可以多预留一些空间，从而降低垃圾回收的开销。

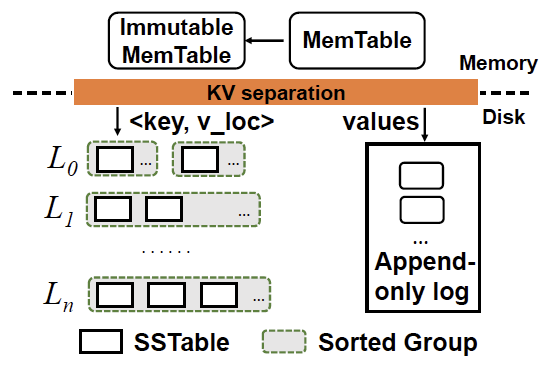


图2键值分离存储结构

**（3）****键值分离结构下的冷热感知缓存**

元数据存储采用键值分离结构的LSM-Tree之后，所有键和指向值的指针缓存在内存中，值存储在硬盘中，每次找到要查询的键时，都需要一次硬盘IO来取对应的值，速度较慢，因此可以考虑将经常访问的值缓存到内存中以提升访问效率。

我们可以根据数据的冷热来做缓存，将热数据缓存到内存中，为统计数据的冷热，为每个元数据条目维护一个计数器，每次访问元数据时，都将被访问的条目对应的计数器加一，所有键值对的计数器会被定期清空，清空时会根据数据的热度调整缓存的内容。根据SSTable内键值对在一定时间内的访问次数，为每个SSTable计算一个热度，当有少量键值对访问很频繁或较多键值对访问较频繁时该SSTable的热度较高，根据热度为所有SSTable维护一个优先队列，当清空计数器的周期到达时，不断取出优先队列的队头SSTable缓存到内存中，如图3所示。以SSTable为粒度可能较为粗糙，也可以在SSTable内对所有键值对的访问次数进行排序，只将访问次数高于阈值的键值对存入缓存，这种方式下存入缓存的键值对一定是热的，可以实现较高的缓存利用率。若写入的数据先前存储在缓存中，可以在写入时将其值更新，提高缓存的命中率。

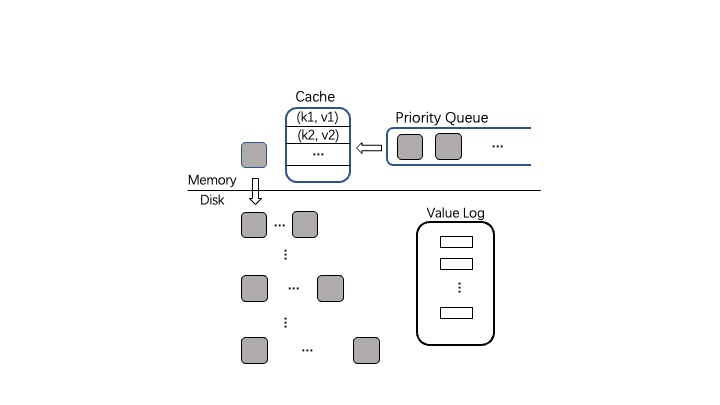


图3键值缓存

### 创新点与困难

为减少元数据的内存占用，我们采用键值化方式将树形结构扁平化，通过RocksDB管理扁平化的元数据，从而便于数据的落盘管理，节省内存空间。对于扁平化的键值化元数据，想要在不损失过多性能的情况下内存空间开销尽可能减少，对元数据分片间与分片内部做更精细的管理存在一定挑战。同时，需要将RocksDB中的数据接入原CubeFS中multi-raft协议的操作，在实现上也是一个难点。

为减少访问元数据时的硬盘IO次数，我们将键值分离方案与元数据管理相结合，以实现便于缓存的元数据存储方案，使得每次元数据访问至多只需要一次硬盘IO即可完成。同时为进一步优化元数据访问性能，我们提出了关于键值对的细粒度缓存方案，将热键值对缓存在内存中，进一步降低元数据操作时需要的硬盘IO次数。

键值分离后需要定期做值日志的垃圾回收，若垃圾回收的开销较大，消耗较多IO资源，可能会影响元数据的正常访问，我们需要设计一种较低开销的垃圾回收方案。细粒度的热数据缓存需要我们维护关于每个键值对的访问数据，并根据这些数据决定存入缓存的值，粒度越细管理就越复杂，如何为全部数据维护访问次数和选出他们中间值得被缓存的数据也是一个难点。

## 工作计划

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 项目阶段 | 起止日期 | 工作描述 | 交付成果 | 验收方法及验收标准 |
| 1 | T -T+3月 | 调研分布式文件系统元数据管理技术，提出基于键值存储的全外存元数据管理方案 | 元数据管理调研报告1份、键值存储设计报告1份 | 通过甲方验收 |
| 2 | T+4月--T+12月 | 实现基于键值存储的CubeFS元数据管理，并分析键值分离技术，对元数据管理进行优化设计 | 键值存储的设计报告1份、系统源码1份、测试报告1份  性能指标：与CubeFS原有的基于B树的全量缓存方案相比，内存占用降低一半以上。 | 通过甲方验收 |
| 3 | T+12月--T+18月 | 实现对元数据的冷热识别和缓存方案 | 基于冷热识别的缓存设计报告1份、实验报告1份、系统源码1份  性能指标：与采用外存KV系统管理元数据的方案相比，基于冷热识别的缓存使元数据访问的IO数量降低10%以上。  知识产权指标：完成CCF B类及以上论文1篇，联合申请专利1项 | 1. 技术交付物通过甲方验收  2. 论文接收证明  3. 专利交底书通过甲方验收 |

T:合同签署日期

## 费用

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | | | 总计（元） | 备注 |
| 直接费用 | 人员支出 | | 500000 | 学生助研费 |
| 研发支出 | 出差/会议费用 | 110000 | 参加国内学术活动与会议差旅费用，举办研讨会费用 |
| 其它费用（由项目PI根据项目情况填写） | 200000 | 设备费，材料费，测试费等 |
| 研发支出小计 | 310000 |  |
| 直接费用合计 | | 810000 |  |
| 间接费用 | 学校管理费等 | | 90000 | 项目费用总额的10% |
| 项目费用总计 | | | 900000 |  |

## 付款计划

甲方应按照下表的约定（“付款计划”）向乙方分期支付各期费用：

|  |  |  |
| --- | --- | --- |
| 阶段 | 支付条件 | 支付金额（人民币•元） |
| 1 | 协议签订后20个工作日 | 300000 |
| 2 | 项目第1阶段完成验收后20个工作日内 | 500000 |
| 3 | 完成结项验收后20个工作日内 | 100000 |

## 团队成员

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 职务 | 角色 | 参与时长（月） | 电话 | 电邮 |
| 许胤龙 | 教授 | 项目负责人 | 8 | 13696504559 | ylxu@ustc.edu.cn |
| 李永坤 | 副教授 | 技术骨干 | 8 | 13034042498 | ykli@ustc.edu.cn |
| 许睿达 | 硕士生 | 技术骨干 | 10 | 18009649152 | xuruida@mail.ustc.edu.cn |
| 朱文喆 | 博士生 | 技术骨干 | 10 | 18855726088 | wzhzhu@mail.ustc.edu.cn |
| 曹宇昂 | 博士生 | 技术骨干 | 10 | 19360216251 | cyadxx@mail.ustc.edu.cn |
| 李雨航 | 硕士生 | 技术骨干 | 10 | 18356139561 | hangge9468@mail.ustc.edu.cn |
| 任德铭 | 硕士生 | 技术骨干 | 10 | 17838512090 | rendm@mail.ustc.edu.cn |

## 合作老师简介

许胤龙，中国科学技术大学计算机学院二级教授、国家教育部软件工程专业教学指导委员会委员、国家高性能计算中心（合肥）常务副主任、中国计算机学会信息存储技术专业委员会委员与高性能计算专业委员会委员。曾任中国科大计算机学院副院长。1983年于北京大学数学系获学士学位，1989、2004年于中国科大计算机系获硕士、博士学位。主持多项国家自然科学基金面上、国家863项目，参与多项国家973、国家自然基金重点、国家863重点、科技部重点研发计划等项目。曾获得国家级教学成果二等奖、安徽省教学成果特等奖与一等奖、宝钢全国优秀教师奖、中国科学院优秀指导教师奖等。指导的学生曾获得全国优秀博士论文提名奖、中国科学院优秀博士论文、阿里星等。主要研究方向有存储系统、数据处理、高性能计算等。在USENIX FAST, USENIX ATC, VLDB, IEEE ICDE, ACM SIGMETRICS, IEEE INFOCOM, Spring, WWW, ACM ToS, IEEE JSAC, IEEE TPDS, IEEE ToC, IEEE, TCAD等国际顶级学术会议与学术杂志上发表了一系列高水平学术论文。

**研究方向**

容错存储系统、高性能计算、并行与分布式系统、云计算等

**教育经历**

1. 1979-09至1983-07，北京大学，数学科学学院，本科
2. 1986-09至1989-07，中国科学技术大学，计算机科学与技术学院，硕士
3. 2001-09至2004-07，中国科学技术大学，计算机科学与技术学院，博士

**科研与学术工作经历**

1. 1994-01至1996-12，德国多特蒙德大学，访问学者
2. 2004-09至今，中国科学技术大学，计算机科学与技术学院，教授

**近三年五篇代表性研究论文**

1. Qiang Zhang, Yongkun Li, Patrick P C Lee, Yinlong Xu, Si Wu. "DEPART: Replica Decoupling for Distributed Key-Value Storage." 20th USENIX Conference on File and Storage Technologies (USENIX FAST 22). 2022. (CCF推荐A类会议)
2. Yongkun Li, Zhen Liu, Patrick P C Lee, Jiayu Wu, Yinlong Xu, Yi Wu, Liu Tang, Qi Liu, Qiu Cui. "Differentiated Key-Value Storage Management for Balanced I/O Performance." 2021 USENIX Annual Technical Conference (USENIX ATC 21). 2021. (CCF推荐A类会议)
3. Qiang Zhang, Yongkun Li, Yinlong Xu, Qiu Cui, Liu Tang. "UniKV: Toward high-performance and scalable KV storage in mixed workloads via unified indexing." 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020. (CCF推荐A类会议)
4. Yongkun Li, Chengjin Tian, Fan Guo, Cheng Li, Yinlong Xu. "ElasticBF: Elastic Bloom Filter with Hotness Awareness for Boosting Read Performance in Large {Key-Value} Stores." 2019 USENIX Annual Technical Conference (USENIX ATC 19). 2019. (CCF推荐A类会议)
5. Helen H. W. Chan, Yongkun Li, Patrick P. C. Lee, Yinlong Xu. "HashKV: Enabling Efficient Updates in KV Storage via Hashing." 2018 USENIX Annual Technical Conference (USENIX ATC 18). 2018. (CCF推荐A类会议)

**近三年代表性研究专利**

1. 吴思，杜清鹏，李诚，李永坤，许胤龙，基于条带合并的局部可修复码冗余度转换方法及存储介质，2022.01.14，中国，CN114564335A
2. 李永坤，王千里，吴思，许胤龙，Redis集群数据移动方法、装置、设备及存储介质，2021.09.22，中国，CN113867631A
3. 许胤龙，王一多，邵新洋，吴思，李诚，基于负载类型感知的分布式文件系统元数据负载均衡方法，2021.08.12，中国，CN113608876A
4. [许胤龙](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=296369);[魏舒展](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=236511);[李永坤](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=256880);[吴思](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=264464);[陈友旭](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=241818)，“一种基于I/O自适应的日志文件系统数据存储方法”，2019-6-11，中国，授权公告号：CN105956090A
5. [许胤龙](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=296369);[孙东东](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=255489);[李永坤](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=256880);[吴思](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=264464)，“基于带宽比较的磁盘阵列扩容时校验更新方式选择方法”，2018-5-29，中国，授权公告号：CN105373352A
6. [许胤龙](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=296369);[魏舒展](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=236511);[李永坤](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=256880);[吴思](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=264464);[陈友旭](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=241818)，“一种基于动态条带构造的分布式编码方法”，2018-4-10，中国，授权公告号：CN105302500A
7. [许胤龙](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=296369);[王能](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=284596);[吴思](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=264464);[梁杰](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=240503);[邹懋](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=318315)，“一种基于分组编码的磁盘阵列构建方法”，2017-7-28，中国，授权公告号：CN104615380A
8. [许胤龙](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=296369);[王能](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=284596);[吴思](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=264464);[梁杰](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=240503);[邹懋](http://information.casip.ac.cn/caspatent/inventor.xhtml?iid=318315)，“一种能快速单盘修复的存储系统构建方法”，2017-7-28，中国，授权公告号：CN104809035A

**荣誉奖励**

1. 2017、2008年度中国科学院朱李月华优秀教师奖
2. 2014年度宝钢优秀教师奖
3. “计算机学科研究生国际化合作培养探索与实践”，2012年度安徽省教学成果二等奖（个人排名第3）
4. “十一五”国家重点图书、中国科学技术大学精品教材《组合数学引论》（第二版）获中国大学出版社第二届优秀教材二等奖，许胤龙、孙淑玲编著，中国科学技术大学出版社，2010年4月第二版
5. 2006年度中国科学院优秀研究生指导教师奖
6. 2020 IEEE SRDS最佳论文奖
7. 并行算法类教学基地建设，2001年度安徽省教学成果一等奖，国家级教学成果二等奖（个人排名第4）
8. 指导的博士论文“分布式存储系统中基于纠删码的容错技术研究”获2015年度ACM中国合肥分会优秀博士论文奖（合肥分会两人获奖，全国分会共16位博士获奖。学生：朱云锋，2013年度阿里巴巴校招阿里星获得者。）

## 参考文献

1. Shvachko, Konstantin V. "HDFS Scalability: The limits to growth." *; login:: the magazine of USENIX & SAGE* 35.2 (2010): 6-16.
2. Ghemawat, Sanjay, et al. "The Google file system." *Proceedings of the nineteenth ACM symposium on Operating systems principles*. 2003.
3. Shvachko, Konstantin, et al. "The hadoop distributed file system." *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Ieee, 2010.
4. Pawlowski, Brian, et al. "NFS Version 3: Design and Implementation." *USENIX Summer*. 1994.
5. Ousterhout, John K., et al. "The Sprite network operating system." *Computer* 21.2 (1988): 23-36.
6. Srivas, M., et al. "Map-Reduce Ready Distributed File System, 2011." *US Patent App* 13.162,439.
7. Weil, Sage A., et al. "Ceph: A scalable, high-performance distributed file system." *Proceedings of the 7th symposium on Operating systems design and implementation*. 2006.
8. Pan, Satadru, et al. "Facebook's tectonic filesystem: Efficiency from exascale." *19th USENIX Conference on File and Storage Technologies (FAST 21)*. 2021.
9. Wang, Yiduo, et al. "CFS: Scaling Metadata Service for Distributed File System via Pruned Scope of Critical Sections." *Proceedings of the Eighteenth European Conference on Computer Systems*. 2023.
10. Shvachko, Konstantin V., et al. "Scaling namespace operations with giraffa file system." *USENIX; login* (2017).
11. Ren, Kai, et al. "IndexFS: Scaling file system metadata performance with stateless caching and bulk insertion." *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2014.
12. Thomson, Alexander, et al. "{CalvinFS}: Consistent {WAN} Replication and Scalable Metadata Management for Distributed File Systems." *13th USENIX Conference on File and Storage Technologies (FAST 15)*. 2015.
13. Lv, Wenhao, et al. "{InfiniFS}: An Efficient Metadata Service for {Large-Scale} Distributed Filesystems." *20th USENIX Conference on File and Storage Technologies (FAST 22)*. 2022.

# 附件二：保密协议

**保 密 协 议**

甲方：OPPO广东移动通信有限公司

地址：东莞市长安镇乌沙海滨路18号

乙方专家：

身份证号：

通讯地址：

鉴于，甲、乙双方即将或正在进行 面向CubeFS的元数据管理优化 的磋商或合作（“协议目的”），在磋商或合作的过程中，双方将向或已经向对方披露或从对方知悉保密信息（“保密信息”定义见下），其中保密信息披露方简称“披露方”，保密信息接受方简称“接受方”。为了保护双方对其保密信息拥有的合法权益，双方协商一致，于东莞市长安镇订立本保密协议如下：

1. **保密信息**

（一）保密信息定义：本协议所称保密信息是指任何一方和/或其关联方（定义见下文）向另一方和/或其关联方披露的以下信息：

1、在与该信息有关的讨论中形成的信息，无论这些讨论是在信息披露之前/同时/之后进行的；

2、以书面形式、电子形式、磁介质（如磁盘存储器）或其他有形形式披露的信息，但条件是该信息上标有披露方的名称、标志、商标或商号，或该信息明显地被标记为专有或保密信息；

3、口头或其他无形形式披露的信息，但前提是此类口头信息在披露时即被指定为专有或保密信息；

4、无论披露方式如何，根据其性质通常被视为专有或保密信息的任何信息。保密信息包括但不限于销售数据、市场营销信息、分销信息、商业计划和其他商业条款、技术信息、财务信息、管理信息、法律信息（包括但不限于许可、诉讼、调解或仲裁）、人事和劳工信息、业务或公司发展策略，以及任何一方的技术、工程设计、电路设计、制造方法、配方、工艺流程、技术规格、计算机程序、软件、数据库、源代码、物理编码、设计、创意、图纸、样品、模型、照片、价格、合同、向政府部门的申请、内容开发、编程、技术参数、说明书、商业计划、图表、流程图、客户信息、营销计划、货运渠道，财务和税务报告等等。双方正在讨论的有关协议目的的事实，以及口头、书面交流所形成的任何类似信息均应被视为保密信息。

1. **非保密信息的范围**

（一）本协议定义的保密信息不包括以下情形：

1、在披露方披露时，已经合法公开或接受方已合法拥有的信息；

2、非因违反本协议约定的行为，已经或正在被公开的信息（但因第三方违反相关保密义务导致公开的除外）；

3、接受方从第三方获取的信息，而当时第三方不受保密义务的约束；

4、接受方在未接触披露方的信息的情况下独立开发的保密信息。

（二）掌握上述信息的一方应提供书面证据证明，否则均视为保密信息。

1. **保密义务**

1、接受方仅有权将保密信息用于双方协议目的。

2、 接受方仅有权为协议目的，向有必要知晓的参与合作或磋商的雇员和关联方披露保密信息。接受方仅可在获得披露方书面同意的情况下，向其他确有必要知悉保密信息的其他第三方披露（“授权第三方”）。接受方保证其雇员、关联方及授权第三方应接受至少与本协议同等严格的保密条款的约束。接受方应为其雇员、关联方及授权第三方在任何情况下、任何时候违反本协议的行为承担责任。本协议所称关联方，是指一方直接控制或间接控制、或直接或间接控制一方、或与一方共同控制同一家法律实体或能对其施加重大影响、或与一方受同一家法律实体直接或间接控制的法律实体。包括但不限于一方的母公司、子公司，与一方受同一母公司控制的子公司等。“控制”是指直接或间接持有百分之五十以上（50%）的表决权，或选举被控制或被共同控制下法律实体的董事或核心管理人员；“重大影响”是指对法律实体的财务和经营政策有参与决策的权利。

3、接受方应采取一切合理且不低于接受方对其类似的保密信息所采取的措施来保守披露方披露的保密信息，防止保密信息被盗窃或（和）泄露、未经授权的使用、因任何第三人的疏忽而导致泄露。

4、接受方如知悉保密信息被未经授权地使用或披露，应立即通知披露方并协助披露方采取救济措施。

5、接受方不得以不正当手段（包括但不限于未经披露方许可对披露方非公开的文件、资料等进行浏览、复制、拍摄、抄录等）获取披露方未向接受方披露的非公开信息。

6、如接受方应法律、行政法规的强制性规定或司法机关之要求必须披露保密信息的，则应尽最大努力事先以书面方式通知披露方，且在法律法规允许的范围内采取合理且足够的救济手段（如申诉等）尽可能限制保密信息的披露范围，保证保密信息的披露仅限于法律法规或司法机关要求披露的最小范围和程度。

7、保密信息可能由披露方或披露方的关联方披露，对于披露方的关联方披露的保密信息，接受方也应按照本协议约定遵守保密义务。

1. **不授予其他权利及许可**

1、未经披露方事先书面授权，接受方不得对保密信息进行反向工程、反编译、反汇编等，不得利用保密信息谋取任何知识产权。

2、除非双方通过协议方式（如合作协议、授权协议）另有约定外，披露方根据本协议向接受方披露保密信息、其他相关信息，不构成授予任何关于专利、著作权、商标权等知识产权的许可行为。

3、未经甲方书面同意，乙方不能擅自使用甲方品牌做任何形式的宣传。所有涉及甲方的宣传，不限于商标、LOGO、名称、文字，图片，视频等方式，都需要获得甲方的事先书面同意，否则甲方有权终止合作或磋商且追究责任。

4、双方签订本协议不能被解释为：

（1）在合同双方间形成合伙或其它任何导致共同、连带、补充责任的关系；

（2）在合同双方间形成代理、合作关系，如对采用或含有保密信息的产品的委托宣传、买卖等；

（3）授权一方为另一方招致费用或其它任何形式的义务。

1. **对保密信息不作保证**

披露方对所有的保密信息均按原样提供，对其准确性、完整性、合法性、适用性不做任何明示、暗示的保证或其他承诺。本协议不得被解释为披露方承诺所披露保密信息不侵犯任何第三方包括专利权、著作权、商标权在内的合法权益。

1. **资料的返还与销毁**

1、所有由披露方披露给接受方的资料，包括但不限于文件、图表、模型、设备、草图、设计、清单、样品、样件等保密信息的任何载体均属于披露方财产。双方未达成合作时，或在双方的合作关系终止时，或在本协议被终止或解除时，或者在披露方以书面形式要求时，接受方应在十（10）日内立即：（1）返还或销毁全部载有保密信息的资料及其复印件、摘要等；（2）消除所有存储在计算机和其他设备中的保密信息；（3）提供一份被销毁资料的清单，证明接受方已履行上述两项义务。

2、尽管保密信息已返还或销毁，接受方及其雇员仍继续受本协议的约束直至本协议约定的接受方保密义务期限终止。

1. **违约赔偿及救济**

1、接受方违反本协议的约定或接受方的雇员、关联方或授权第三方违反本协议的约定的，接受方应向披露方支付人民币按照本合作项目总金额的100%的违约金。

2、如披露方遭受的损失超过上述约定的违约金数额的，接受方还应向披露方赔偿超过违约金的那部分损失以及披露方因采取救济手段发生的合理费用（包括但不限于调查取证费、执行该赔偿的费用、合理的律师费）。

3、尽管有上述约定，双方理解并同意，保密信息是披露方所有的独特而重要的资产，任何违反本协议的行为可能给披露方造成金钱所不足以弥补的无法挽回的损失。双方同意，若违约行为发生或可能发生时，披露方有权无需证明实际遭受的损害而采取限制和暂停双方业务合作事项或者采取其他合理的法律行动（包括但不限于申请禁止令或保全措施等）。

1. **保密期限**

1、本协议自双方签字盖章之日起生效，且有效期自生效之日起为三（3）年。本协议届满或终止后，双方还需继续履行保密义务两（2）年。

2、双方对在本协议签订之前及合作或磋商期间知悉的保密信息均须相互承担保密义务。

1. **适用法律与争议解决**

1、本协议适用中华人民共和国的法律，并排除冲突规范的适用。

2、任何基于本协议产生的或与协议相关的争议应由双方友好协商解决。如协商不成，可将该争议提交至本协议签订地法院以诉讼方式解决，诉讼费用应由败诉方承担。

1. **其他**

1、本协议中的权利和义务不可转让。

2、各方应采取一切适当措施，遵守所有适用的国家出口管制法律、法规和规则。任何一方应根据请求方的书面要求，向对方提供根据本协议交换的任何技术、软件或物项的出口管制分类资料。

3、本协议涵盖了双方对本协议内容的全部理解，在订立本协议前双方任何与本协议相关的洽谈、往来书信、约定、或与本协议事项相同的协议如与本协议矛盾，以本协议为准。

4、任何对本协议的修改或增补，双方应另行签订补充协议，否则视为无效。

5、披露方未行使或延迟行使本协议约定的任何权利均不构成对该权利的放弃，任何对该权利的单独或部分行使均不妨碍其他权利或继续行使其权利。本条款所规定的权利和救济措施可以逐渐累加，因而不排除任何其他权利或救济措施（不论是法律规定或通过其他方式）。

6、双方就协议目的而另行签署主业务合同的，则本协议将作为双方主业务合同的附件，与主业务合同具有相同法律效力；本协议未尽事宜，适用主合同约定；主业务合同关于保密违约责任的约定与本协议约定有冲突的，以责任较重的约定为准。双方未就协议目的而另行签署主业务合同的，则本协议独立生效。

7、根据法律对本协议某一条款无效的认定，并不影响本协议其他条款的有效性。披露方和接受方同意，将用最接近该无效条款目的、经济效益的有效条款替代该无效条款；但此种替换，不应剥夺协议一方或双方在本协议中的实质利益。

8、本协议可通过扫描件、传真件、电子签章等方式签署或通过邮件方式传签，双方同意此种文件与原件具有同等的法律效力。

9、本协议经双方签署后生效。

10、本协议一式 肆 份，双方各执 贰 份，具有同等的法律效力。

（以下无正文）

甲方：OPPO广东移动通信有限公司 乙方专家：

签字： 签字：

日期： 日期：