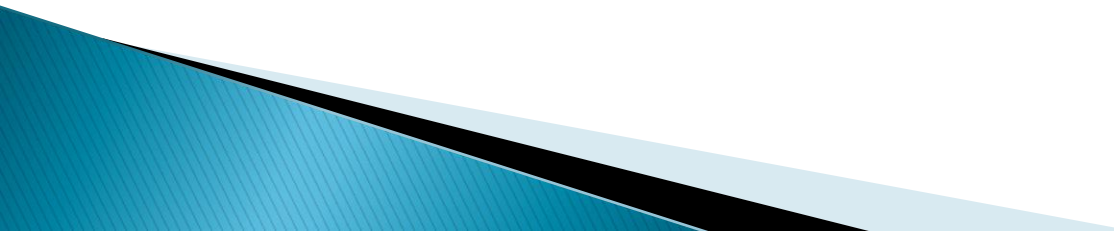


ELEMENTE INTRODUCTIVE PRIVIND INFORMAȚIA

– *Curs3* –

Subiecte abordate:

- definirea și măsurarea informației;
 - codificarea informației: coduri de lungime fixă, coduri de lungime variabilă și eficiența codificării;
 - detectarea și corectarea erorilor.
- 

Măsurarea informației

- ▶ Memorarea, regăsirea și prelucrarea informației reprezintă operații de bază întâlnite în studiul oricărui capitol al științei calculatoarelor.
- ▶ *Domeniul ingineriei consideră că informația înlătură/elimină incertitudinea. Astfel, informația nu are nici o legătura cu cunoștința sau semnificația. Informația este, în mod natural, “ceea ce nu se poate prezice”.*
- ▶ În cadrul procesului de edificare a domeniului teoriei informației s-au dat și precizat o serie de definiții formale privind conținutul informațional al unui mesaj (Hartley - 1928, Kolmogorov - 1942, Wiener - 1948, Shannon - 1949). Sub forma cea mai generală, informația este considerată ca o înlăturare/eliminare a incertitudinii.

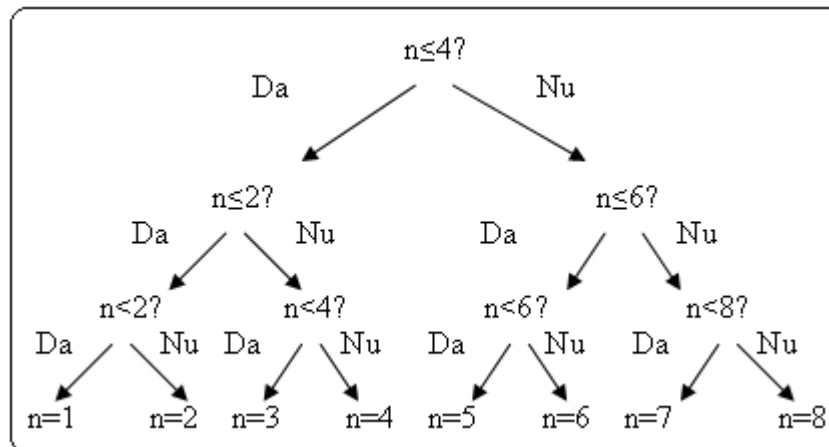
- ▶ Se consideră un sistem, care reprezintă o mulțime formată din **n** obiecte, având proprietatea că fiecare obiect **i** posedă o probabilitate independentă **p_i** de apariție. Incertitudinea **H**, asociată acestui sistem, este definită ca:

$$H = - \sum_{i=1}^n p_i \times \log_2(p_i)$$

- ▶ Se presupune existența unei urne cu bile numerotate de la 1 la 8. Probabilitatea de a extrage o cifră dată, în urma unei trageri, este egală cu 1/8. Incertitudinea/informația medie asociată cu numărul selectat poate fi calculată cu ajutorul formulei de mai sus:

$$H = \sum_{i=1}^8 \left(\frac{1}{8} \times \log_2 \left(\frac{1}{8} \right) \right) = - \log_2 \left(\frac{1}{8} \right) = \log_2 8 = 3$$

- ▶ *Pentru a măsura incertitudinea asociată sistemului s-a folosit o unitate de măsură numită bit. Un bit este o măsură a incertitudinii sau a informației asociate unei condiții cu două stări: fals/adevărat, închis/deschis etc.*
- ▶ Cantitatea de informație care se câștigă este egală cu cantitatea de incertitudine înlăturată (în cazul de față, prin aflarea numărului extras din urnă). În situația numărului selectat din urna, s-a calculat că sunt necesari trei biți de informație pentru a afla numărul de pe bila extrasă.
- ▶ Din punctul de vedere al definiției bitului, aceasta înseamnă că este permisă punerea a trei întrebări cu răspuns de tipul Da - Nu, pentru a cunoaște numărul extras:



Interogări și răspunsuri binare privind selectarea unei bile numerotate dintr-o urnă care conține 8 bile.

Numărul bilei	Schema 1 a mesajului	Schema 2 a mesajului
1	000	001
2	001	010
3	010	011
4	011	100
5	100	101
6	101	110
7	110	111
8	111	000

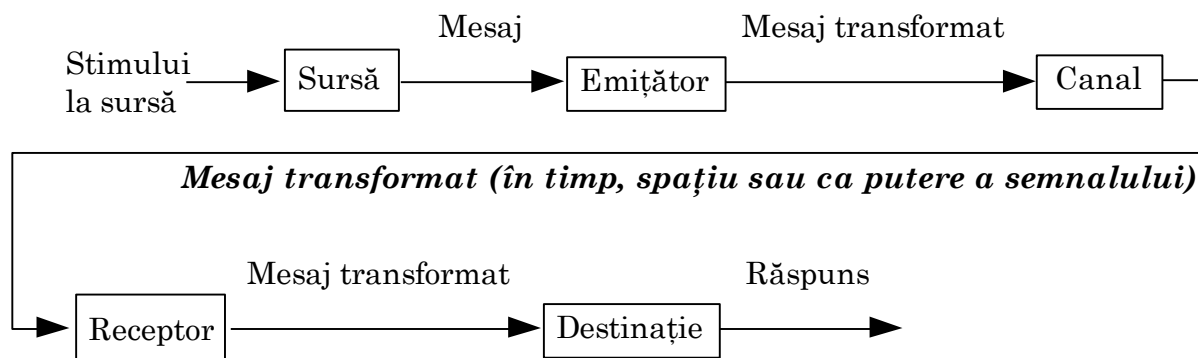
Schemele posibile de codificare pentru cele opt numere înscrise pe bilele din urnă

- ▶ În orice schemă de codificare, care reprezintă o mulțime cu n elemente, cu probabilități egale de selectare, cel puțin unul din coduri trebuie să aibă o lungime egală sau mai mare decât măsura informației asociată mulțimii date, adică:

$$H = - \sum_{i=1}^n p_i \cdot \log_2(p_i)$$

Astfel, în cazul unui sistem fizic, care se poate afla în 13 stări distincte, codificarea stărilor se va realiza cu mesaje având lungimea de 4 biți. Mesajele de 4 biți lungime vor putea codifica 16 stări distincte.

În teoria comunicațiilor se consideră că mesajele recepționate, dar incomplet înțelese, conțin zgomot. Diagrama transmiterii semnalelor, după Shannon, este următoarea:



- Pentru a asigura transferul informației între sursă și destinație trebuie considerate trei aspecte:

- ❖ sintactic

- ❖ semantic

- ❖ pragmatic.

Aspectul sintactic este legat de forma fizică de reprezentare a informației transmise.

Semantica se referă la semnificația atașată reprezentării sintactice.

Aspectul pragmatic implică acțiunea întreprinsă, ca urmare a interpretării (sensului atașat) informației.

O comunicație corectă trebuie să considere toate cele trei aspecte mai sus menționate.

- ▶ Dacă un sistem se caracterizează prin **n** evenimente cu probabilități egale de apariție, în condițiile în care s-au obținut informații, care au redus cele **n** evenimente la **m** evenimente s-au obținut: $\log_2 \left(\frac{n}{m} \right)$ biți de informație.

- ▶ **Entropia** este cantitatea medie de informație conținută într-un șir de date.

$$entropia = \sum_{i=1}^N \left(\frac{M_i}{N} \right) \cdot \log_2 \left(\frac{N}{M_i} \right)$$

unde:

$\left(\frac{M_i}{N} \right)$ ▪ reprezintă probabilitatea mesajului **i**

$\log_2 \left(\frac{N}{M_i} \right)$ ▪ constituie informația din mesajul **i**

Codificarea informației

- ▶ Codificarea informației se referă la reprezentarea acesteia. O codificare corespunzătoare și eficientă se reflectă pe mai multe niveluri:
 - ★ la nivelul *echipamentelor de calcul și de memorare* influența se manifestă în legătură cu numărul de componente
 - ★ la nivelul *eficienței*, influența se referă la numărul de biți utilizați
 - ★ la nivelul *fiabilității* influența se manifestă sub aspectul zgomotului
 - ★ la nivelul *securității* influența se referă la criptare

CODIFICAREA CU LUNGIME FIXĂ

- ▶ Se pot utiliza în cazul în care evenimentele au aceeași probabilitate de apariție.
- ▶ Un asemenea cod trebuie să folosească un număr suficient de biți pentru a putea reprezenta conținutul informațional.
- ▶ **Exemplu:**
- ▶ —> în cazul cifrelor zecimale {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} se folosește un cod de 4 biți (binar-zecimal), denumit BCD (Binary Coded Decimal).
- ▶ Lungimea codului (numărul de biți) rezultă ca fiind 3,322 conform relației:

$$\sum_{i=1}^{10} \left(\frac{1}{10} \times \log_2 10 \right) = 3,322$$

CODIFICAREA NUMERELOR

- ▶ Numerele pozitive se pot codifica direct, sub forma unei secvențe de biți, cărora li se asociază ponderi diferite. De la dreapta la stânga, aceste ponderi reprezintă, în ordine crescătoare, puteri ale lui 2. Valoarea v a unui număr de n biți, codificat în acest mod, este dată de expresia:

$$v = \sum_{i=0}^{n-1} 2^i \cdot b_i$$

unde b_i constituie rangul i (bitul i) al reprezentării.

2^{11}	2^{10}	2^9	2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
0	1	1	0	1	0	1	1	1	0	0	1

Se obține numărul 1721 în zecimal

Codificarea în bazele 8 și 16

Baza 8

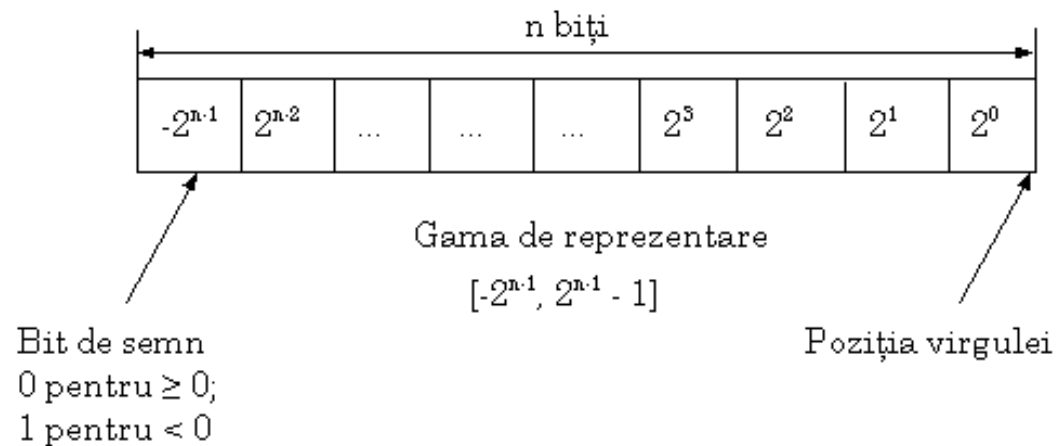
Triada binară	Cifra octală
000	0
001	1
010	2
011	3
100	4
101	5
110	6
111	7

Baza 16

Tetrada binară	Cifra hexa	Tetrada binară	Cifrahexa
0000	0	1000	8
0001	1	1001	9
0010	2	1010	a
0011	3	1011	b
0100	4	1100	c
0101	5	1101	d
0110	6	1110	e
0111	7	1111	f

NUMERE CU SEMN, REPREZENTAREA ÎN COMPLEMENT FAȚĂ DE 2

► Structura:



$$47_{(10)} = 00101111_{(2)}$$

$$-47_{(10)} = 11010001_{(2)}$$

CODIFICAREA CU LUNGIME VARIABILĂ

- ▶ În situațiile când evenimentele au probabilități diferite de apariție, se obține mai multă informație atunci când se produce un eveniment cu probabilitate mică de apariție, decât în cazul producerii unui eveniment cu probabilitate mare de apariție.

❖ Informația furnizată de apariția evenimentului : $i = \log_2 \left(\frac{1}{p_i} \right)$ biți

❖ Entropia informației este : $\sum p_i \times \log_2 \left(\frac{1}{p_i} \right)$

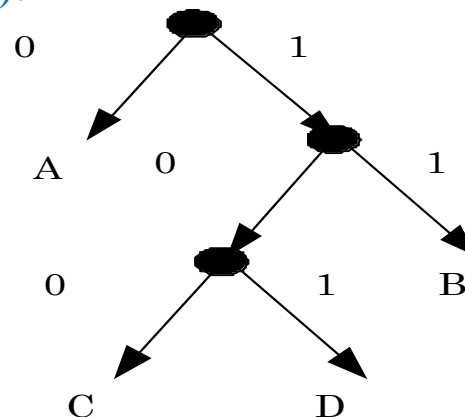
Exemplul 1:

Se consideră 4 evenimente A,B,C,D, cu probabilitățile de apariție și cu codurile asociate, conform tabelului de mai jos:

Eveniment	A	B	C	D
Prob. Apariție (p_i)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
Codificare	0	11	100	101

Informația medie = $0,5 \times 1 + 0,25 \times 2 + 2 \times 0,125 \times 3 = 1,750$ biți

O schema de decodificare a unui mesaj ce conține codurile unei secvențe de evenimente A, C, D, A, B, C etc, se conformează următoarei structurii arborescente (**arborele de decodificare Huffman**):



Exemplul 2:

Se consideră suma a două zaruri, în sensul evaluării conținutului de informație existent în suma obținută la o aruncare.

Suma	Posibilități	
2	1+1	
3	1+2, 2+1	
4	1+3, 2+2, 3+1	
5	1+4, 2+3, 3+2, 4+1	
6	1+5, 2+4, 3+3, 4+2, 5+1	
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1	
8	2+6, 3+5, 4+4, 5+3, 6+2	
9	3+6, 4+5, 5+4, 6+3	
10	4+6, 5+5, 6+4	
11	5+6, 6+5	
12	6+6	

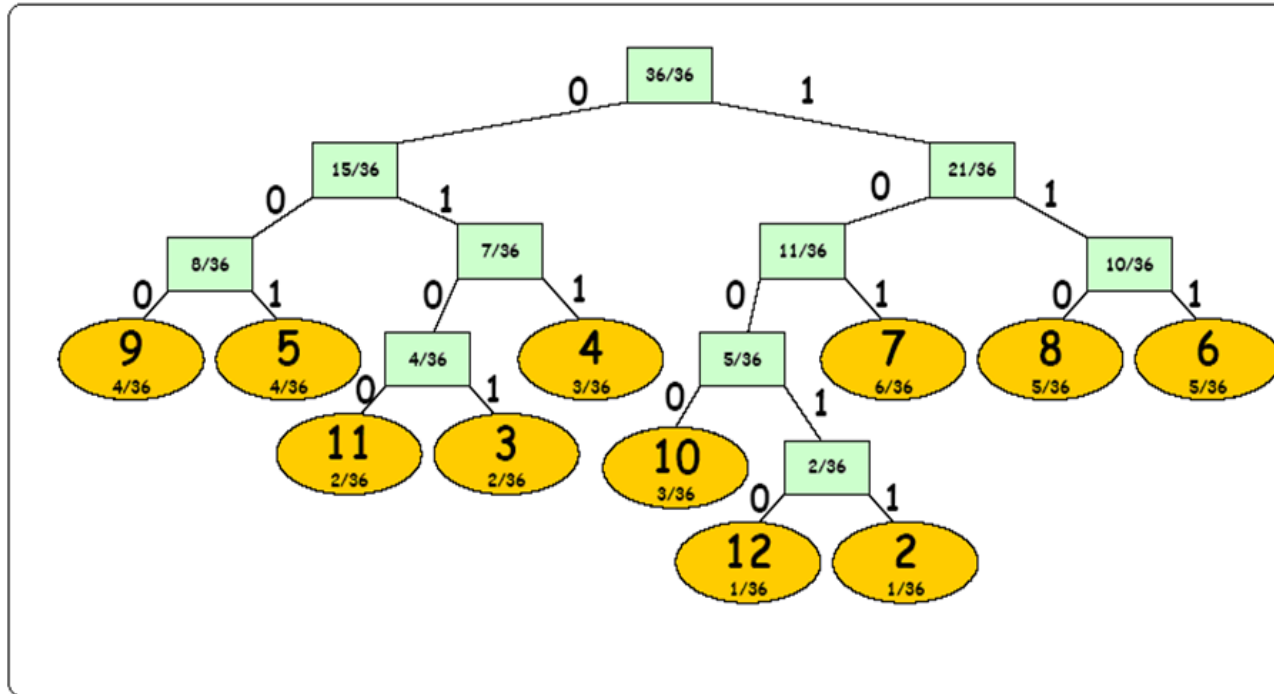
$$i_2 = \log_2 \frac{36}{1} = 5,170 \quad i_3 = \log_2 \frac{36}{2} = 4,170 \quad i_4 = \log_2 \frac{36}{3} = 3,585 \quad i_5 = \log_2 \frac{36}{4} = 3,170 \quad i_6 = \log_2 \frac{36}{5} = 2,848$$

$$i_7 = \log_2 \frac{36}{6} = 2,585 \quad i_8 = \log_2 \frac{36}{5} = 2,848 \quad i_9 = \log_2 \frac{36}{5} = 2,848 \quad i_{10} = \log_2 \frac{36}{3} = 3,585$$

$$i_{11} = \log_2 \frac{36}{12} = 4,170 \quad i_{12} = \log_2 \frac{36}{1} = 5,170$$

$$i_{med} = \sum_{j=2}^{12} \left(\frac{M_j}{N} \times \log_2 \left(\frac{N}{M_j} \right) \right) = \sum_{j=2}^{12} \left(p_j \times \log_2 \left(\frac{1}{p_j} \right) \right) = 3,275$$

Construirea arborelui binar



Arborele de decodificare Huffman : 2 - 10011; 3 – 0101; 4 – 011; 5 – 001; 6 – 111; 7 – 101; 8 – 110; 9 – 000; 10 – 1000; 11 – 0100; 12 - 10010.

Eficiența codificării

- ▶ *Eficiența unei metode de codificare poate fi măsurată prin diferența între conținutul informațional (entropia) al unui șir de simboluri și dimensiunea medie a codului.*
- ▶ Dimensiunea medie a codului stabilit pentru sumele obținute la aruncarea a două zaruri se calculează astfel:

$$\begin{aligned}d_{med} &= \frac{1}{36} \times 5 + \frac{2}{36} \times 4 + \frac{3}{36} \times 3 + \frac{4}{36} \times 3 + \frac{5}{36} \times 3 + \frac{5}{36} \times 3 + \frac{4}{36} \times 3 + \frac{3}{36} \times 4 + \frac{2}{36} \times 4 + \\ &+ \frac{1}{36} \times 5 = 3,306\end{aligned}$$

- ▶ Rezultatul obținut se apropie destul de mult de informația medie (3,275)

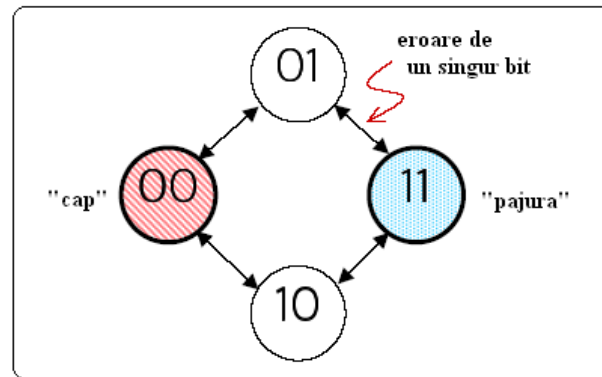
DETECTAREA ȘI CORECTAREA ERORILOR

Pentru a asigura detectarea unei erori de un singur bit, într-un cod de lungime oarecare, se adaugă un *bit de paritate*.



Distanța Hamming între două coduri valide devine egală cu 2.

Diagrama de codificare a rezultatului aruncării unei monede.

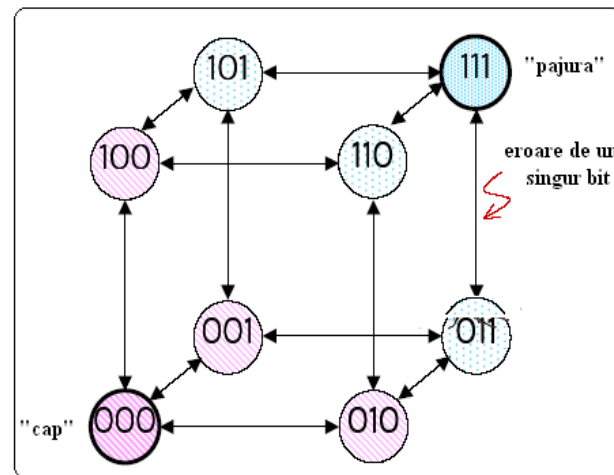


Dacă D este distanța Hamming minimă între două cuvinte cod, atunci se pot detecta până la $(D-1)$ erori la nivel de bit.

Mărind distanța Hamming între două cuvinte – cod valid la 3, se poate garanta faptul că seturile de cuvinte generate de erori la nivelul unui singur bit nu se suprapun. În cazul în care se detectează o eroare, aceasta se poate corecta, întrucât poziția bitului eronat poate fi localizată.



**Localizarea poziției
bitului eronat folosind o
diagrama-hipercub.**

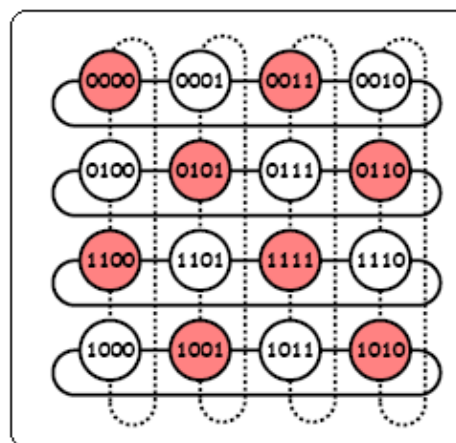
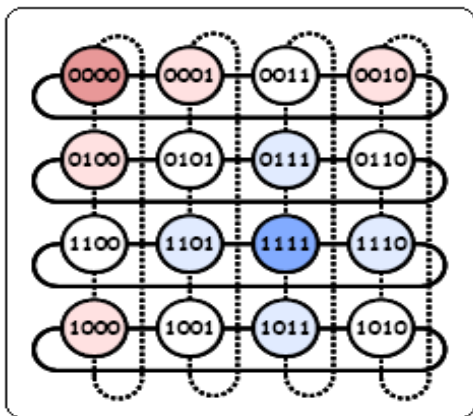


Dacă D este distanța Hamming minimă între două cuvinte-cod valide, se pot corecta $\left\lfloor \frac{D-1}{2} \right\rfloor$ biți eronați.

SCHEMA DE CODIFICARE CU 4 BIȚI A ERORILOR

Prin folosirea a 4 biți, pentru 3 biți de informație, se pot genera 8 coduri cu distanța Hamming egală cu 1, ceea ce permite detectarea unei erori la nivelul unui singur bit.

**Reprezentarea plană a
hipercubului generat
pentru codificare.**



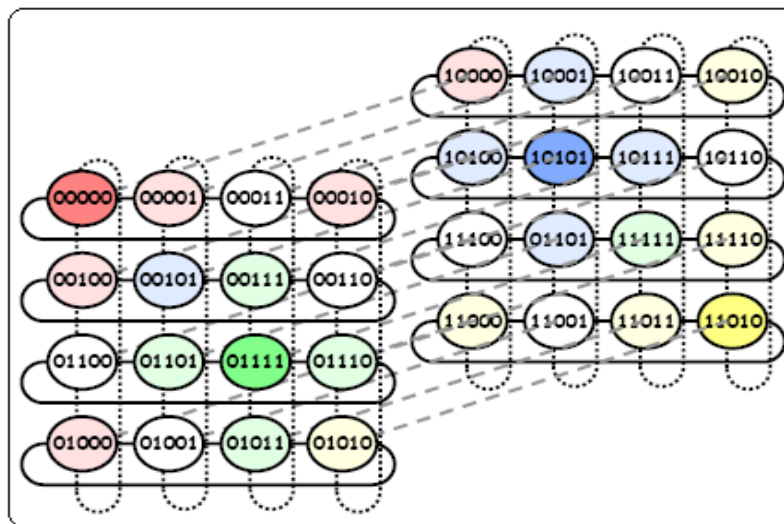
Există doar două coduri, care sunt separate printr-o distanță Hamming egală cu 4. Aceasta va permite corectarea erorilor de 1 bit, cât și detectarea erorilor de 2 biți.

SCHEMA DE CODIFICARE CU 5 BIȚI A ERORILOR

Asigură obținerea a mai mult de două coduri separate printr-o distanță Hamming egală cu 3. Hiper-cubul de mai jos conține 4 coduri de câte 5 biți {00000, 01111, 10101, 11010} separate printr-o distanță Hamming ≥ 3 .



Se poate corecta o eroare de 1 bit și se pot detecta unele dintre erorile de 2 biți, dar nu toate.



Codificarea permite soluționarea multor probleme:

- ❖ detectarea erorilor multi-bit: verificarea redundanței ciclice (CRC);
 - ❖ corectarea erorilor în rafală: coduri Reed-Solomon;
 - ❖ îmbunătățirea raportului semnal/zgomot
- 