P-code位矩阵方法与扩展研究

杜建军 李 华 陈 明
(重庆大学计算机学院 重庆 400030)

摘 要 以磁盘冗余阵列(Redundant Array of Inexpensive Disks,RAID)技术中新出现的P-code编码为主要对象,进行了其构造方法、编码及译码算法的详细分析,并首次运用位矩阵(Binary Distribution Matrix,BDM)的方法分析和研究了P-code码。在此基础上,对当前主要RAID-6编码的扩展即更多磁盘数量的容错问题进行了总结与探讨,提出了P-code等垂直最大距离可分码(Maximum Distance Separable,MDS)的扩展将是该领域未来研究的新方向和难点。

关键词 RAID-6,P-code编码,位矩阵,编码扩展
中图法分类号 TP333 **文献标识码** A

P-code Analyzing Based on BDM Model and Extension Research

DU Jian-jun LI Hua CHEN Ming
(College of Computer Science and Engineering,Chongqing University,Chongqing 400030,China)

Abstract P-code is one new member of Redundant Array of Inexpensive Disks(RAID) coding family,and its structure, coding method and decoding algorithm were analyzed in this paper. Further more the Binary Distribution Matrix(BDM) was firstly applied in P-code study. Based on this work,the approach of RAID-6 extension for tolerating more than three disks failure was concluded. Including P-code,the research of vertical RAID-6 extension of Maximum Distance Separable (MDS) codes should be a new study direction and a difficult research problem in this domain.

Keywords RAID-6,P-code,Binary distribution matrix,Coding extension

1 引言

磁盘冗余阵列(RAID)技术通过对阵列中的磁盘数据进行高效的冗余存储,保护数据不会因一个或多个磁盘故障而造成损坏,它被广泛地应用于现代的存储系统中[1]。由多个数据中心组成的分布式存储网络中,通过将每个数据中心映射为RAID模型中的一个磁盘进行扩展,就可以实现整个网络数据存储的容错保护。

目前RAID分为几个不同的标准,RAID-1,RAID-4和RAID-5能够精确恢复单个磁盘的故障,RAID-10和RAID-01通过磁盘镜像能够容忍多个磁盘故障,但是源数据磁盘和镜像同时发生故障时,就会造成数据的永久丢失,其数据存储效率只有50%。RAID-6能够容忍任意两个磁盘的故障,存储效率和可靠性都很高,P-code就是一种全新的RAID-6编码。本文在对P-code编码进行详细阐述的基础上,首次运用位矩阵的方法对其进行了分析,并针对当前主要RAID-6码的扩展问题,即能容许3个设备故障的编码策略进行了总结归纳和探讨研究。

2 P-code编码及位矩阵分析

2.1 P-code编码方法

P-code是一种新出现的RAID-6的垂直编码,其结构与最低密度编码类似[2]。最低密度编码的设计借用了图论和生成矩阵的方法,而P-code采用标签的编码方式更有利于系统的实现。P-code定义在一个$[(p-1)/2]*(p-1)$的矩阵上,其中p为质数。矩阵中,每列代表一个磁盘,用$d_1,d_2,\cdots,d_{p-1}$来标识,每个磁盘进行分块,其中第一个块为奇偶块,由各个对应数据块通过XOR运算得到,剩下的$(p-3)/2$个块为源数据块。这样,每个数据块就可以用一组无序的元组来(m,n)表示,数据块的标签集合C可以表示为式(1)。

$$C=\{(m,n)\mid(1\leqslant m,n\leqslant p-1),m\neq n,m+n\neq p\} \quad (1)$$

集合C可以划分为子集$C_i(1\leqslant i\leqslant p-1)$,即

$$C_i=\{(m,n)\mid(m,n)\in C,m+n\equiv i \bmod p\} \quad (2)$$

集合C和C_i有如下的关系:

$$1.\ C=\bigcup_{i=1}^{p-1}C_i; \quad (3)$$

$$2.\ C_i\cap C_j=\emptyset(i\neq j); \quad (4)$$

$$3.\ |C_i|=\frac{(p-3)}{2},1\leqslant i\leqslant p-1; \quad (5)$$

这样,每一个元组(m,n)对应一个磁盘源数据分块,加上按顺序排列的奇偶块,就构成了如图1所示的P-code编码。

d1 (m1=4)	d2 (m2=1)	d3 (m3=5)	d4 (m4=2)	d5 (m5=6)	d6 (m6=3)
(1)	(2)	(3)	(4)	(5)	(6)
(3,5)	(3,6)	(4,6)	(5,6)	(1,4)	(1,5)
(2,6)	(4,5)	(1,2)	(1,3)	(2,3)	(2,4)

图1 P-code编码($p=7$)6个磁盘的分块标签

到稿日期:2010-04-25 返修日期:2010-08-16 本文受国家科技支撑计划:可信任互联网(2008BAH37B04)资助。
杜建军(1974-),男,博士生,主要研究方向为计算机安全、网络应用及计算机体系研究,E-mail:dujianjun@cqu.edu.cn;陈 明(1978-),男,博士生,主要研究方向为可信计算、计算机网络应用研究。

每一列(磁盘)的标签中,都有一个缺省数字 m,可用式(6)计算。

$$2m_i=i \bmod p \text{ 或 } m_i=\begin{cases}\dfrac{i}{2}(i\text{ 为偶数})\\ \dfrac{i+p}{2}(i\text{ 为奇数})\end{cases} \tag{6}$$

这样,以每个奇偶块标签为起始,就组成了 $p-1$ 条数据逻辑链,每个源数据块分别属于两条不同的数据链,每条数据链贯穿了 $p-2$ 个磁盘,图 2 中用虚线表示。

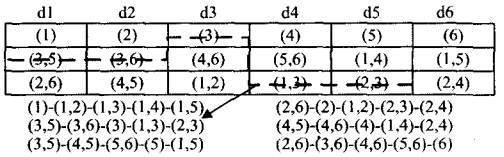

图 2 P-code 编码中的逻辑数据链

P-code 是一种线性码,所以其最小行距 d 可用 Singleton[3]式(7)计算得到,$d\leqslant 3$。

$$d\leqslant n+1-\log_q{}^h N \tag{7}$$

$$\begin{cases}a(k)=\{(-1)^k(k+\dfrac{1}{2})(n_1-n_2)+\dfrac{n_1+n_2}{2}\} \bmod p\\ \{a(k)\mid k=0,1,\cdots,p-1\}\\ 0\leqslant n_2<n_1\leqslant p-1\end{cases} \tag{8}$$

$$a(k_1)-a(k_2)=\{((-1)^{k_1}(k_1+\frac{1}{2})-(-1)^{k_2}(k_2+\frac{1}{2}))(n_1-n_2)\} \bmod p\neq 0 \tag{9}$$

由于 p 是质数,通过式(8)和式(9)可以证明 d 不小于 2,即 $d=3$,满足边界条件,所以 P-code 编码是最大距离可分码(MDS)。

2.2 P-code 磁盘故障重建算法

当有一个磁盘故障发生时,由于该磁盘的每个块都处于 P-code 编码的至少一个逻辑数据链中,而且这些数据链至多有一个数据丢失,因此直接通过 XOR 运算就可以恢复因故障丢失的数据。

当有两个磁盘发生故障时,由于知道发生故障磁盘在阵列中的索引值 d_i(与奇偶块编号对应),通过索引值很容易计算出缺省数字 m,从而建立恢复数据链。通过恢复数据链的回溯,即从起点到终点依次进行 XOR 计算,就可以恢复数据。

图 3 中,假设磁盘 d_3 和 d_4 损坏(可以是任意两个磁盘),相应的恢复数据链为 5-6-4-0-3-1-2,即(5,6)→(4,6)→(4)和(1,2)→(1,3)→(3),其算法如图 4 所示。

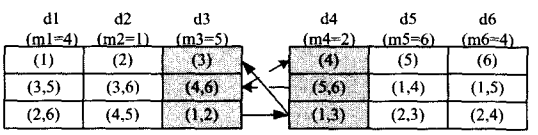

图 3 两个磁盘故障的数据链恢复链顺序

```
1.用公式(6)计算出两个磁盘的缺省数字mi和mj;
2.计算出两个磁盘的恢复数据链,以mi为起点,mj为终点;
3.并行执行重建线程;
 线程1: 取恢复链起点值,u←mi;
 While(u≠0){
通过在同一个数据链P(u)中的其它数据块重建丢失数据;
继续取恢复链中的下一个u值;}
线程2: 取恢复链起点值,v←mj;
 While(v≠0){
通过在同一个数据链P(v)中的其它数据块重建丢失数据;
继续取恢复链中的下一个v值;}
```

图 4 两个磁盘故障的 P-code 重建算法

2.3 P-code 性能及位矩阵分析

由 RAID-6 编码的代表 X-code 和 RDP 算法可以分析得到,其构建每个数据块分别需要 $2-2/(p-2)$ 和 $2-2/(p-1)$ 次 XOR 运算。而 P-code 编码算法表明,每个奇偶冗余位的建立需要($p-4$)次 XOR 运算,完成整个编码共需要($p-1$)($p-4$)次运算。因为共有($p-1$)($p-2$)/2 个数据块,所以对每个数据块而言构建编码的时间复杂度是 $2-2/(p-3)$次 XOR 运算,要优于 X-code 和 RDP。通过位矩阵的方法,可以更直观地分析 P-code 的性能。

用位矩阵构建奇偶类编码的方法最先被应用于 Cauchy Reed-Solomon 码[4]。假设有 k 个源数据设备和 m 个编码方案增加的设备,每个设备存储 w 位数据,就可以用一个 $w(k+m)\times wk$的在群 $GF(2)$上的矩阵来表示编码方案,这个矩阵叫二进制分布矩阵(BDM)或位矩阵,图 5 就是一个用位矩阵表示的编码方案。

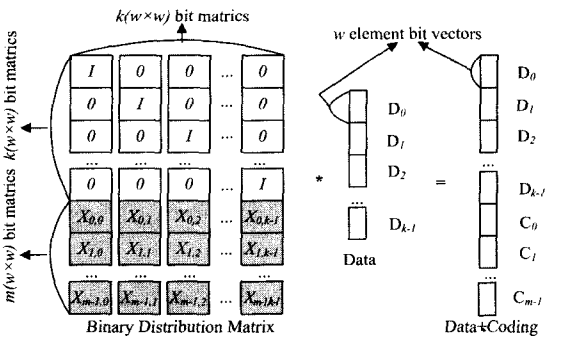

图 5 位矩阵编码方案

图 5 中左边是 BDM,其上面 wk 行由 $w\times w$ 的单位阵和 0 矩阵组成,下面 mw 行是由不同编码方案决定的由 $w\times w$ 的 bit 组成的矩阵 $X_{mw\times kw}$。BDM 与 wk 维的源数据 Data 相乘就可以得到最后的($k+m$)$\times w$ 维的编码系统。位矩阵中单位阵和 0 矩阵的组成是固定的,运用位矩阵编码实际就是对矩阵 $X_{mw\times kw}$ 的研究。矩阵 $X_{mw\times kw}$ 由 0 和 1 的 bit 组成,0 表示不进行运算,1 表示 XOR 运算,所以矩阵中 1 的总和 H 就是编码时进行 XOR 运算的总次数。显然,H 的值越低,编码的效率越高,根据线性编码原理可以得到 H 的最小理论值为 mkw。用位矩阵描述 RAID-6 编码中的 EVENODD,RDP 和 Cauchy Reed-Solomon 编码可以得到图 6 的对应 $X_{mw\times kw}$矩阵,其中阴影表示 bit 为 1,白色为 0。可以很容易得出,Cauchy Reed-Solomon 编码的效率最高,其解码相对要求更高。

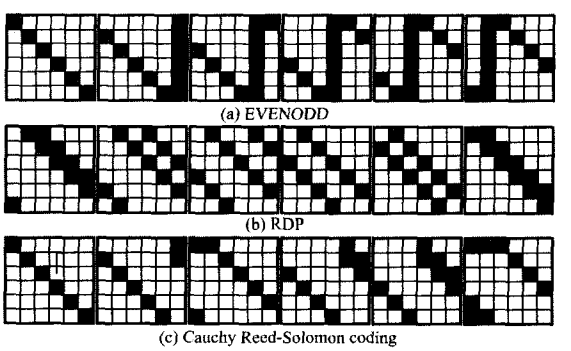

图 6 3 类编码方案的 $X_{mw\times kw}$矩阵表示($k=6$,$w=6$)

用位矩阵的方法可以很直观地得出水平编码(EVENODD,RDP 和 Cauchy Reed-Solomon 编码)的效率,而 P-code 属于垂直 RAID-6 编码,无法直接应用位矩阵描述,需进行形式转

换，图 7 是应用位矩阵的思想描述出的 P-code 编码方案。

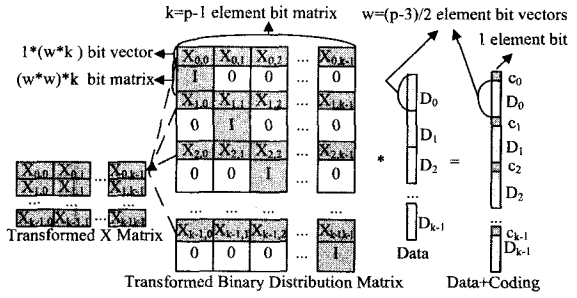

图 7　转换形式的 P-code 位矩阵编码方案

对比图 7 和图 5 的位矩阵可以发现，由于垂直编码的奇偶冗余校验数据分布在各个磁盘上，不同于水平编码单独的磁盘存储，因此其 X 矩阵的各个行与固定格式的单位矩阵和零矩阵交错分布。由于只有 X 矩阵的位信息代表了进行 XOR 运算的次数，因此将其各个行重新合并就可以采用与位矩阵等价的研究方法。对应图 1 的 P-code 编码标签分布，通过变换形式的位矩阵方法就得到了图 8 的 X 矩阵。该矩阵中的 XOR 次数与理论值是一致的，其性能要优于 EVENODD，RDP 和 Cauchy Reed-Solomon 编码。由于 P-code 编码的磁盘阵列分布不同于 EVENODD 等编码，因此其 X 矩阵的几何分布也完全不同。

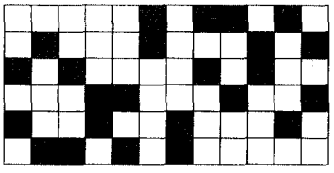

图 8　P-code 编码的 X 矩阵（$p=7,k=6,w=2$）

3　RAID-6 与 P-code 编码扩展

3.1　RAID-6 扩展方法

计算机高速发展造成数据存储量的几何级增加，同时设备故障率也加大，这就对磁盘和数据中心存储可靠性提出了更高的要求。RAID-6 的标准是容忍两个磁盘的故障，对 RAID-6 编码的扩展就是指能容许 3 个或更多磁盘同时故障的编码方案研究，该方向的研究目前还没有形成标准化的定义。

水平编码的数量在当前 RAID-6 编码方案中占到了三分之二以上，其代表有 EVENODD，RDP 和 Cauchy Reed-Solomon 编码等。借鉴和利用 RAID-6 编码的思想和方法就可以扩展出能容许 3 个设备同时故障的编码策略，目前的水平编码扩展方案中有 Blaum 码、EEOD 码、类 Reed-Solomon 编码等。Blaum 码[5]是通过多项式环上的一组线性方程实现的，其解码算法不易实现、复杂度高。EEOD 码[6]通过校验方程组上的回路表示策略实现了对 EVENODD 码的扩展，译码过程涉及到对图回路的叠加，空间利用率和性能有损失。类 Reed-Solomon 码[7]通过环形序列矩阵（circular permutation matrices，CPM）对 Reed-Solomon 进行扩展，使得编码能容许 3 个磁盘故障，而通过 Cauchy 矩阵和 CPM 的方法能扩展出容忍 4 个及更多磁盘故障的策略。类 Reed-Solomon 码的优点是其属于 MDS 码，但是其编码及译码在实现和性能上都存在难题。

另一类扩展方法是混合型的，即在水平编码的基础上加入了垂直方向上的校验信息，WEAVER 码、HoVer 码、HDD 码等都属于这类扩展。这类编码都不是 MDS 码，即空间效率不是最佳理论值且冗余磁盘数目呈非线性增长，还需要高斯消元求解等，所以该类扩展方法有明显的局限性。

笔者研究 EVENODD 码的几何结构时发现，其编码方案中的冗余校验信息由两部分组成，即数据逻辑链包含水平和逆对角线两个方向。考虑增加一个正对角线方向上的冗余校验信息数据链，能否扩展成容许同时 3 个磁盘故障且是 MDS 码的方案？其扩展方法的思路如图 9 所示。

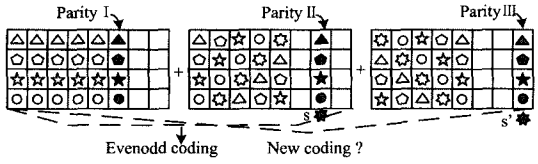

图 9　EVENODD 编码扩展结构

经研究发现，在 EVENODD 码逻辑结构上增加这样的校验信息数据链的确能够使得新的编码策略容忍 3 个磁盘故障。美国学者也用类似的思路对 EVENODD 码进行了扩展，并且命名为 Star 码[8]。Star 码很好地继承了 EVENODD 码的结构和优点，并且易于实现。

3.2　P-code 编码扩展

P-code 码本身在使用时就必须考虑扩展问题，即不满足编码条件的磁盘阵列分布如何使用 P-code 码的问题。该问题可以通过编码时增加逻辑空磁盘，使得整个磁盘阵列几何分布满足 p-1 的编码条件得到解决。

在对 P-code 码的研究中，很自然地会思考如何将其扩展到容许 3 个磁盘故障的情况。P-code 属于垂直 RAID-6 编码，很奇怪的是目前却不存在一个完全的垂直 MDS 码能容许 3 个或更多磁盘故障的编码方案。

笔者在 P-code 多种扩展策略设计中始终无法找到一种覆盖所有磁盘故障的方案，同时典型的垂直 RAID-6 编码 X-code 在设计时也强调将其扩展到更多数量的磁盘容错是很困难的。究其原因可以发现，现在对 RAID-6 编码的扩展主要是两类方法。一是几何方法，即通过对编码结构的几何特性的观察进行扩展，该方法在面对 3 个及以上磁盘故障时非常受限，并且编码和译码算法的验证也非常困难。二是代数法，即把每个数据磁盘定义为一个已知向量，将存储校验冗余信息的磁盘定义为求解的未知向量，在群 $GF(2^w)$ 上建立多项式方程组，通过计算满足特定数量解的情况建立编码策略。代数法很适用于水平编码的扩展，类 Reed-Solomon 码便是使用该方法的典型例子。但是垂直编码时由于冗余校验信息与源数据一起分布在各个磁盘空间，无法将每个磁盘定义为单独的向量从而使用现行的代数法寻找扩展方案。所以，对垂直 RAID-6 编码，包括 P-code 码的扩展问题将期待进一步的研究探索。

结束语　在计算机存储系统中，磁盘冗余阵列（RAID）技术一直具有重要的理论和应用研究价值。本文对 RAID-6 编码家族中的新成员 P-code 码在编码结构、编码及译码算法等方面进行了详细阐述。在此基础上，首次运用了位矩阵（BDM）的研究方法，通过位矩阵的形式变换和抽象组合对 P-code 编码的结构性能进行了分析。研究表明 P-code 在冗余率、编码及译码复杂度上都优于现有的主要 RAID-6 编码策略，但是其几何结构不同于主流磁盘阵列分布，需要在运用中

拓展。以 RAID-6 编码策略为基础，总结了已有的能容许 3 个设备故障的多种编码方案，发现扩展都是基于水平及混合编码的。概括了造成该现象的原因和难点，垂直 RAID-6 编码的扩展应用将是今后工作的研究方向。

参 考 文 献

[1] Plank J S. The RAID-6 Liberation Codes [C]// Proceedings of FAST '08:6th USENIX Conference on File and Storage Technologies. USENIX Association, 2008:97-111

[2] Chao Jin, Hong Jiang, et al. P-Code: A New RAID-6 Code with Optimal Properties [C]// Proceeding of ICS'09: International Conference of Supercomputing. York Town Heights, New York, USA, 2009

[3] Blaum M, Roth R M. On Lowest Density MDS Codes [J]. IEEE Transactions on Information Theory, 1999:45(1):46-59

[4] Blomer J, Kalfane M, et al. An XOR-based Erasure Resilient Coding Scheme [R]. TR-95-048. International Computer Science Institute, August 1995

[5] Blaum M, Bruck J, Vardy A. MDS Array Codes with Independent Parity Symbols [J]. IEEE Trans. Inform. Theory, 1996, 42(2):529-542

[6] 万武南，索望，等. 基于 EEOD 码的一种有效的数据分布策略[J]. 电子科技大学学报，2007，36(5)：834-838

[7] Feng Gui-liang, Deng R H, et al. New Efficient MDS Array Codes for RAID Part I: Reed-Solomon-Like Codes for Tolerating Three Disk Failures [J]. IEEE Trans. on Computers, 2005, 54(9):1071-1080

[8] Cheng Huang, Xu Li-hao. STAR: An Efficient Coding Scheme for Correcting Triple Storage Node Failures [J]. IEEE Trans. on Computers, 2008, 57(7):889-901

[9] 余成波，张冬梅，许超明，等. 缺陷硬盘数据恢复新技术中的影子方法[J]. 重庆工学院学报：自然科学版，2009，23(6)：40-44

(上接第 270 页)

3 层，相应地知识源也分为领域知识源、环境知识源和控制知识源 3 层。控制机构是黑板系统中最复杂同时也是最重要的部分，主要由刺激、响应、判断和控制 4 大部分模块组成，用于监督和调度知识源和黑板之间的交互[10]。

(1) 刺激模块

刺激模块根据黑板状况的变化自动刺激合适知识源。黑板的层次关系通过刺激顺序来体现，先从最低级黑板开始刺激知识源，依次类推。

(2) 响应模块

响应模块将知识源受刺激后产生的中间结果存入匹配链表，其算法步骤如下：

Step 1　根据刺激模块产生的中间结果找到相应匹配链表；

Step 2　在该链表中寻找匹配节点：

a) 若有，则对该节点的匹配记数加 1，并比较节点与中间结果的可信度大小，若节点可信度小，则用中间结果的可信度代替之；

b) 否则，创建新节点并赋值，即写入该中间结果内容(结论、可信度及决策优先级)，匹配累计记数设置为 1，并将新节点添加到匹配链表末尾。

(3) 判断模块

判断模块算法对匹配链表节点逐一比较，找出匹配累计记数和可信度都为最大者。若匹配累计记数和可信度最大的节点不是同一个，则应进行取舍。相应的算法步骤如下：

Step 1　搜索匹配链表中匹配累计记数及可信度均最大项；

a) 若有且仅有一项，则该项即为判断结果；

b) 若有且多于一项，则取决策优先级最高者(称为决策优先级法)；

c) 若匹配累计记数及可信度最大者不为同一节点，则采用可信度最大-最小阈值法；若匹配累计记数最大节点的可信度小于某阈值，而可信度最大节点大于某阈值，则取后者为判断结果；反之，取前者。

Step 2　用判断结果更新黑板中的相应内容。

(4) 控制模块

完成总控，决定另外 3 个模块的执行顺序。

结束语　信息融合最初起始于 C^3I 系统，并由多传感器融合发展而来。与单传感器数据处理相比，多传感器数据融合(多源信息融合)具有很多突出的优点。当然，相应地，多传感器数据融合(多源信息融合)的复杂性也大大增加了，由此在设计算法和实现系统时也就产生了一系列不利因素，例如成本提高、功耗增加和隐蔽性降低等。相比而言，本文混合式实时智能信息融合系统具备以下特点：

(1) 由于采用混合式数据处理结构，因此数据处理量适中、实时性强；

(2) 由于采用知识库系统，且运用神经网络技术来实现知识的自动获取，因此智能性强，具有良好的自学习、自适应能力；

(3) 由于采用 3 层数据库结构，因此可兼顾系统的实时高效与安全可靠；

(4) 由于采用黑板进行数据和知识的调度，因此可减少交叉访问，提高访问和处理速度。

参 考 文 献

[1] Waltz E, Llinas J. Multisensor data fusion[M]. Boston: Artech House, 1990

[2] 刘同明，等. 信息融合技术及其应用[M]. 北京：国防工业出版社，1998

[3] Hall D L. Mathematical techniques in multisensor data fusion [M]. New York: Artech House, 1992

[4] 罗明，等. 多源数据融合系统中基于知识处理的推理策略[J]. 电讯技术，1999(3)：42-45

[5] Wang Jun, et al. COM-based software architecture for multisensor fusion system[J]. Information Fusion, 2001(2):261-270

[6] 李晓强，等. 基于关系数据库的知识库结构设计[J]. 计算机工程与应用，2001(24)：102-103

[7] 王宗军，等. 嵌入神经网络专家系统的智能化城市评价 DSS[J]. 系统工程理论与实践，1995(4)：25-31

[8] Ricardo M, Fricks A. Performance analysis of distributed real-time databases[J]. Performance Evaluation, 1999, 35:145-169

[9] Llinas J, et al. Blackboard concepts for data fusion applications [J]. International Journal of Pattern Recognition and AI, 1993, 7(2):285-308

[10] 徐从富，等. 面向通侦信息融合的多层黑板模型[J]. 电子学报，2001，29(3)：361-363