网络存储系统容错编码技术进展

Erasure Codes for Networked Storage Systems

/LIN Sheng
(天　工大学　与　工　学　,天　300191)
刘晓光/LIU Xiaoguang
刚/WANG Gang
(南开大学 信息技　学学　，天　300071)
(College of Information Technical Science, Nankai University, Tianjin 300071, China)

中图分　号：TP393.03　文　志　：A　文　号：1009-6868 (2010) 05-0000-00
基　：国家　技　发展(“ 863 ”)　划(2008AA01Z401)；国家　学基　(60903028)

摘　：
前,专业　大型　存储　均发展为包含多块　大型　列　。　中　数　不断增加，　失效引　数据丢失　可　性　大。对于　存储　中　分　失效所引　数据丢失　，　前业　公　好　决方　是使　冗余容　技　实　容　。在工　实　中，　前广　应　方　大多局　于双容　列　.　一　加大，3容　更多容　方　已引　。今后　5　10年　,对于3容　或多容　方　将会成为新　。

关　：
存储　；容　；　列

Abstract:
Large storage systems are generally large array systems consisted of a great number of hard disks. With the increasing of the number of the hard disks, the chance of data loss also increased. To address this problem, researchers have conducted a lot of research. For the problem of data loss caused by the disk fault in the storage system, the recognized solution is to use coding technique to achieve fault tolerance.Nowadays, only double-erasure array codes are widely used in engineering practice. With the increasing of the system size, triple or more erasure coding schemes will draw more attentions. Experts in this field generally agreed that triple-erasure coding scheme will become the new hot spot in next 5 to 10 years.
Key words:
storage systems; fault-tolerant coding; array code

1 存储容 与 价指

20 年 ， 技 发展，大 存储 发展也十分 。当前，普 PC 存储器 容 已 到了太 别， 之 20 年前 20 MB 提 了 10 000 倍。 了传 动器之外，新型 固态存储（SSD）存储器也已 向市场。尽 单个存储器 容 发展 ，但是却仍 不上人们对存储容 增 度。 大型 “以 为中心”向 “以信息处 为中心” 变，以及信息 式增 ，人们对 存储 日 提 。 存储 上是将很多 单个存储器件（下 均以 为例）， 接口， 接整合为一个 拟 容 巨大 单一存储器，即 列。

列中 数 增多， 可 性也 之下 。工业 一 使 平均数据丢失时（MTTDL） 列 可 性。 单个 平均失效时 为 $MTTF_{disk}$，则对于包含 n 块 无冗余 列 ，其 $MTTDL$ 可 单估 为：$MTTDL=MTTF_{disk}/n$。可 ，当 n 大时，整个 可 性成 例下 。 对于 大 是不可接受 。利 冗余数据 提 可 性是公 决 一 好方 。 巧妙 将 m 块 准大小 上 数据，增加 分冗余 信息， 后存放于 n 块 上，使得 ：对于任意 k 块 失效， 可以 其他 $n-k$ 块 失效 中 数据 恢复，则 整个 是 **k 容** ，或 k 为 容 数。 分 明[1]，对于 k 容 ，$MTTDL$ 可以 似估 为：

$$MTTDL=MTTF^{k+1}_{disk}/n(n-1)\cdots(n-k)MTTR^{k} \quad (1)$$

因 ，在大 中，容 数可以 是另一 对 可 性 描 方式。市场中一 $MTTF_{disk}$ 为 10^5 左右， 修复时 $MTTR$ 一 为 10^1 左右。 据（1）式可以 出，当 数为 10^3～10^4 时，一 2 容 或是 3 容 就基 上可以 存储 容 。

于增加容 力 加 冗余 多， 外 价也将 。因 在具有 同容数 前提下，人们往往 更小 冗余度，即 $(n-m)/n$ 值，其中 n 为 数、m 为存储 户数据 数。 据 Singleton ，k 容 最小冗余度为：k/n。到 一最小值 方 做最大最小 可分(MDS) 。 前多数存储 中于 不同参数下 MDS 。

了上 指 ，任何 度与效 是 指 。 我们不 如何有效 并 处 多 中 数据 取（ 是另外一个 大 ）， 于冗余 带 外 开 。对于即便是 同 方 ， 于 / 不同，可 效 差异 大。 于在 中，最 会反映为一些二 制 ，因 常使 总 二 制异或 数 于 外冗余 带

开 。对于一个 存取 存储 ， 小块信息写操作 性 尤为 。
中 个单元所参与 平均异或 数可以 一指 ，我们 其为 **更新复 度**。

合上 ，存储 容 可以归 为寻 对如下指 优化 方 ：

- 容 性 ，容 数为 *k* 。
- 有 小（或最优） 冗余度。
- 有 小（或最优） /更新复 度。

2 性

对于单容 ， 单 奇偶 使得上 3 到最优。 典 RAID1,4,5 是使 方 。 对于 *k*>1 情况， 决就不是 么显 易 了。从 信 丰富成 中，两 有代 性 方 人们挑 出 ，并 于 决存储容 ，他们是 2 制 性 和 RS 。

2.1 多 列

图 1 所 是二 列 及 。二 列 是奇偶 推广， 图 1 很容易出它是双容 。

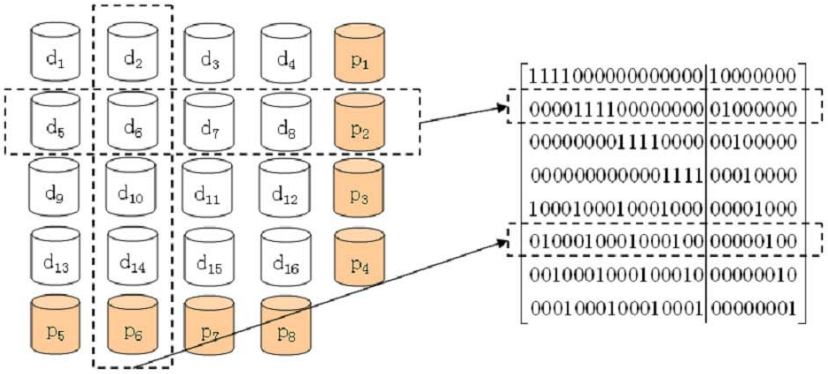

图 1 二 列 及

二 列 保持了单容 时奇偶 最优 复 度 性，但是他 冗余度不再是最优 了。

二 列 也很容易推广为 *k* 列，并且容易 明 *k* 容 性，但是 *k* 增大，冗余 大[2,3]。

2.2 Full

图 2 所 是 FULL-2 。FULL-2 可 做是二 列 推广。

$$\left[\begin{array}{c|c}
10101010101010101010 & 100000 \\
10010101101010010101 & 010000 \\
10010101010101101010 & 001000 \\
01101001011001100101 & 000100 \\
01100110100101011001 & 000010 \\
01011010010110010110 & 000001
\end{array}\right]$$

图 2 FULL-2

FULL 依 保持了最优 复 度，并且冗余度 列 好很多。 不幸 是，当 *k*>3 时，FULL-*k* 不再是 *k* 容 [4]。

2.3 RS

图 3 所 是 RS 。RS 从最佳 冗余 性出发。 到 Singleton RS 人们提出并广 应 。

$$\begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{q-2} \\ 1 & \alpha^2 & (\alpha^2)^2 & \dots & (\alpha^2)^{q-2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & (\alpha^{d-1})^2 & (\alpha^{d-1})^2 & \dots & (\alpha^{d-1})^{q-2} \end{pmatrix}$$

图 3 RS

上　　性变换可以化为　　，存储　　亦即显式区分出　　中哪些单元　于存储　　单元。可以　出，　　中　元　不再是 01，　为有　域元　　幂，故　　使　有　域　　。在　　中，有　域元　最后　是会　映射为 01 单元，时　个有　域元　一　会　映射为多个 01 单元，有　域　　也可以　分　为　些 01 单元　复　　。我们仍　以　　所　　异或　　为基　单位，则　　所　　异或　　数和　　巧妙　度有关。　前　好　域　　所　　异或　数大　为 $O(n^3)$[5]，　　复度　当　。RS　是 MDS　，故冗余度是最优　。

3　列

上　几　　各有优　，么是否存在对于多指　同时最优　k 容　　方　呢？　文[5]提出　EVENODD　，一大　只使　异或　　列　　提出并广　　。多　　列或 FULL　　二　制　性　　块　　只取一个　　单元　　。　列　则在　块　　上取多个　　单元，一　交叉　　。　　同 2　制　性　一　，只使　二　制异或　　，但冗余度却可以与 RS　　同。

3.1 EVENODD

EVENODD　　想　很　单，　块　　中取　干单元，排成方　，　后将　些单元分成不同　　，另外　加两块　　于存储　　单元。所有　　均使　　单　二　制奇偶　。

1　平　　与对

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6	Disk7
D15	D16	D17	D18	D19	P1	P5
D26	D27	D28	D29	D25	P2	P6
D37	D38	D39	D35	D36	P3	P7
D48	D49	D45	D46	D47	P4	P8
0	0	0	0	0	0	P9

平　　与对　　如　1 所　。　1 中 D 代　　户数据单元，P 代　冗余　　单元。可以出，Disk 1～5 存储　户数据单元；Disk6、7 存储冗余　　单元。Disk6　各单元为　户数据各　　平　　和，　Disk7　各单元为　户数据　　对　　和。

存储　户数据　　数　为 p（如上例中 p=5），则　　包含 p+2 块　　，前 p+1 块　　中最后一个单元为　拟 0 元，故　　实　包含 p-1 个单元，最后一块　　包含 p 个单元。可以　明，当 p 为　数时　　是双容　　。

单　　可　　时　　冗余度为（2p-1）/((p+2)(p-1)+1)。　于最后　　多出一个单元，所以冗余度　　大于最优　2/(p+2)。为了　到最优值，文　[5]中使　如下技巧：将多出　单元（即　对　交　和）叠加到　　其他单元上，　　MDS　EVENODD　如　2 所　。

2　　MDS　EVENODD

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6	Disk7

D15	D16	D17	D18	D19	P1	P5+ P9
D26	D27	D28	D29	D25	P2	P6+ P9
D37	D38	D39	D35	D36	P3	P7+ P9
D48	D49	D45	D46	D47	P4	P8+ P9

2 也可　　为如　3 所　。

3　　MDS　EVENODD

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6	Disk7
D15	D16	D17	D18	D15678	P1	P5
D26	D27	D28	D25678	D25	P2	P6
D37	D38	D35678	D35	D36	P3	P7
D48	D45678	D45	D46	D47	P4	P8

也就是　当　一　对　　和为 1 时，其他各对　　为奇　；当　一　对　　和为 0 时，其他各对　　为偶　。　就是它　命名为 EVENODD　　原因。

3.2 RDP

从　2 可以　出，为了得到冗余最优，EVENODD　　对　上　单元　更新复　度很　。　　更新　些单元　数据时　　同时更新其他 *p* 个　　单元，　对于双容　　，最优值为 2。文　[6]中　　RDP　　将　些单元　更新复　度均　到　个单元，从　有效地　　了小写操作中更新性　　不均　。包含　平　　对　　如　4 所　。

4 包含　平　　对

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6
D159	D169	D179	D189	P1	P5
D256	D256	D258	D259	P2	P6
D367	D368	D369	D356	P3	P7
D478	D479	D457	D467	P4	P8
0	0	0	0	0	P9

与 EVENODD 不同处在于，做对　　时也包含了　平　　单元　一列（因　，数据单元也　EVENODD 少了一列）。

同　　，RDP　最后一个　　多出一个单元，使得整个　　不是 MDS　。但 RDP　优势在于，　单地将多出　单元删去，　　仍　为双容　　。即得到如　5 所　　列。

5 RDP　MDS

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6
D15	D16	D17	D18	P1	P5
D256	D256	D258	D25	P2	P6
D367	D368	D36	D356	P3	P7
D478	D47	D457	D467	P4	P8

从　5 可以　出，所有数据单元　更新　　为 2 或 3，分布　EVENODD　　均匀，不会产　　方式带　　外　，但　　平均更新复　度是　同　。

3.3 Liberation

从前　几　　可以　出，所使　　方　　是　平　　加其他一　　共同　成双容　。不同之处就在于“另一　　”　不同　择。如将另一　　上　　元　作一个 01 向　，　块数据　上　单元对　些　　元　影响可　一个 01　　。如　5 中　　1 列　4 个数据单元对 Disk7 中　各　　元　影响可　　为如图 4 所　　。

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

图 4 列

在 下，前 所 更新复 度就对应 中 1 个数。于是 一个双容 列 就 变为：寻找 干个 ，使得其中 1 个数尽 少，并且任意 2 个之和为 。

在 p 为 数时，文 [7]中 Liberation 使得 $p \times p$ 1 数 不 p+1，其 p 个 可 单地描 为：各对 加一个 外单元。 k 个 外 1 单元 位 可描 为 (k(p-1)/2Mod p,1+k(p=1)/2 Mod p)。得到 如 6 所 。

6 Liberation

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6	Disk7	Disk8	Disk9
D1a	D1g	**D1fg**	D1e	D1d	D1c	D1b	P1	Pa
D2b	D2a	D2g	D2f	**D2ef**	D2d	D2c	P2	Pb
D3c	D3b	D3a	D3g	D3f	D3e	**D3de**	P3	Pc
D4d	**D4cd**	D4b	D4a	D4g	D4f	D4e	P4	Pd
D5e	D5d	D5c	**D5bc**	D5a	D5g	D5f	P5	Pe
D6f	D6e	D6d	D6c	D6b	**D6ab**	D6g	P6	Pf
D7g	D7f	D7e	D7d	D7c	D7b	D7a	P7	Pg

3.4 PDHLatin

前 些 为 MDS 充 件均为： 与 数 关（RDP 为 p+1，其他为 p+2）。它们 双容 方 均为 据一个已 单元， 后 关 与失效单元形成 式关依 恢复所有单元。 使人们 到其容 力 是任意两列 可以形成 关 。文 [8]中利 拉丁方 了 PDHLatin ，使得 不再必 关 一个 数。

所 拉丁方是指 $n \times n$ 方 中填入 n 个不同 号，使得 列 号 不 复。显 拉丁方 两列 成一个 n 元 换，所 密尔 拉丁方是指拉丁方 任何两列 成 换为单 。图 5 为一个 9 密尔 拉丁方。

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 4 & 8 & 9 & 3 & 5 & 1 & 7 & 6 \\ 3 & 1 & 9 & 2 & 8 & 7 & 5 & 6 & 4 \\ 4 & 5 & 2 & 3 & 1 & 8 & 6 & 9 & 7 \\ 5 & 7 & 4 & 1 & 6 & 9 & 8 & 3 & 2 \\ 6 & 9 & 5 & 8 & 7 & 4 & 2 & 1 & 3 \\ 7 & 8 & 6 & 5 & 9 & 2 & 3 & 4 & 1 \\ 8 & 6 & 1 & 7 & 4 & 3 & 9 & 2 & 5 \\ 9 & 3 & 7 & 6 & 2 & 1 & 4 & 5 & 8 \end{bmatrix}$$

图 5 密尔 拉丁方

从一个 定 密尔 拉丁方，我们可以 与 EVENODD 似 方 ，只不各单元对于 二 关 不再依单元所在对 位 决定， 是 据拉丁方 应位 号决定。 据图 5，得到 7 所 PDHLatin 。

7 PDHLatin

Disk1	Disk2	Disk3	Disk4	Disk5	Disk6	Disk7	Disk8	Disk9	Disk10	Disk11
Da1	Da2	Da3	Da4	Da5	Da6	Da7	Da8	Da9	Pa	P1
Db2	Db4	Db8	Db9	Db3	Db5	Db1	Db7	Db6	Pb	P2

Dc3	Dc1	Dc9	Dc2	Dc8	Dc7	Dc5	Dc6	Dc4	Pc	P3
Dd4	Dd5	Dd2	Dd3	Dd1	Dd8	Dd6	Dd9	Dd7	Pd	P4
De5	De7	De4	De1	De6	De9	De8	De3	De2	Pe	P5
Df6	Df9	Df5	Df8	Df7	Df4	Df2	Df1	Df3	Pf	P6
Dg7	Dg8	Dg6	Dg5	Dg9	Dg2	Dg3	Dg4	Dg1	Pg	P7
Dh8	Dh6	Dh1	Dh7	Dh4	Dh3	Dh9	Dh2	Dh5	Ph	P8
0	0	0	0	0	0	0	0	0	0	P9

3.5 X

上 介 几 方 到了冗余 最优，但在更新复 度方 均 于最优值，么是否可以 到两 同时最优呢？文 [9]提出 X 是一 双容 。

X 想 也很 单，仍 是在 列中 主对 和 对 两 ，但是 巧妙地将 单元分布到各个 中（ 不是像其他方 中， 单元 分 出 ， 存放于 ），使得 同时 到了两方 指 同时最优。

为了 双容 ，X 也 列中包含 列数（或 ）为 数。 为 数 p X 中， 一列包含 p-2 个 户数据单元，2 个冗余 单元。

3.6 B

是否 存在与 X 同 性 其他 方 呢？显 将两个 X 列 叠， 仍 保持最优冗余与最优更新复 度。

得到 新 ，在 数 不变 情况下， 块 关 单元数 加倍。 在实 中，为了 化实 ，我们实 上 块 关 单元数 尽 少。对于 n 块 ，在保持最优冗余与最优更新复 度 件下， 块 最少 多少个单元 关 呢？文 [10]提出 B 在双容 情况下， 决了 一 。

将 同于图 中 完全图完 一因子分 。并 据图 已有 ， 出一 各方 性 均 到最优 。从一个完全图 K_6 一 完 1 因子分 方 ，我们可以 如 8 所 双容 。

8 B

Disk1	Disk2	Disk3	Disk4	Disk5
D23	D34	D14	D12	D13
P1	P2	P3	P4	D24

， 块 包含 多 1 个 单元，并且只有一块 不包含 单元。它将 n 个 号 所有 2 元 分划为各列，并且 双容 ，因 在保持了最优冗余度与更新复 度 前提下， 到最 。因 也 做**最 最低密度 列** 。

3.7 T

对于 3 容 最 最低密度 列 之双容 复 很多，文 [11]最先 出了一 ，并利 助 明了 些参数下，3、4 容 最 最低密度 列 MDS 性。我们在文 [12]中 了同 并利 合 NRB（ 乎可分 不完全区 ） 出了 合 ，同时也 出了 明 代数 明。

T 从形式上与 B 同， 块 包含 多 1 个 单元，并且只有一块 不包含 单元。文 [12] 明了对于任意容 最 最低密度 列 均 性 。

对于普 参数 T ，或任意容 最 最低密度 列 ，仍是困 。

3.8 Weaver

前 将优化冗余 最为 一 ，同时兼 /更新复 度。但在一些 中，如 冗余 当损失可换 更好 性 或更易于 ，则也是可 择 。文 [13]从优先 /更新复 度 度，提出了易于 Weaver 。

B 、T 也可以 出，在保持更新复 度最优 前提下， 单元分布在各 中 容易 。为了 化 ，文 [13] 择具有循 对 性 列 ，也就是 ：（1）所有数据单元参与 数为常数；（2）所有 包含 单元数 为常数；（3）如 i 上 数据单元 j 参与 k 上 单元 p 所代 ，则必有对于任何 $0 \le x < n$ $i+x \bmod n$ 块 上 数据单元 j 参与 $k+x \bmod n$ 单元 p 所代 。

为了更容易地得到 k 容 ，文 [13]放宽了冗余 ，只 块 中，冗余 单元不少于 户数据单元 情况。 ，Weaver 最好冗余 只有 50%。

4

列 尽 有 很多性 优势，但在 前 存储 中， 是 RS 及层叠 RAID（如 RAID1+0 ）使 得 多。 为其原因主 为以下几个方 ：

先是实 上 单性因 ：RS 已 是工业 技 ，无 件 有成 实方 ， 层叠 RAID 原 十分 单。所以 两 实施最 单易 。与之 对， 列 多 多 、原 复 ，实施 一定 投入。 前 存储 处于发展 ，什么是“最好 ” 尚不 形成定 ，因 就 前 ，最 单 就是最好 。

其 ，受到 前大 分应 存储 影响：尽 将多个单个 件合成一个 一 拟 件会有好处，但也会有 应 。如对 10 000 块 是合成 1 个 好呢？ 是 成 10 个包含 1 000 块 小 好呢？ 据 判断。一 小一些 会更容易 、 护。 前只有 少 应 对 1 000 块 容 数据并 处 ，因 将 分为多个 小 是有 。

三， 价 低且发展 ： 使得人们可以 “奢侈”地使 存储 ，因 大型存储 建 前 处于“ 旷 ” 。 对于易实施性、易 护性、易扩展性，当前 冗余 并不是主 决定因 。

但是， 单 容 日 和、 对性 、容 、 不断变化， 存储 应 也会不断发展。明天 存储 将会 具备什么 性 形式， 我们不断地探 。

5 参 文

[1] HARTLINE J R, RAO K K. Notes on Reliability Models for Non-MDS Erasure Codes [R]. Research Report. RJ10391(A0610-035). San Jose, CA,USA: IBM. 2006.

[2] 新 , 国 . ——原 与方 [M]. 安： 安 子 技大学出 , 2001.

[3] . 存储 容 及 列 [D]. 天 ：南开大学, 2010.

[4] HELLERSTEIN L, GIBSON G A, KARP R M, et al. Coding Techniques for Handling Failures in Large Disk Arrays [J]. Algorithmic, 1994,12(3/4):182-208.

[5] BLAUM M, BRADY J, BRUCK J, et al. EVENODD: An Optimal Scheme for Tolerating Double Disk Failures in RAID Architectures [J]. ACM SIGARCH Computer Architecture News, 1994,22(1):245-254.

[6] CORBETT P, ENGLISH B, GOEL A. Row-diagonal Parity for Double Disk Failure Correction [C]//Proceedings of the 3rd USENIX Conference on File and Storage Technologies(FAST’04), Mar 31-Apr 2, 2004, San Francisco, CA,USA. Berkeley, CA, USA: USENIX Association, 2004: 14p.

[7] PLANK J S. The RAID-6 Liberation Codes [C]//Proceedings of the 6th USENIX Conference on File and Storage Technologies(FAST’08), Feb 26-29, 2008, San Jose, CA, USA. Berkeley, CA, USA: USENIX Association, 2008:97-110.

[8] WANG Gang, LIN Sheng, LIU Xiaoguang, et al. Combinatorial Constructions of

Multi-erasure-correcting Codes with Independent Parity Symbols for Storage Systems [C]//Proceedings of the 13th IEEE Pacific Rim International Symposium on Dependable Computing(PRDC’07), Dec 17-19, 2007, Melbourne, Australia. Piscataway, NJ, USA: IEEE, 2007:61-68.
[9] XU Lihao, BRUCK J. X-code: MDS Array Codes with Optimal Encoding [J]. IEEE Transactions on Information Theory, 1999,45(1):272-276.
[10] XU Lihao, BOHOSSIAN V, BRUCK J, et al. Low Density MDS Codes and Factors of Complete Graphs [J]. IEEE Transactions on Information Theory, 1999,45(6): 1817-1826.
[11] LOUIDOR E, ROTH R M. Lowest-density MDS Codes over Extension Alphabets [C]//Proceedings of the IEEE International Symposium on Information Theory (ISIT’03), Jun 29-Jul 4,2003, Yokohama, Japan. Piscataway, NJ, USA: IEEE, 2003:58.
[12] LIN Sheng, WANG Gang, STONES D S, et al. T-code: 3-erasure Longest Lowest-density MDS Codes [J]. IEEE Journal on Selected Areas in Communications, 2010, 28(2):289-296.
[13] HAFNER J L. WEAVER Codes: Highly Fault Tolerant Erasure Codes for Storage Systems [C]//Proceedings of the 4th USENIX Conference on File and Storage Technologies(FAST’05), Dec 13-16,2005, San Francisco, CA,USA. Berkeley, CA, USA: USENIX Association, 2005.

收 日 ：2010-07-07
作 介
，南开大学 专业博士 业，天 工大学副教授， 方向为存储 、 合 。
刘晓光，南开大学 专业博士 业，南开大学信息技 学学 副教授， 方向为 性 、 信息存储技 。
刚，南开大学 专业博士 业，南开大学信息技 学学 教授， 方向为 信息存储技 、并 。