存储EC编码杂谈

Accela Zhao Accela推箱子 2017-11-29

现代分布式存储，尤其是云存储，在成本和可靠性考量下经常使用EC（Erasure-Coding）编码存储数据。EC中的Erasure指数据丢失（而不是纠错），对应存储节点挂掉。除了经典的RS（Reed-Solomon）编码（RS编码是最简单的MDS编码，同数据体积下容错最好），还有很多种改进。大体方向有

1） 减少网络传输量
2） 减少通信的结点数
3） 减少磁盘读取数据量
4） 一次修复多个丢失的symbol
5) 更好地分散修复负载，并发修复
6） 减少修复时间

上面的分类可以在F.Oggier的survey的chapter5找到。不过实际EC编码常常有覆盖多个方面的。该文很不错，覆盖了EC编码基础知识和当前主要现状

[Coding Techniques for Repairability in Networked DistributedStorage Systems](http://phdopen.mimuw.edu.pl/lato12/longsurvey.pdf)

---- Regenerating Code ----

Regenerating code可以属于分类（1），它节省网络带宽达到50%以上。但是，它需要联络更多结点，磁盘和网络的IO数量增加。云存储反而是从IO数量减少获益更多。另外，regenerating code基本上一次只能修复一个丢失的symbol；尽管普通的RS编码网络消耗网络带宽多得多，但可以一次修复多个丢失，利用先批量修复再把symbol复制出去的方式，同样可以节省大量带宽。Regenerating code的经典论文有

[Network Coding for Distributed Storage Systems](http://users.ece.utexas.edu/~dimakis/RC_Journal.pdf) 第一次提出regenerating code，用information flow graph算出了regenerating code夸张的带宽下界

[Optimal Exact-Regenerating Codes for Distributed Storage at the MSR and MBR Points via a Product-Matrix Construction]

(https://people.eecs.berkeley.edu/~rashmikv/papers/product_matrix_codes.pdf) 解出了达到带宽下界的regenerating code编码构造方式。当然，编码有更多的计算开销。

[Explicit Constructions of High-Rate MDS Array Codes With Optimal Repair Bandwidth](https://arxiv.org/pdf/1604.00454.pdf) 用array code的方式构造了更简单、计算量更小的编码。不过它支持的symbol数量较小。

---- LRC Code ----

微软用的LRC系列编码背后其实历史渊远。它可以分类到（2）。早期的"Rethinking Erasure Codes for Cloud File Systems:
Minimizing I/O for Recovery and Degraded Reads"通过预计算选择最优修复方程（recovery scheduling），还有RS编码作个旋转，已经能节省不少IO。但还是被后期的Pyramid Codes超越。

Pyramid Codes原本的RS编码的一个parity拆成两个，各覆盖一部分data symbol。通用化后，Generalized Pyramid Codes可以各个parity只覆盖各自locality区域内的data symbol；覆盖区域也可以重叠。这个包含了LRC系列编码。"Erasure Coding in Windows Azure Storage"进一步对Azure存储中用的LRC编码做了总结。

[Pyramid Codes: Flexible Schemes to Trade Space for Access Efficiency in Reliable Data Storage Systems] (https://staff.ie.cuhk.edu.hk/~mhchen/papers/nca.07.pyramid.codes.pdf)

之后这个论文进一步推导的Locality的理论。在不放宽存储开销的情况下，有n − k >= roof(k/r) + d - 2，它确定了locality r在给定海明距离d（即最多可恢复d-1的丢失）的下界。好的编码能够达到下界，也可以牺牲数据体积改进延迟。

[On the Locality of Codeword Symbols](https://arxiv.org/pdf/1106.3625.pdf)

LRC编码更加强大的是，它是基于RS编码的，因此所有RS编码的优化，它都可以用。这个论文有对RS编码矩阵的经典优化。首先，我们以Cauchy矩阵作编码矩阵，因为Galois域(GF(2^w))的特性，它可以把大的GF(2^w)的域映射到GF(2)；原本Galois域的计算难点是乘法开销大，映射到GF(2)后，乘法变成了XOR。第二，该论文发现，含有更少的1的编码矩阵，编码效率更高。再次，Intel CPU的SSE/AVX向量计算指令，也使得编码速度大大加快。

[Optimizing Cauchy Reed-Solomon Codes for Fault-Tolerant Storage Applications](http://web.eecs.utk.edu/~plank/plank/papers/CS-05-569.pdf)

另外还有一些有意思的优化。这个论文把大Galois域拆成多个小域，用经过优化的查表法处理乘法，小表也可保留在cache里。

[Optimizing Galois Field Arithmetic for Diverse Processor Architectures and Applications](http://www.kaymgee.com/Kevin_Greenan/Publications_files/greenan-mascots08.pdf)

编码Locality一般认为是由三篇论文共同提出的。除了上面的“On the Locality of Codeword Symbols”，还有Simple Regenerating Code。下面的论文在推导理论。与“On the Locality of Codeword Symbols”不同之处在于，它放宽存储空间限制，结点存储a=(1+e)M/k数据，进一步压低locality下界： d<=n-roof(k/(1+e))-roof(k/(r(1+e)))+2。最终给出一个编码构造，Simple Regenerating Code。这个编码虽然消耗1/3额外存储空间，但无论有多少symbol，都只需要联络4个结点去修复，而且与RS编码容忍同等数量的symbol丢失，而且单丢失只需要XOR修复。

[Locally Repairable Codes](https://arxiv.org/pdf/1206.3804.pdf)

第三个Locality论文是O.Fragger的Self-repairing codes。它通过多项式构造编码，通过p(a+b)=p(a)+p(b)的特性进行恢复，恢复时只需要XOR操作，locality非常小。不过编码本身不是MDS的（给定存储空间开销，恢复能力最优），也不是systematic（编码后初始数据还在）的。

[Self-repairing Homomorphic Codes for Distributed Storage Systems](https://arxiv.org/pdf/1008.0064.pdf)

---- XOR Code ----

另外一大分支仅使用XOR进行编码。RS编码虽然有最佳容错能力，但是需要Galois域上的开销较大的乘法运算。如果仅使用XOR，编码解码速度极快。XOR类编码一般速度快，恢复时通信结点少，但是需要更多的存储空间，而且不能容忍太多symbol丢失。这类XOR编码

大量用于RAID存储，而云存储基本全用RS编码。这是因为云存储对存储空间成本敏感，而且需要高容错。

早期提出的EVENODD编码算是XOR编码的开山鼻祖，将磁盘排一排，横着XOR一遍生成一个parity，对角线来一遍生成一个parity。它是n+2的模式，容忍最多两个磁盘丢失，达到MDS，编码和恢复性能都很好。另外还有个与EVENODD不同，用两条对角线编码的“X-Code: MDS Array Codes with Optimal Encoding”，两条对角线，就成了X。而另一个是STAR编码，通过加进来第三个对角线的parity，从而能够容忍3个磁盘丢失：“STAR: An Efficient Coding Scheme for Correcting Triple Storage Node Failures”。

[EVENODD: An Efficient Scheme for Tolerating Double Disk Failures in RAID Architectures](https://authors.library.caltech.edu/29320/1/BLAieeetc95a.pdf)

XOR编码比RS编码有个不同的地方，恢复的时候往往有多条路径，而需要选择最好的，一般叫recovery scheduling。EVENODD后来有了更优的recovery schedling策略，见“Rebuilding for Array Codes in Distributed Storage Systems”。类似地，另一种只用XOR的编码叫RDP（row-diagonal parity），它同样用是n+2模式，一条横向一条稍不同的对角线编码，容忍最多2个磁盘丢失。下面的论文提出了改进的recovery schedling策略，达到最少磁盘IO数和各磁盘的恢复负载平衡。这些应该可以算进开头的分类（3）。另外，EMC的XtremIO使用了类似RDP的XDP编码。

[Optimal Recovery of Single Disk Failure in RDP Code Storage Systems](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.439.9023&rep=rep1&type=pdf)

一个比较经典的XOR编码是Flat XOR-based code。LDPC编码（后面讲）在symbol数量很多时（>50），才展现超越RS编码的表现，才有较稳定的表现。而Flat XOR像是给出了以symbol较少的LDPC编码。它有明确的编码性质，牺牲一定的存储空间，容忍2到3个丢失，恢复丢失的性能比RS好得多，通信结点数比RS少得多。Flat XOR应该可以算进分类（2）。

[Flat XOR-based erasure codes in storage systems: Constructions, efficient recovery, and tradeoffs](https://pdfs.semanticscholar.org/09be/d5a75cbdba4b930cdca6bd2499d61121e030.pdf)

有一类非常有名的编码较LDPC（low-density parity check）编码。围绕它有大量研究，它在通信领域里用得较多，能够恢复大量错误（lossy channel），进入了10GBase-T Ethernet和Wi-Fi 802.11标准（见wiki）。但由于它只有在symbol数量很多（>50）时才显现超过RS编码的优势，存储领域用得较少。编码方式是，画一个二分图，数据symbol在左边，编码symbol在右边，从数据到编码结点的连线表示要XOR。连线是通过概率分布随机生成的。表示成编码矩阵，就是一个只含有1和0的稀疏（low-density）矩阵。下面的论文分析了市面上的80多种LDPC，对于综合了解很有用

[A Practical Analysis of Low-Density Parity-Check Erasure Codes for Wide-Area Storage Applications](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.5556&rep=rep1&type=pdf)

最后有个RAID6中使用的XOR编码，Liber8tion编码，拥有优异的性能。它只使用XOR，编码矩阵中有最少的1，达到MDS，并且在一些场景的表现下超过RDP编码。论文中优化了recovery scheduling。另外，这个论文也为其它XOR编码提供了很好的总结。

[The RAID-6 Liber8tion Codes](https://www.usenix.org/legacy/event/fast08/tech/full_papers/plank/plank_html/)

---- Distributed Repair ----

对于分类（5），它是RAID场景提出的，在云时代已经是常用的了。Ceph中把PG的副本分散在不同结点，一个结点挂掉，整个集群都并行地参与修复。减少的修复时间，有利于提高MTTF。而对于分类（6），几乎所有优化都有是在减少修复时间。

最后，EC编码是为了在平衡成本的情况下提高云存储的可靠性。Google有篇论文提供了丰富的总结，其中包括如何通过编码方式计算MTTF。更重要的是，这篇论文着重讨论了correlated failure对可靠性的影响，也讨论了multisite时的不同可靠性。

[Availability in Globally Distributed Storage Systems](https://www.usenix.org/legacy/event/osdi10/tech/full_papers/Ford.pdf)

致癌！事关书包！

人民网科普

全国悲痛！巨星陨落，他是14亿国人都该致谢的救星！

宪法小卫士