

徐亮亮

✉ llxu@mail.ustc.edu.cn · ☎ (+86) 15656547361 · 🌐 <https://lianglxu.github.io>

个人介绍

我的主要研究方向是分布式存储系统、数据修复、纠删码 (EC) 以及内存存储系统。博士期间，我的工作主要包括设计有效的数据布局，高效的故障修复算法，设计实际的纠删码部署策略，最后在开源存储系统中实现。相关论文发表 (或已投稿) 至 IEEE INFOCOM, IEEE IPDPS, USENIX HotStorage, ACM TOS, IEEE TC, IEEE TIT 以及 IEEE TPDS 等国际顶级会议及著名期刊上。其中，共发表 (或已投稿) 论文 7 篇以及专利 1 项，包含 CCF 推荐列表 A 类论文 5 篇 (第一作者 (或学生一作))。

教育背景

博士，中国科学技术大学，合肥

2017 – 至今 (预计 2022.06 毕业)

- 计算机科学与技术，先进数据系统实验室
- 导师：许胤龙教授，吕敏副教授

本科，安徽大学，合肥

2013 – 2017

- 信息与计算科学
- GPA: 3.69/4.0 (前 5%)
- 导师：施敏加教授

实习经历

华为技术有限公司，云存储 lab，内存存储项目组，深圳

2020 年 10 月 – 2021 年 01 月

- 导师: 左鹏飞
- 实习内容: 纠删码在内存存储系统中的应用，具体如下:
 - 调研了以下几个方面工作: EC 在 cache 场景下的应用、RDMA 环境下的高性能 key value 数据库、RDMA 的一些高级硬件结合级操作 (如 doorbell)、EC 在 x86 架构下的应用最多的两个库 (Jerasure 和 ISA-L) 的设计文档与实现、EC 在 data-intensive 场景下的应用 (如 memcached) 以及 Disaggregating Persistent Memory 等场景。
 - 测试了大量 Jerasure 和 ISA-L 库的性能。如: 小对象编解码 latency、throughput, 多线程编解码性能, 预加载 SSE 指令集的编解码性能, EC 细粒度参数 word、packet、buffer size 等灵敏度性能, pipeline 编解码性能, 不同 EC 编码类型的性能 (RS、CRS 等)。
 - 发现了一些现象并分析了性能瓶颈: 如采用预加载指令集后, EC 在小对象时延较低 (<8k 对象在 1us 以下); EC 编码对象时延的增加比太大, 甚至大于线性增长, 主要原因可能是大对象计算编码时 L1 和 L3 等 cache miss 快速增加; 单边的 RDMA 读写性能在小对象时延高于 EC 编码, 但 RDMA 读写时延增加缓慢; EC 编码与 RDMA 写之间的 latency 存在 GAP, 小对象 EC 编解码时延较低, 但大对象 RDMA 读写时延较低。
 - 提出了一些解决方案: 考虑 in-memory KV 数据库场景: 如关于写性能使用动态 pipeline 编解码, 先把大对象 split 为多个小对象, 串行的编解码性能更好, 这样能够把编码时延降下来, 接着再动态的合并小 split (基于网卡端 QP 的负载以及 CQE 状态), 进一步优化网络消息转发; 关于读性能长尾延迟优化, 可采用多读校验块、设计低复杂度生成矩阵 (一个 parity 计算简单), reorder 到达的消息来避免不必要的降级读; 多线程来优化降级读解码开销等策略。

项目

1. **MonEC**. 该项目提出了一个在内存存储系统中部署 EC 的工作流程。我设计了 cache 局部性友好的编解码流程, 可将编解码瓶颈时间减少至 1/3。该项目正在进行中, 实现系统为 Apache Crail。
2. **SelectiveEC [HotStorage 2020, TOS 2021]**. 该项目提出了一个均衡调度模块, 以动态选择要批量修复的任务来组成一批条带, 并为每个重构任务选择源节点和替代节点。当发生故障时, 它可以实现均衡的网络修复流量以及计算资源和磁盘 I/O。我设计了调度模块及在 Hadoop 3.1.4 中的实现原型。
 - 概要: EC 广泛应用于实际的存储系统中, 以提供低存储开销的容错能力。当出现节点故障时, 丢失的块会被批量恢复。由于每一批中有限的条带数量, 其数据分布是不均匀的。再加上恢复

任务的源节点和替换节点的随机选择，每一批的恢复工作负载是倾斜的，这严重降低了故障恢复的速度。我们提出了一个新的恢复任务调度模块 **SelectiveEC**，该模块为基于 EC 的大规模存储系统提供可证明的网络流量和恢复负载均衡。它依靠二部图模型来调度幸存节点之间的恢复流量，动态地在重构队列中选择来形成一批重构任务，并进一步调度需要读取源数据块或存储恢复的块。**SelectiveEC** 支持单节点故障和多节点故障恢复，可以部署在同构和异构网络环境中。我们在 HDFS 中实现了 **SelectiveEC**，并在一个 18 节点的本地集群和 AWS EC2 中评估其恢复性能。与最先进的工作比较，**SelectiveEC** 提高了 30.68% 的恢复吞吐量。在异构网络环境下，由于实时调度避免了 straggler，进一步实现了 HDFS 的 $1.32\times$ 修复吞吐量和 $1.23\times$ 前端应用的吞吐量。

3. **PDL [INFOCOM 2020, TC 2021]**. 该项目提出了一种有效的基于组合设计工具 PBD (Pairwise Block Design) 的数据布局 PDL，以加快混合 EC 分布式存储系统中单节点故障的数据修复。由于减少了机架间的通信量，并在修复过程中实现了读写 I/O 的负载平衡，因此它实现了几乎均匀的数据分布以及更高的修复性能。我设计了数据的放置策略以及相应的故障恢复方案，并且在 Hadoop 3.1.1 中实现了它们。

- 概要：EC 在分布式存储系统 (DSSes) 中越来越受欢迎，因为它以低的存储开销提供了高可靠性。然而，传统的随机数据放置方式在故障恢复过程中会造成大量的跨机架流量和严重的负载不平衡，严重影响恢复性能。此外，在 DSS 中共存的各种 EC 策略加剧了上述问题。在本文中，我们提出了一种基于成对平衡设计 (PBD) 的数据布局 PDL 来优化 DSSes 中的故障恢复性能。基于 PBD 组合设计的性质，PDL 给出了统一的数据布局。在此基础上，提出了一种基于 rPDL 的故障恢复方案。rPDL 通过均匀选择替代节点和检索确定的可用块来恢复丢失的块，有效地减少了跨机架流量，并提供了几乎均衡的跨机架流量分布。我们在 Hadoop 3.1.1 中实现了 PDL 和 rPDL。实验结果表明，与现有的 HDFS 数据布局相比，rPDL 平均减少了 62.83% 的降级读延迟，提供了 $6.27\times$ 的数据恢复吞吐量，为前端应用提供了更好的支持。

4. D^3 [IPDPS 2019, TPDS 2020]. 该项目提出一种单 EC 的数据布局 D^3 ，可以在大规模 EC 的分布式存储系统中的节点之间均匀地分配数据/奇偶校验块，并可以最小化单节点故障时的跨机架修复流量。我将 D^3 集成到 Hadoop 3.1.0 的 HDFS-EC 模块中，并基于 Reed-Solomon (RS) 码以及 Locally Repairable Codes (LRC) 测试了修复性能。

- 概要：由于单个不可靠的商品组件，故障在大型分布式存储系统中很常见。EC 广泛应用于实际的存储系统中，以提供低存储开销的容错能力。然而，随机数据分布 (RDD) 通常用于 EC 存储系统，它会导致严重的跨机架流量、负载不平衡和随机访问，从而对故障恢复产生不利影响。在本文中，利用正交矩阵，定义了确定性数据分布 (D^3)，使数据/校验块均匀分布于节点之间，并提出了一种基于 D^3 的高效故障恢复方法，使单节点故障下的跨机架修复流量最小化。由于 D^3 的均匀性，所提出的恢复方法不仅可以均衡机架内节点之间的修复流量，还可以均衡机架之间的修复流量。我们在本地 28 台节点的 Hadoop 分布式文件系统 (HDFS) 中实现了基于 D^3 的 Reed-Solomon 码和 Locally Repairable Codes 码。我们的实验表明，与 RDD 相比， D^3 显著提高了 RS 码的故障恢复速度，达到了 $2.49\times$ ；并提高了 LRCs 的故障恢复速度，达到了 $1.38\times$ 。此外， D^3 在正常和恢复状态下都比 RDD 更好地支持前端应用。

5. **A note on one weight and two weight projective Z_4 -codes [TIT 2017]**. 该项目是我本科大学生科研训练计划时的项目。一方面我解决了关于代码编码的一个公开问题，另一方面完成了对一个丢番图问题的有效分类，得到了在空间 Z_4 上关于类型为 $4^{k_1}2^{k_2}$ 的 two-Lee weight projective codes 不存在的条件。

发表论文

1. **SelectiveEC: Towards Balanced Recovery Load on Erasure-coded Storage Systems.**
Liangliang Xu, Min Lyu, Qiliang Li, Lingjiang Xie, Cheng Li and Yinlong Xu.
Submitted to ACM Transactions on Storage (TOS 2021).
2. **A Data Layout and Fast Failure Recovery Scheme for Distributed Storage Systems with Mixed Erasure Codes.**
Liangliang Xu, Min Lyu, Zhipeng Li, Cheng Li and Yinlong Xu.
Submitted to IEEE Transactions on Computers (TC 2021).
3. **PDL: A Data Layout towards Fast Failure Recovery for Erasure-coded Distributed Storage Systems.**
Liangliang Xu, Min Lv, Zhipeng Li, Cheng Li and Yinlong Xu.

IEEE International Conference on Computer Communications (INFOCOM 2020).
(AR: 268/1354 = 19.8%, CCF A)

4. Deterministic Data Distribution for Efficient Recovery in Erasure-Coded Distributed Storage Systems.
Liangliang Xu, Min Lyu, Zhipeng Li, Yongkun Li and Yinlong Xu.
IEEE Transactions on Parallel and Distributed Systems (TPDS 2020), 31.10: 2248-2262.
(CCF A)
5. SelectiveEC: Selective Reconstruction in Erasure-coded Storage Systems.
Liangliang Xu, Min Lyu, Qiliang Li, Lingjiang Xie and Yinlong Xu.
12th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 2020)
6. D3: Deterministic Data Distribution for Efficient Data Reconstruction in Erasure-Coded Distributed Storage Systems.
Zhipeng Li, Min Lv, Yinlong Xu, Yongkun Li and **Liangliang Xu**.
33rd IEEE International Parallel & Distributed Processing Symposium (IPDPS 2019).
(AR: 102/372 = 27.7%, CCF B)
7. A note on one weight and two weight projective Z_4 -codes.
Minjia Shi, **Liangliang Xu** and Gang Yang.
IEEE Transactions on Information Theory (TIT 2017), 63.1: 177-182.
(CCF A)

专利

1. 一种基于纠删码存储系统的负载均衡修复调度方法
吕敏, **徐亮亮**, 李启亮, 谢灵江, 许胤龙
专利号: 202010313968.5, 申请时间: 2020.04.20, 公开时间: 2020.08.07

邀请报告

- 2020.07: Paper Presentation in INFOCOM 2020, PDL: A Data Layout towards Fast Failure Recovery for Erasure-coded Distributed Storage Systems, Online.
- 2020.07: Paper Presentation in HotStorage 2020, SelectiveEC: Selective Reconstruction in Erasure-coded Storage Systems, Online.
- 2020.06: Invited Talk in the 18th ChinaSys workshop, PDL: A Data Layout towards Fast Failure Recovery for Erasure-coded Distributed Storage Systems, Online.

专业技能

- 编程语言: Java. C/C++. Matlab. Python. Linux Shell.
- 分布式系统: HDFS. Crail.

获奖情况

- 深交所奖学金, 2020.
- INFOCOM Student Conference Award, 2020.
- 品学兼优毕业生, 2017.
- 优秀本科毕业论文, 2017.
- 学术科技一等奖学金, 2016.
- 国际大学生数学建模比赛一等奖, 2016.
- 校级三好学生, 2016.
- 国家励志奖学金, 2014/2015/2016.
- 团学奖学金, 2015.
- 挑战杯校级二等奖, 2014.