heig-vd
Haute Ecole d'Ingénierie et de Gestion
du Canton de Vaud

RESDS
Reconfigurable & embedded
Digital Systems

CSE - Conception des systèmes embarqués – 8

Multimedia : Audio / Video

Michel Starkier

CSE

ENCODAGE ET TRANSPORT DE L'AUDIO NUMÉRIQUE

Bases d'acoustique

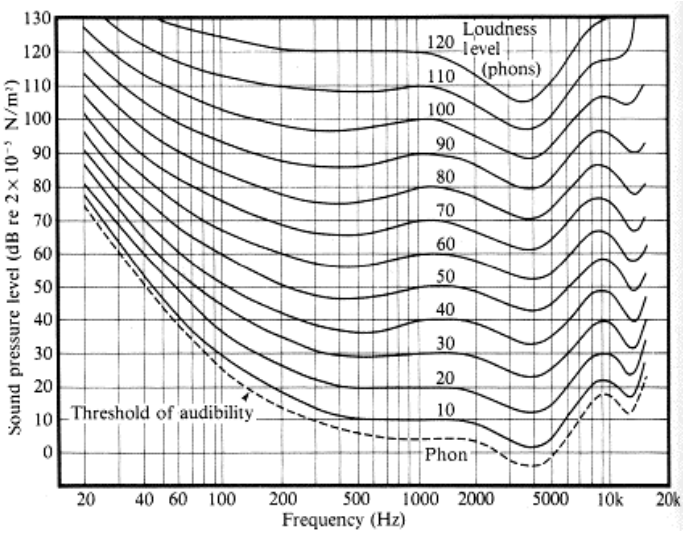
- Le son provient d'un objet qui entre en vibration mécanique et cause une variation de la pression atmosphérique dans l'air.
- Vitesse du son dans l'air : 340 m/s
- Niveau sonore mesuré en dB SPL

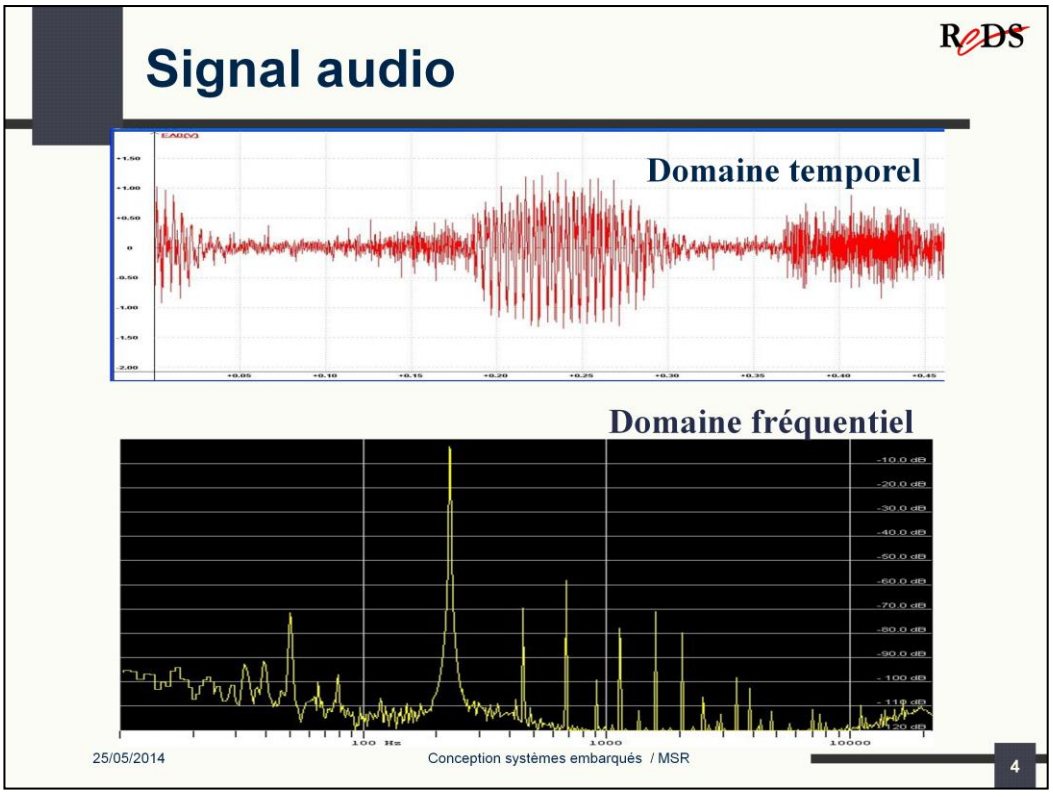
$S_{dB} = 20\ \log_{10}(P_{eff}/P_{ref})\ dB$ => rapport de pression, Pref est la pression de référence de 20 µPa (micro Pascal)

- Dynamique: 0dB (seuil d'audition) à 120dB (seuil de la douleur)
- Fréquence : 20Hz à 20KHz

25/05/2014 Conception systèmes embarqués / MSR 3

La perception du niveau sonore dépend de la fréquence. La courbe de Fletcher et Munson indique les lignes ou les niveaux sonores perçus sont identiques dans la plage des fréquences audibles.





Le signal audio peut être affiché dans le domaine temporel sous forme du niveau sonore en fonction du temps, ou dans le domaine fréquentiel sous forme du niveau des composantes fréquentielles pour une gamme de fréquence et une période de temps donné (transformée de Fourier du signal).

Audio numérique

- **Fréquence d'échantillonnage (non compressé) :**
  - Radio : 22/ 32 KHz / 16 bits
  - CD : 44.1 KHz / 16 bits
  - Audio pro : 48 KHz /16 ou 24 bits
  - Audio HQ : 96 KHz / 24 bits
  - Rapport signal /bruit > 90 dB
- **Théorème de Shannon**
  - fréquence d'échantillonnage d'un signal doit être égale ou supérieure au double de la fréquence maximale

Une sinusoïde de fréquence F_0 est échantillonnée à une fréquence F_s, la période d'échantillonnage est : $T = \frac{1}{F_s}$

$x[n] = sin(2\pi F_0 t_s)$ avec $t_s = nT$ (n est un entier)

Nous pouvons écrire :

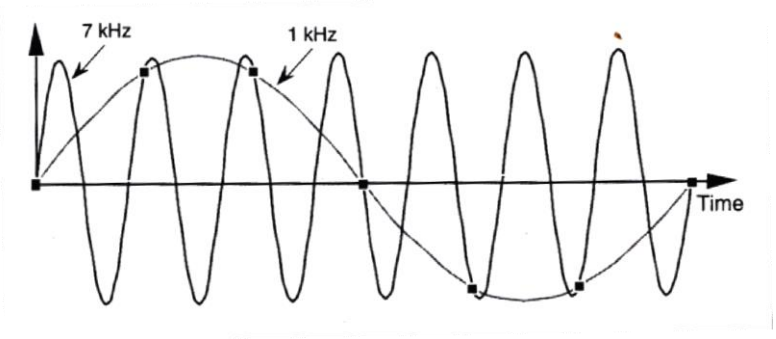
$x[n] = sin\left(2\pi F_0 nT + 2\pi kn\frac{T}{T}\right)$, avec k entier, car le deuxième terme prends les valeurs $2\pi, 4\pi, 6\pi$ ….

d'où :

$x[n] = sin(2\pi F_0 nT + 2\pi knF_s T) = sin(2\pi(F_0 + kF_s)nT)$

Avec un sampling rate de Fs samples/s, et k un entier (positif ou négatif), nous ne pouvons pas distinguer les valeurs échantillonnées d'une sinusoïde de fréquence F0 Hz de celles d'une sinusoïde de fréquence (F0 + Fs)

Fs = 6 kHz



Compression lossless (sans perte)

- Le plus simple : codage différentiel
- Plus compliqué : prédiction linéaire
- Run lenght (encodage des silences ...)
- Compression 50 %

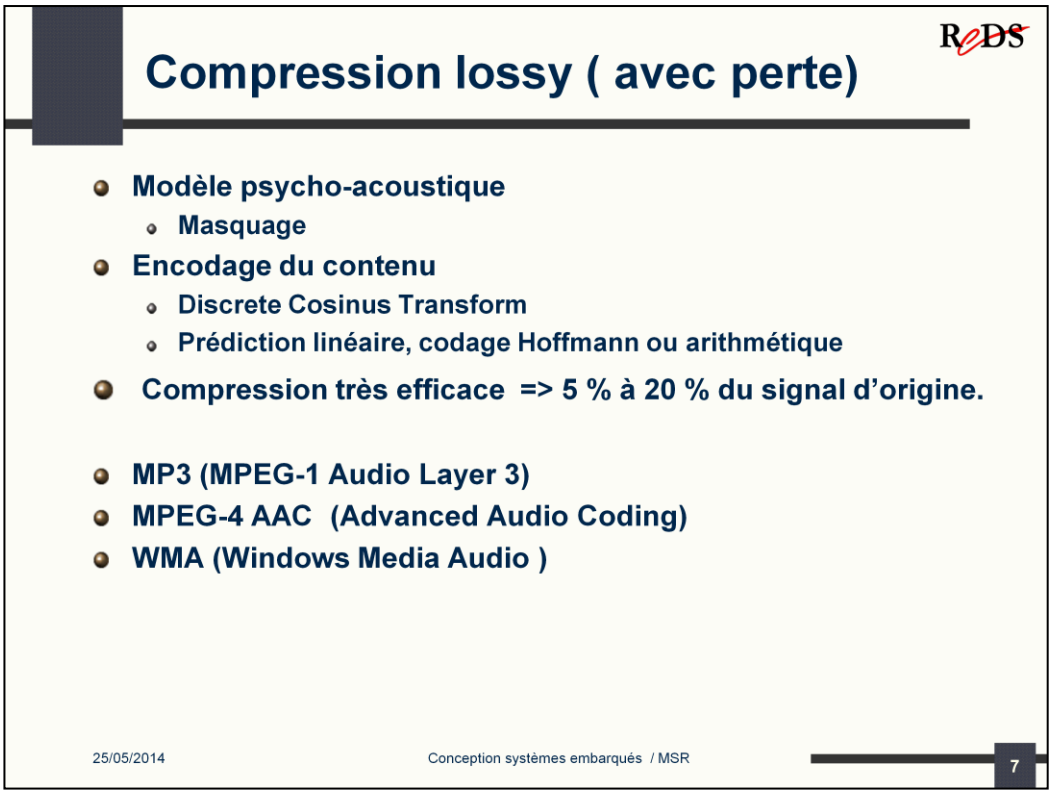
- Free lossless audio codec
- Shorten
- ALE (Apple)

25/05/2014 Conception systèmes embarqués / MSR 6

Le codage différentiel s'appuie sur le fait que les échantillons successifs sont fortement corrélés. Le codage différentiel, le plus simple consiste à encoder seulement la différence entre deux échantillons successifs. La prédiction linéaire, plus compliquée, consiste à prédire l'échantillon suivant et de ne coder que l'erreur entre l'échantillon prédit et l'échantillon réel suivant.

Liste non-exhaustive de codec lossless (Wikipedia)

Apple Lossless – ALAC (Apple Lossless Audio Codec)
apt-X Lossless
ATRAC Advanced Lossless
Direct Stream Transfer – DST
Dolby TrueHD
DTS-HD Master Audio
Free Lossless Audio Codec – FLAC
Meridian Lossless Packing – MLP
Monkey's Audio – Monkey's Audio APE
La famille des MPEG
OptimFROG
RealPlayer – RealAudio Lossless
Shorten – SHN
TTA – True Audio Lossless
WavPack – WavPack lossless
WMA Lossless – Windows Media Lossless

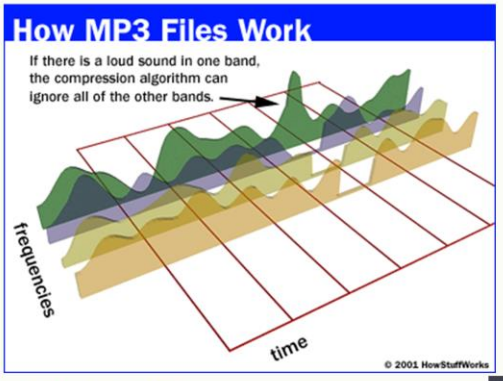


La compression avec perte s'appuie sur un modèle psycho-acoustique. Certains sons sont masqués par d'autres. Le cerveau ne les prends pas en compte (pour la majeure partie des êtres humains. On peut donc ne pas les transmettre (ces informations sont perdues => codage avec pertes.

La transformé en cosinus discrète est utilisée pour passer dans le domaine fréquentiel, ce qui permet de sélectionner les parties du signal à conserver. On effectue ensuite une compression «classique» des données.

MP3

- MP3 = MPEG-1/2 Audio Layer 3
- 128 Kbits <=> 1400 Kbits non-compressés
- Méthode :
  1. Transformation domaine fréquentiel
  2. Masquage
  3. Quantification
  4. Hoffman coding

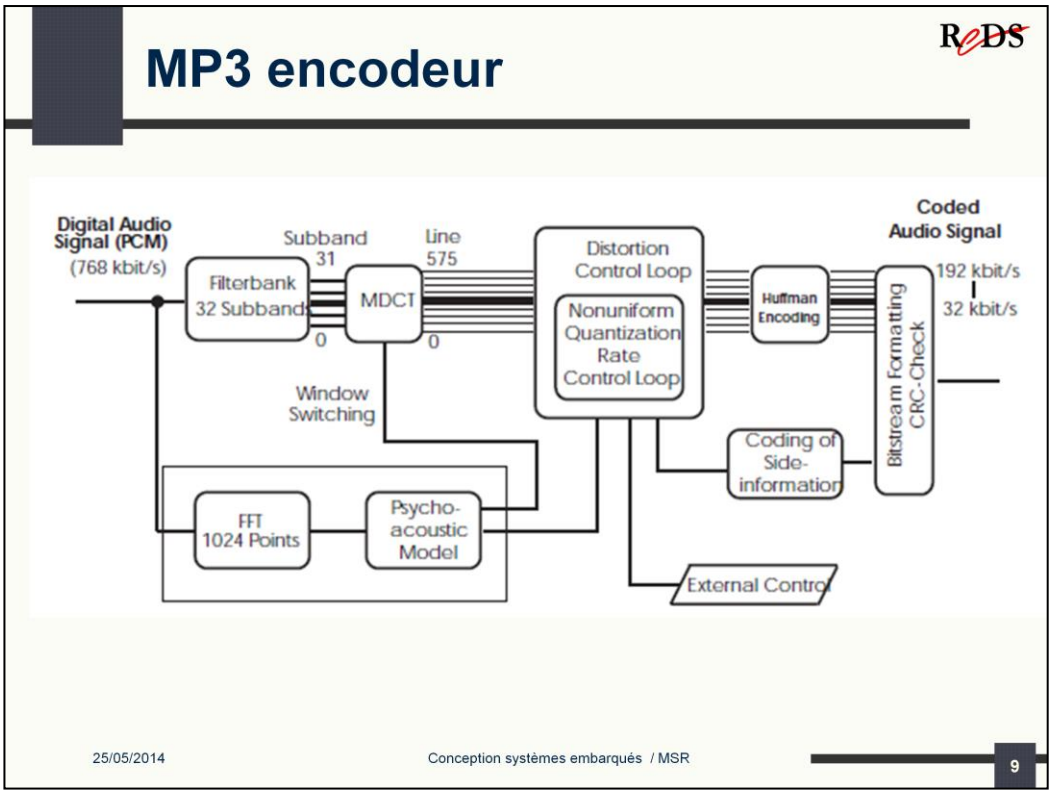


Un excellent tutorial sur le MP3 à l'université de Columbia

A Tutorial on MPEG/Audio Compression

Davis Pan
Motorola

http://www.ee.columbia.edu/~dpwe/e6820/papers/Pan95-mpega.pdf



Analyse du bloc diagramme d'un encodeur MP3 :

Le signal est séparé en 32 bandes, chacune est transformée dans le domaine fréquentiel par une transformée en cosinus discrète (DCT). Parallèlement le signal complet subit une transformée de Fourier, et est analysé dans son ensemble sur le plan fréquentiel. Le modèle psycho-acoustique est alors utilisé pour déterminer les parties du signal inutiles ou peu utiles pour une perception correcte du signal. Le signal provenant des 32 sous-bandes est quantifié en fonction des informations de masquage. C'est à cette phase que la quantité de données est réduite de façon importante. Les données quantifiées sont encodés de la façon la plus efficace possible par codage de Hoffman.

Interconnections audio par câble



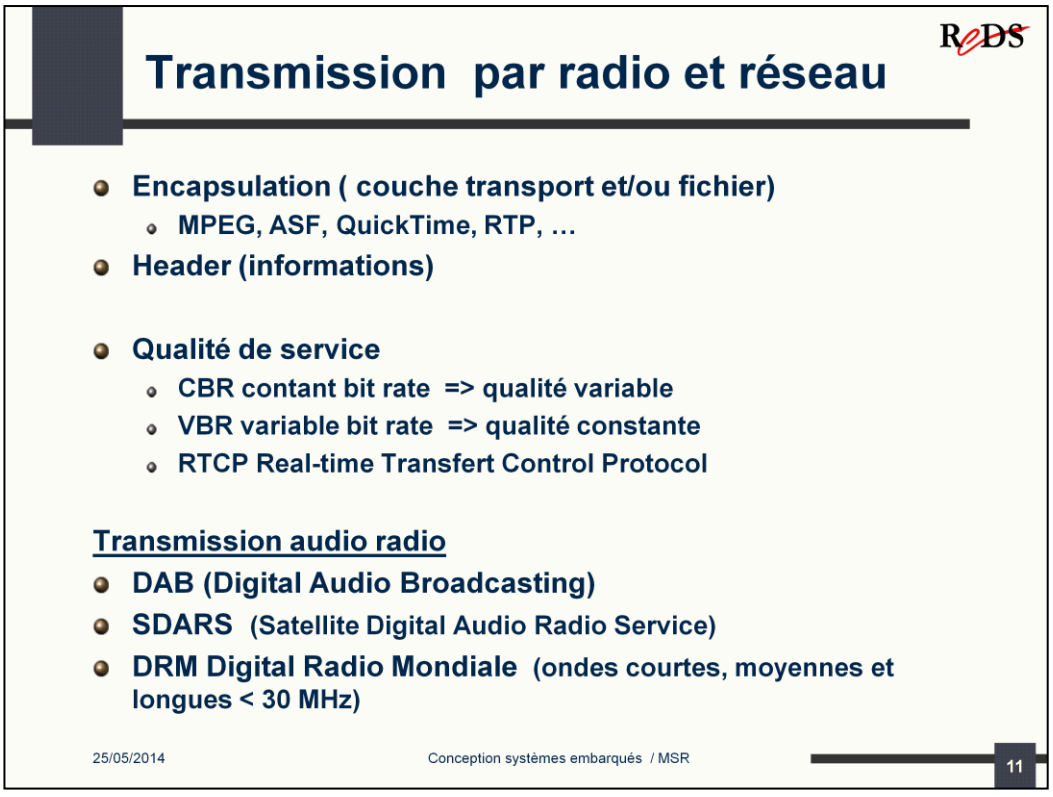
Analogique

- Symétrique – XLR
- Asymétrique - Cinch , Jack

Numérique

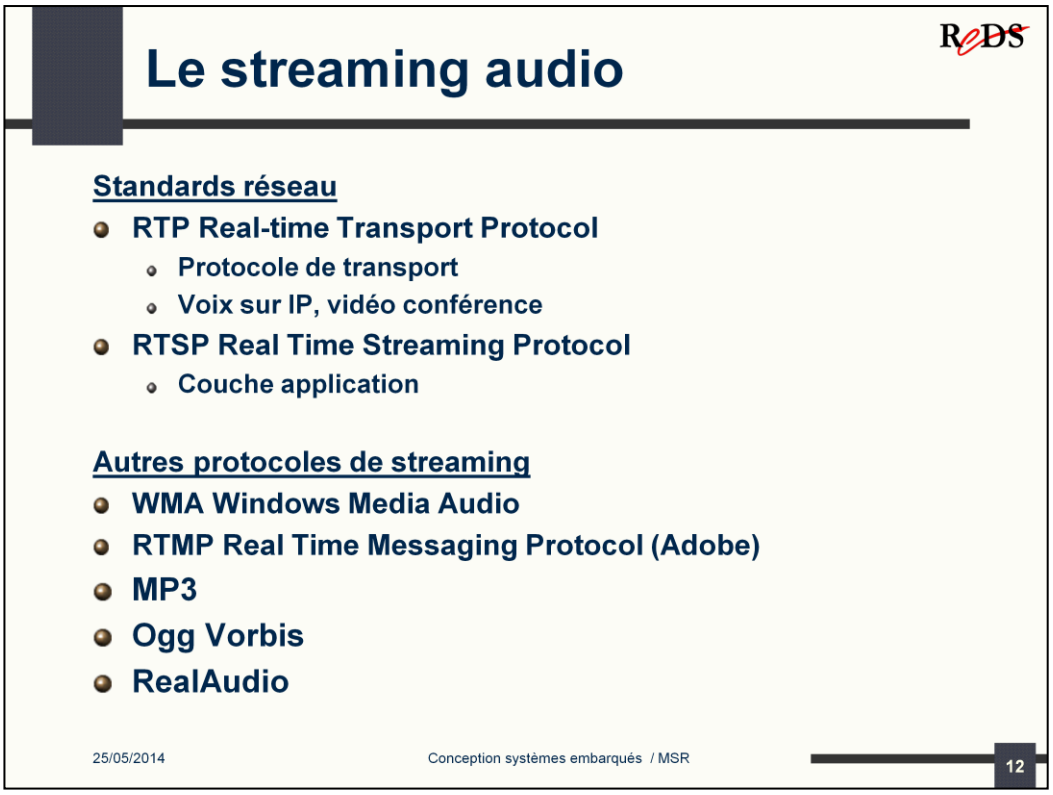

- AES-EBU 2 canaux,
  - Symétrique – XLR
  - Asymétrique - Cinch , Fibre

- S/PDIF cinch ou fibre
- HDMI – 8 canaux audio
- MADI /AES10 - 28, 56, ou 64 canaux - BNC ou fibre
- Ethernet, USB, FIreWire, … (fichiers ou streaming)



Les échantillons audio ne peuvent pas être transmis sans encapsulation. Cette encapsulation s'effectuer au niveau de la couche transport. On ajoute aux données audio des informations qui garantissent l'intégrité du transport (gestion des paquets, corrections des erreurs, time tags, ..).

La qualité de service est très importante en audio. Il s'agit de garantir le débit du canal de transport et les latences. La perte d'échantillons n'est pas admissible. Le VBR garantie une qualité constante, c'est-à-dire un taux de compression fixe, au prix d'un débit variable. Par contre le CBR garantit un débit fixe aux prix d''une qualité variable, c'est-à-dire d'un taux de compression variable.



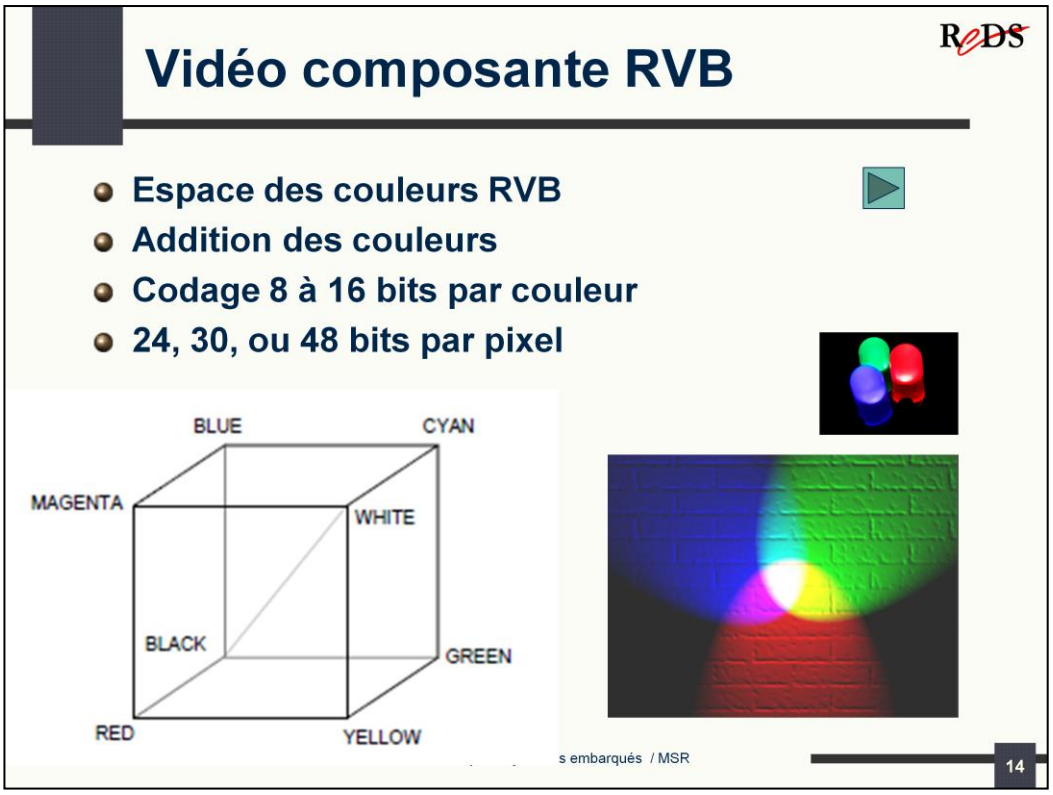
Streaming Media With Linux by Dave Phillips
http://linuxdevcenter.com/pub/a/linux/2001/03/23/streaming_media.html

SILICON LABORATORIES
USB AUDIO CLASS TUTORIAL
http://www.silabs.com/Support%20Documents/TechnicalDocs/AN295.pdf

OpenSL ES - The Standard for Embedded Audio Acceleration
http://www.khronos.org/opensles/

CSE

ENCODAGE ET TRANSPORT DE LA VIDÉO NUMÉRIQUE

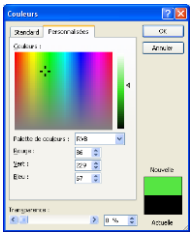


Les systèmes vidéo utilisent une synthèse additive des couleurs à partir de 3 couleurs primaires : rouge, vert et bleu.

- jaune = rouge + vert,
- violet (magenta) = bleu + rouge,
- bleu turquoise (cyan) = bleu + vert .

L'ensemble des couleurs que l'on peu obtenir par synthèse est représenté sur 3 axes par un cube, l'espace des couleurs.. Notez que rouge + vert + bleu = blanc. La quantification des couleurs se fait sur 8 à 16 bits, soit un mot de 24 à 48 bits par pixel (16 millions de couleurs en 24 bits).

Essayez de vous déplacer dans l'espace de couleur RVB de la fenêtre "Autres couleurs de.../Personnalisées" d'un programme Office. Observez les valeurs RVB correspondant à votre position.



Vidéo composante YUV/YCrCb

- YCrCb Luminance + Chrominance (Rouge Bleu)

$$[Y \quad Cb \quad Cr] = [R \quad G \quad B] \begin{bmatrix} 0.299 & -0.168935 & 0.499813 \\ 0.587 & -0.331665 & -0.418531 \\ 0.114 & 0.50059 & -0.081282 \end{bmatrix}$$

- YUV

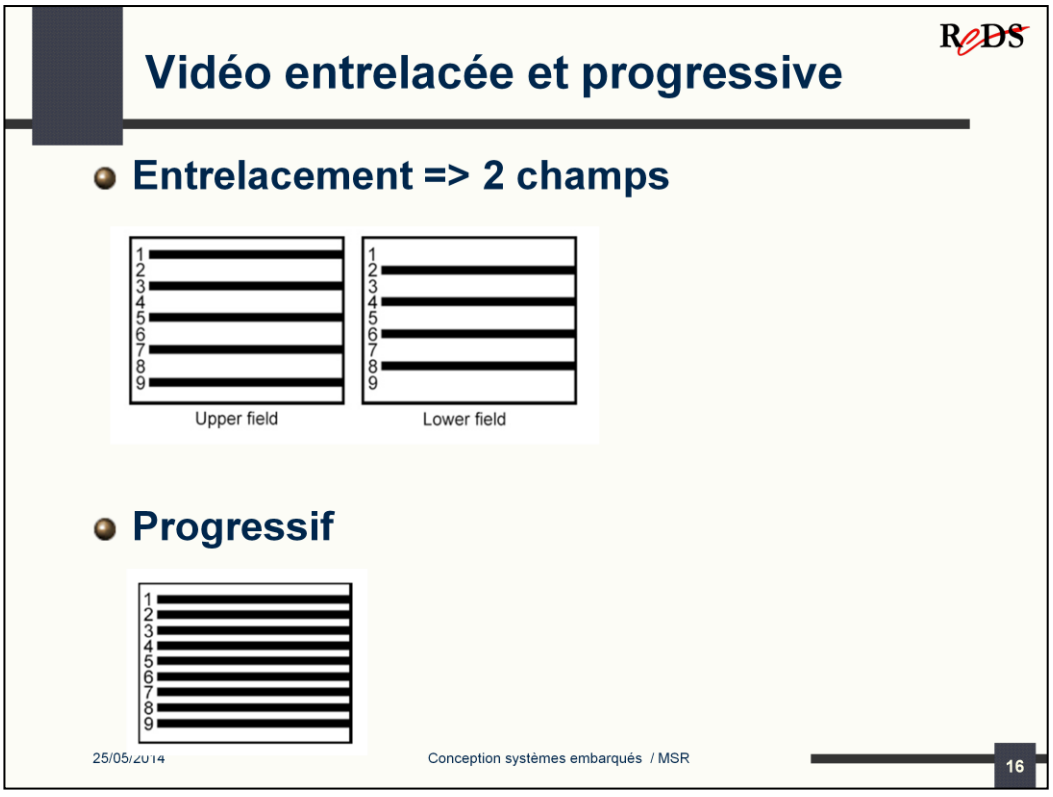
$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.14713 & -0.28886 & 0.436 \\ 0.615 & -0.51499 & -0.10001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

25/05/2014 Conception systèmes embarqués / MSR 15

L'espace de couleur peut être représenté en RVB (RGB en anglais), mais il existe une autre représentation appelé «vidéo composante». La couleur est définie par sa luminosité (luminance) et sa couleur (chrominance) composé du rouge et du bleu.

L'avantage de cette représentation est de pouvoir contrôler séparément la luminosité et la couleur. Ainsi l'ensemble des couleurs possibles pour un niveau de luminosité est représenté sur deux axes seulement (voir la boite de dialogue couleur de Windows). On passe de la représentation RVB à la représentation composante par une matrice.

Notez qu'il existe plusieurs représentations composante qui différent par la valeur des coefficients, par exemple YCrCb et YUV.



Les systèmes vidéo utilisaient un mécanisme appelé entrelacement pour limiter le nombre d'images par seconde. Les images diffusées par tube cathodique étaient balayées, c'est-à-dire diffusées point après point. L'entrelacement et le balayage tendent à disparaître avec l'utilisation des écrans LCD. Le progressif s'impose actuellement.

Entrelacement

Les lignes impaires de l'image sont d'abord balayées (trame impaire) , puis les lignes paires sont balayées(trame paire). Ce mécanisme évite le scintillement en donnant l'impression que le nombre d'image est doublé, par exemple 25 images/s correspond à 50 trames/s.

Progressif

L'image est diffusées en une seule fois. Le nombre d'image par seconde doit être au moins 50 i/s pour un confort visuel acceptable.

Formats vidéo

Format	Video Resolution
SQCIF	128 × 96
QCIF	176 × 144
SCIF	256 x 192
SIF(525)	352 x 240
CIF/SIF(625)	352 × 288
4SIF(525)	704 x 480
4CIF/4SIF(625)	704 × 576
16CIF	1408 × 1152
DCIF	528 × 384

Basse résolution (progressif)

Video Mode	Frame size in pixels (W×H)	Pixels per image[1]	Scanning type
576p[2]	704x576	405,504	Progressive
720p	1280×720	921,600	Progressive
1080i	1920×1080	2,073,600	Interlaced
1080p	1920×1080	2,073,600	Progressive
2K	2048x1536	3,145,728	Progressive
2160p	3840×2160	8,294,400	Progressive
4K	4096x3072	12,582,912	Progressive
2540p	4520x2540	11,480,800	Progressive
4320p	7680x4320	33,177,600	Progressive

Haute résolution

Wikipédia

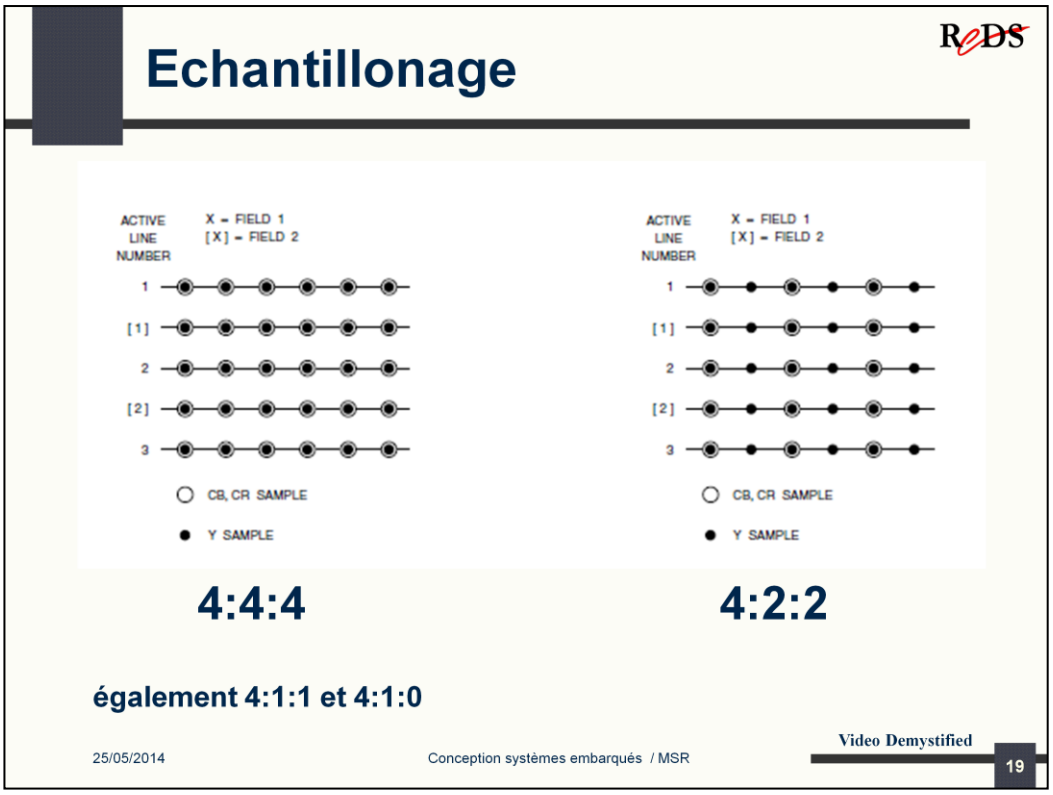
Les formats vidéo sont définis par le nombre de pixels horizontal et vertical, le système utilisé,- entrelacé (i) ou progressif (p) -, et le nombre d'images par seconde

Formats vidéo 4K

- **4K est le standard Ultra High Definition Television (UHDTV).**

Format	Resolution	Display aspect ratio	Pixels
UHD-1 Ultra high definition television	3840 × 2160	1.78:1 (16:9)	8,294,400
UHD Ultra wide television	5120 × 2160	2.37:1 (21:9)	11,059,200
WHXGA	5120 × 3200	1.60:1 (16:10)	16,384,000
DCI 4K (native resolution)	4096 × 2160	1.90:1 (256:135) ~17:9	8,847,360
DCI 4K (CinemaScope cropped)	4096 × 1716	2.39:1	7,020,544
DCI 4K (flat cropped)	3996 × 2160	1.85:1	8,631,360
UHD-2	7680 × 4320	16:9	33,177,600

Wikipédia



Chaque pixel de l'image est défini comme un échantillon du signal vidéo. Les valeurs d'échantillon sont transmises ligne par ligne de haut gauche vers bas droit de l'image.

4:4:4 : La fréquence d'échantillonnage de la luminance et celle de la chrominance sont identiques. Autant de pixels chrominance que de pixels luminance. Utilisé pour le RVB

4:2:2 : La fréquence d'échantillonnage de la luminance est deux fois plus élevée que la fréquence d'échantillonnage horizontal de la chrominance (un pixel couleur sur deux). Utilisé pour le YUV.

4:1:1 : La fréquence d'échantillonnage de la luminance est quatre fois plus élevée que la fréquence d'échantillonnage horizontal de la chrominance (un pixel couleur pour quatre pixels luminance)

4:2:0 : La fréquence d'échantillonnage de la luminance est deux fois plus élevée que la fréquence d'échantillonnage horizontal et que la fréquence d'échantillonnage vertical de la chrominance (un pixel couleur sur deux, une ligne couleur sur deux)

En 4:2:2 par exemple, le signal vidéo est encodé avec 4 mots (8 ou 19 bits) par pixel : Cb /Y0/Cr/Y1 => chrominance (bleu)/ luminance/chrominance/luminance

Compression vidéo

- **Standards courants : MJPEG, MPEG-1, MPEG-2, MPEG-4**

- **Méthodes de compression:**

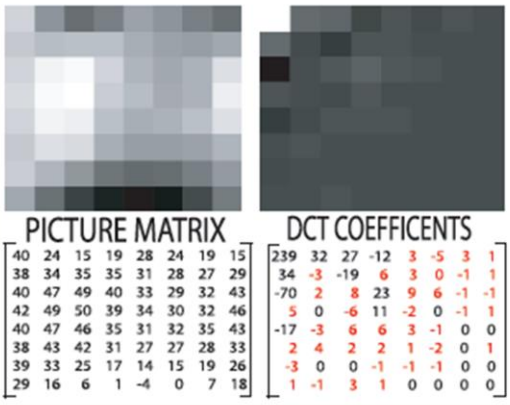
1. **Découpage de l'image en blocs (8x8 ou 16x16)**
2. **Transformation**
   - **DCT (TDC) Discrete Cosinus Transform**
   - **Ondelettes (moins utilisé)**
3. **Quantification => compression**
4. **Prédiction (MPEG)**
5. **Codage des données**
   - **Codage arithmétique**
   - **Codage de Huffman**

Méthodes de compression (Wikipedia)

Year	Standard	Publisher	Popular Implementations
1984	H.120	ITU-T	
1988	H.261	ITU-T	Videoconferencing, Videotelephony
1993	MPEG-1 Part 2	ISO, IEC	Video-CD
1995	H.262/MPEG-2 Part 2	ISO, IEC, ITU-T	DVD Video, Blu-ray, Digital Video Broadcasting, SVCD
1996	H.263	ITU-T	Videoconferencing, Videotelephony, Video on Mobile Phones (3GP)
1999	MPEG-4 Part 2	ISO, IEC	Video on Internet (DivX, Xvid ...)
2003	H.264/MPEG-4 AVC	Sony, Panasonic, Samsung, ISO, IEC, ITU-T	Blu-ray, HD DVD Digital Video Broadcasting, iPod Video, Apple TV,
2009	VC-2 (Dirac)	SMPTE	Video on Internet, HDTV broadcast, UHDTV
2013	H.265	ISO, IEC, ITU-T	

DCT, Discrete Cosine transform

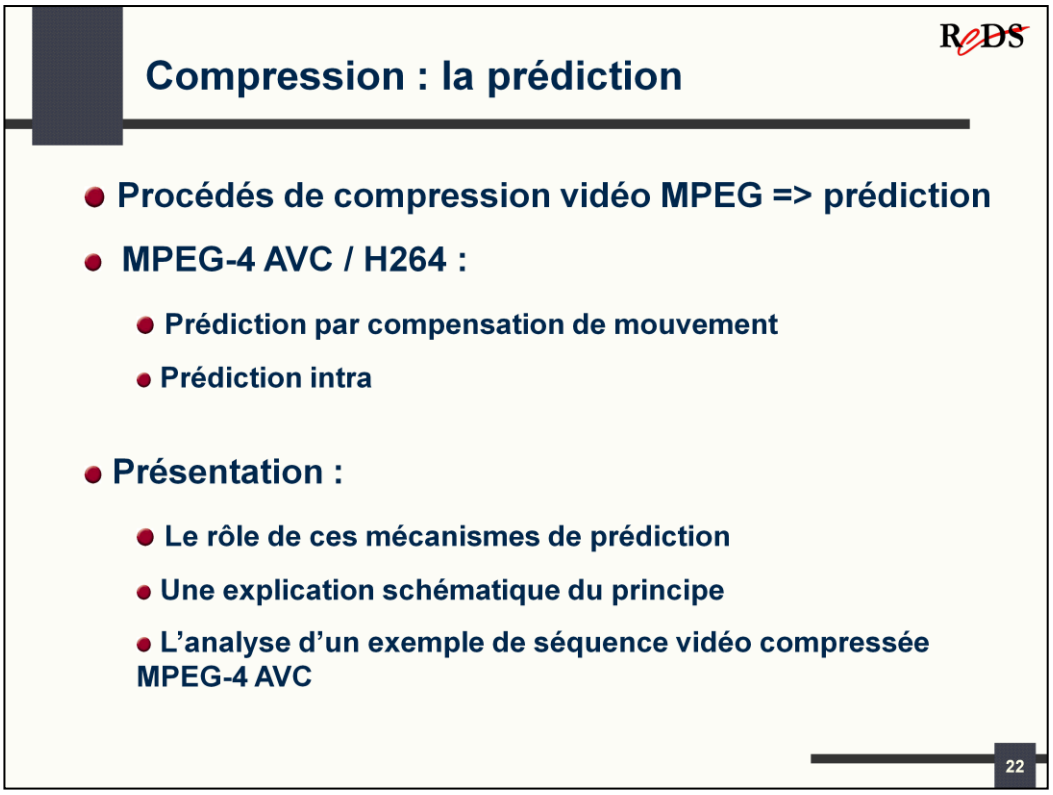
- DCT => composantes fréquentielles du bloc
- Conserver les basse fréquences (haut gauche)



La transformée en cosinus discrète (DCT ou Discrete Cosine Transform) est une transformation proche de la transformée de Fourier. La DCT s'applique à une matrice carrée. Le résultat est représenté dans une matrice de même dimension. Il s'agit de représenter un bloc de 8x8 pixels par une matrice de même dimension mais dans le domaine fréquentiel. Au décodage, le signal est ramené dans le domaine temporel par une DCT inverse.

Les basses fréquences se trouvant en haut à gauche de la matrice, et les hautes fréquences en bas à droite.

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right]$$

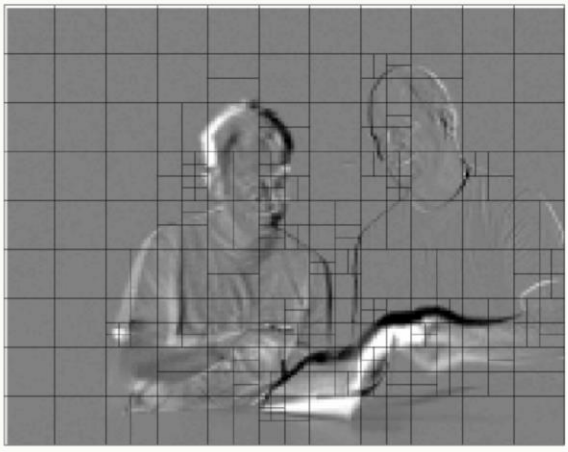


Les slides suivantes présentent un exemple de compression faisant appel au mécanisme de prédiction : MPEG-4 AVC.

AVC signifie Advanced Vidéo Coding. MPEG-A AVC est l'algorithme de compression le plus utilisé à l'heure actuelle.

Division des images en blocs

- Les images sont divisées en macroblocs : 16 x 16 pixels, divisés en blocs plus petits - 8x8, 4x8, 8x4, 4x4



Choix optimum ?

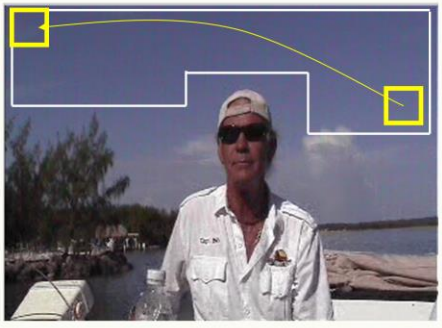
23

Les systèmes de compression moins sophistiqués que le MPEG-4 AVC utilisaient des blocs de taille fixe (en général 8x8). En fait, les blocs de petite taille sont plus adaptés à un niveau de détail important.

L'algorithme choisit la taille de bloc la plus adapté en fonction du niveau de détail. Plus le niveau de détail est important, plus le bloc contient de composantes haute fréquence.

Prédiction intra d'un bloc

- C'est calculer les valeurs des pixels de ce bloc, à partir des pixels provenant d'un autre bloc de cette même image



= > Réduire les redondances spatiales (les répétitions de valeurs semblables de pixels dans une même image)

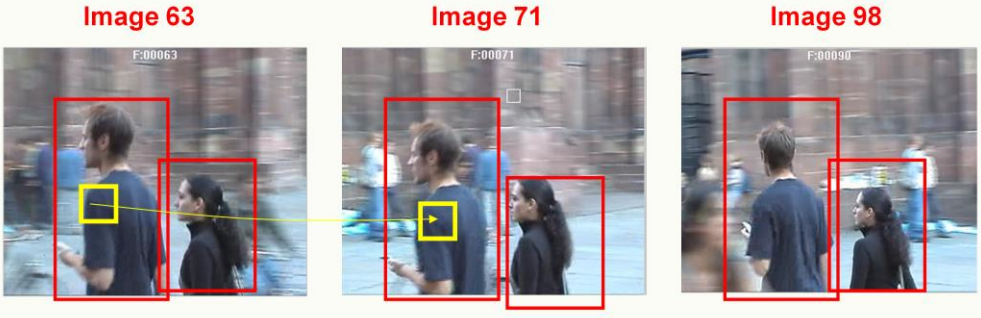
24

Afin de réduire la quantité d'information à transmettre, on déduit un bloc d'un autre bloc dans une image. Ceci est possible quand les blocs sont similaires : par exemple, un ciel bleu, un mur, ….

Dans l'exemple ci-dessus, les blocs composant le ciel sont quasi identiques. Il est inutile de les encoder tous. Il suffit de coder les (faibles) différences entre ces blocs. Cette technique s'appelle : **prédiction intra**.

Prédiction inter d'un bloc

- C'est calculer les valeurs des pixels de ce bloc, à partir des pixels provenant d'un bloc d'une autre image de la séquence vidéo

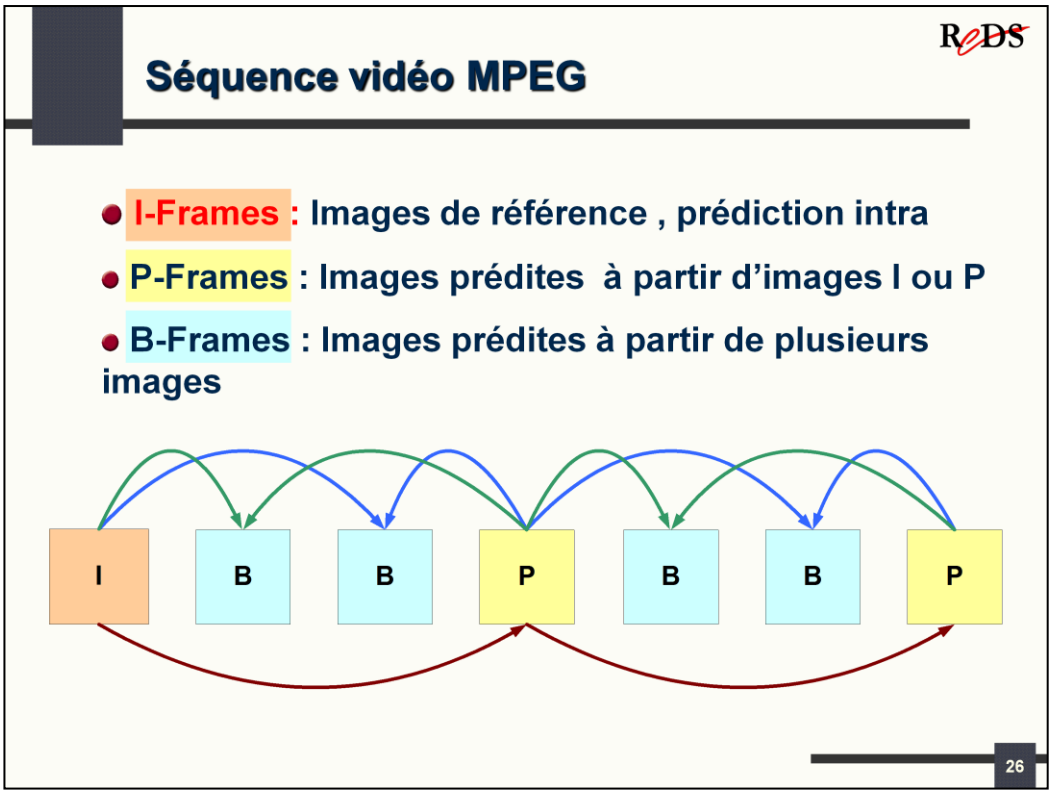


= > Réduire les redondances temporelles (les répétitions de valeurs semblables de pixels entre images successives)

25

Toujours pour réduire la quantité d'information à transmettre, on déduit un bloc d'un autre bloc dans une autre image. Ceci est possible quand une partie d'une image se retrouve dans les images suivantes. C'est le cas de nombreuses images : soit le fond reste fixe (personnages dans une pièces par exemple), soit le fond change mais certains éléments sont toujours présents (voiture en mouvement par exemple).

Dans l'exemple ci-dessus, des promeneurs sont filmés. Les blocs composant l'image de ces promeneurs sont quasi identiques d'une image à l'autre. Comme précédemment, il est inutile de les encoder tous. Il suffit de coder les (faibles) différences entre ces blocs et la position des blocs (qui peut avoir changé d'une image à l'autre). Cette technique s'appelle : **prédiction inter**.



Dans une séquence vidéo MPEG-4 AVC, les images ne sont pas toutes du même type:

Les **I-Frames** sont les images de référence. Le contenu d'une I-Frame ne dépend pas d'une autre image. Seule la prédiction intra est utilisée dans ces blocs. Ces images sont peu compressées (la compression intra n'est pas toujours possible) par quantification des blocs après DCT. La quantité de donnée est importante et la perte d'une I-Frame provoque un freeze de l'image (plusieurs images dépendant de l'I-Frame sont perdues).

Le **P-Frames** sont prédites à partir des I-Frames (ou parfois de P-Frames) précédentes. Ce sont des images de référence intermédiaires.

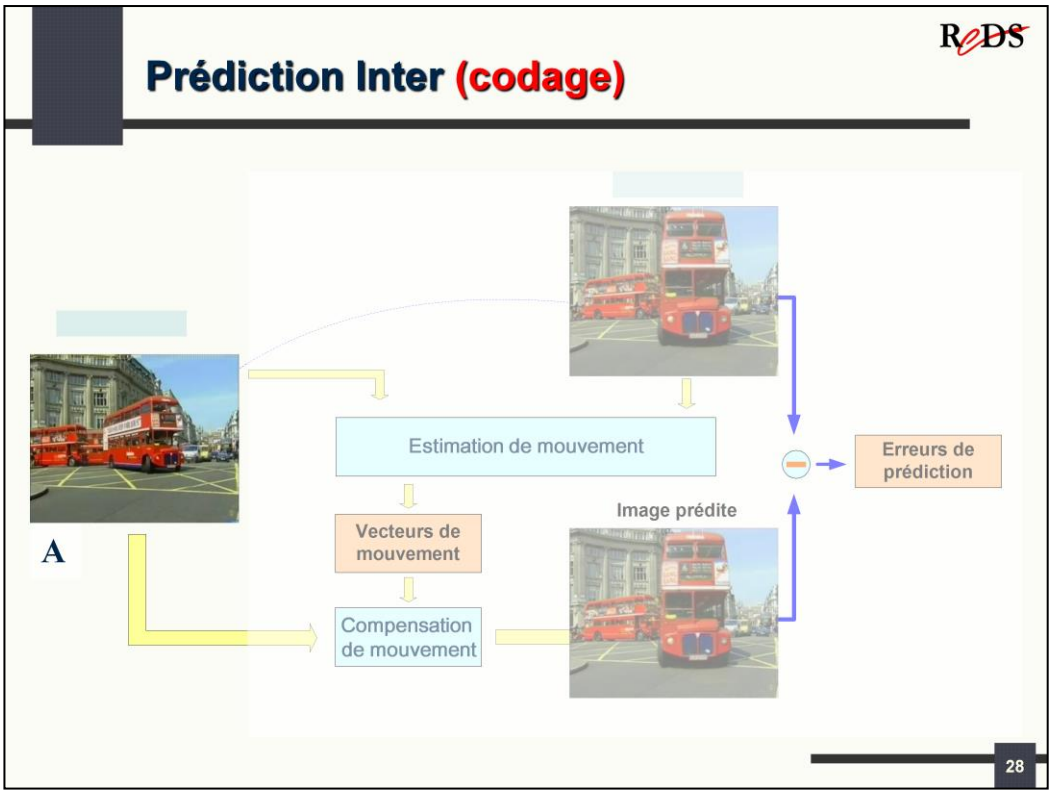
Les **B-Frames** peuvent être prédites à partir de I-Frames et de P-frames précédentes ou suivantes. Les B-Frames sont les images codées avec le minimum d'information.

Exemple

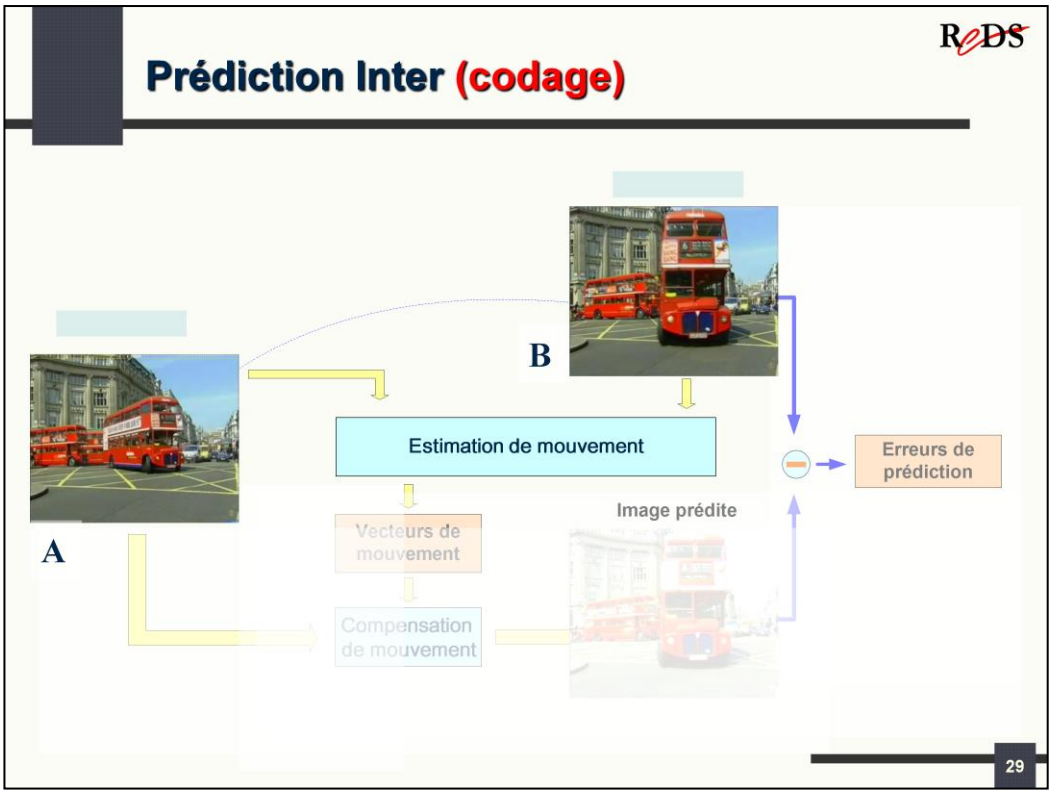
- **Séquence vidéo : CIF 352x288 pixels -15 i/s - 1,5 Mb/s**
  - **Compression MPEG-4 AVC / H264**
  - **Un plan fixe avec des surfaces homogènes**
  - **Véhicules en mouvement**

27

Dans l'exemple suivant, on utilise un film de test très particulier : Des bus rouges sont filmés en mouvement devant un fond fixe avec des grandes surfaces homogènes.



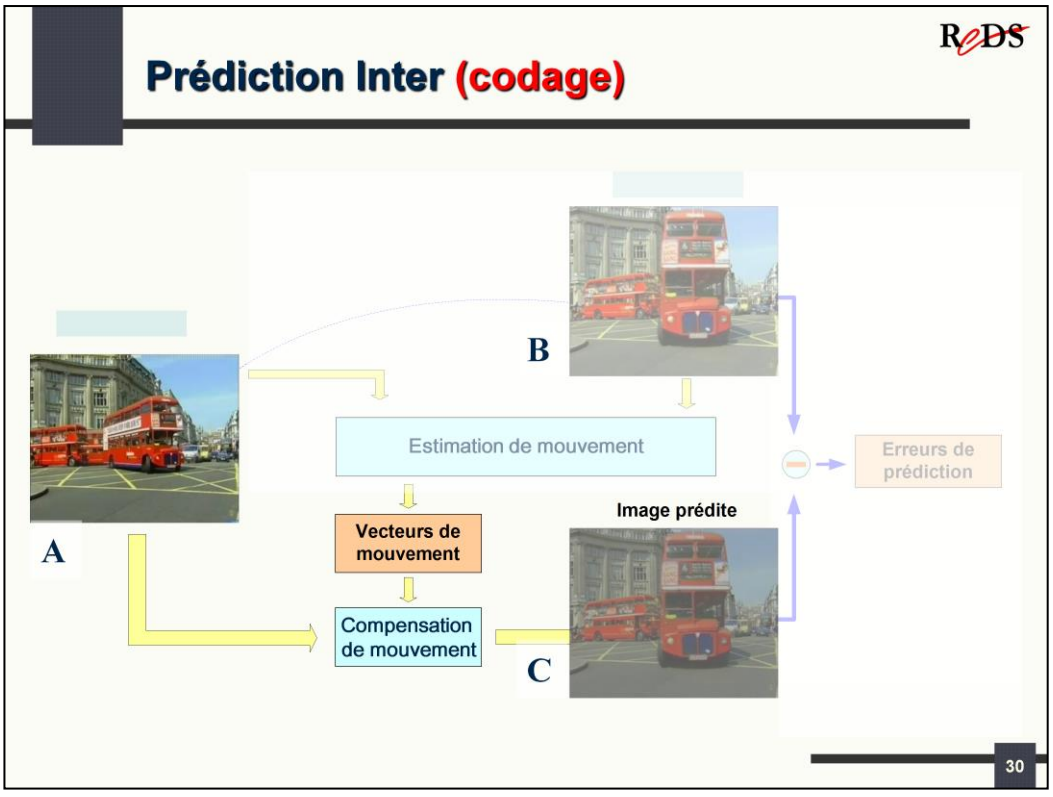
1- L'image de référence est l'image A.



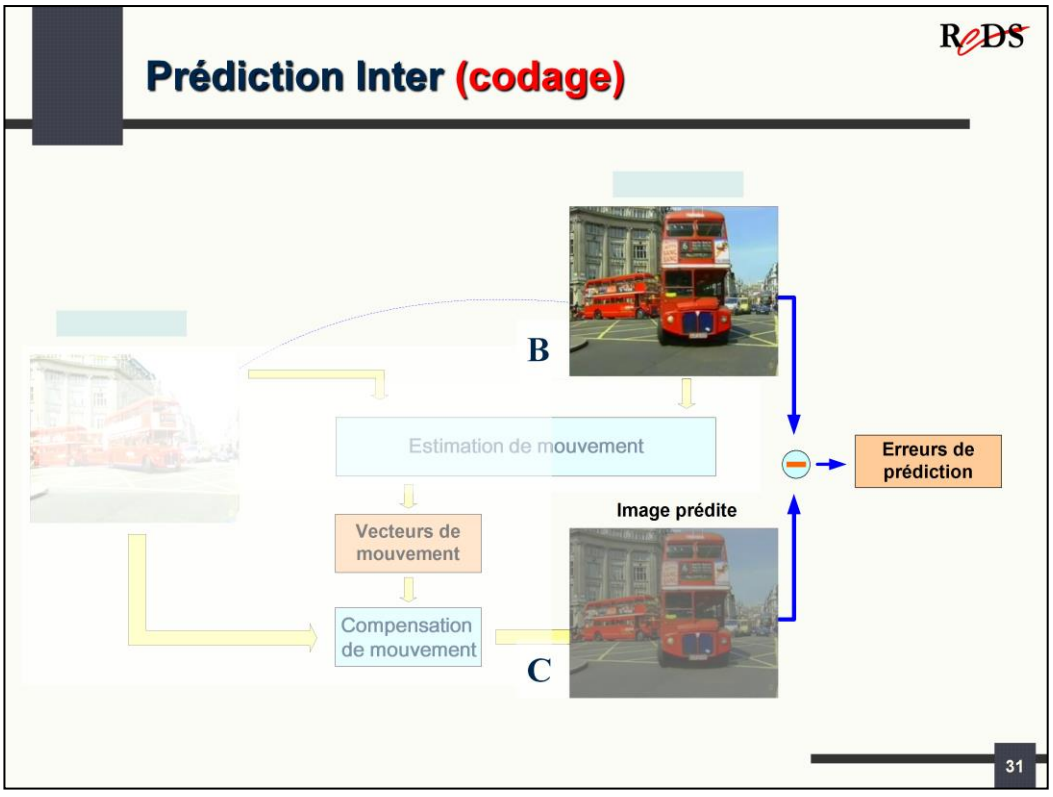
2- L'image qui doit être prédite à partir de l'image de référence A est l'image B située un peu plus loin dans la séquence vidéo.(image avec le bus de face).

L'algorithme de compression effectue une estimation de mouvement : A chaque bloc de l'image B, l'algorithme recherche un bloc plus ou moins identiques dans l'image B.

Il fait correspondre un (ou plusieurs blocs) de l'image B en calculant des vecteurs de mouvement. Les vecteurs de mouvement indiquent la direction et la distance de déplacement à appliquer à un bloc de l'image A pour obtenir le bloc correspondant de l'image B.

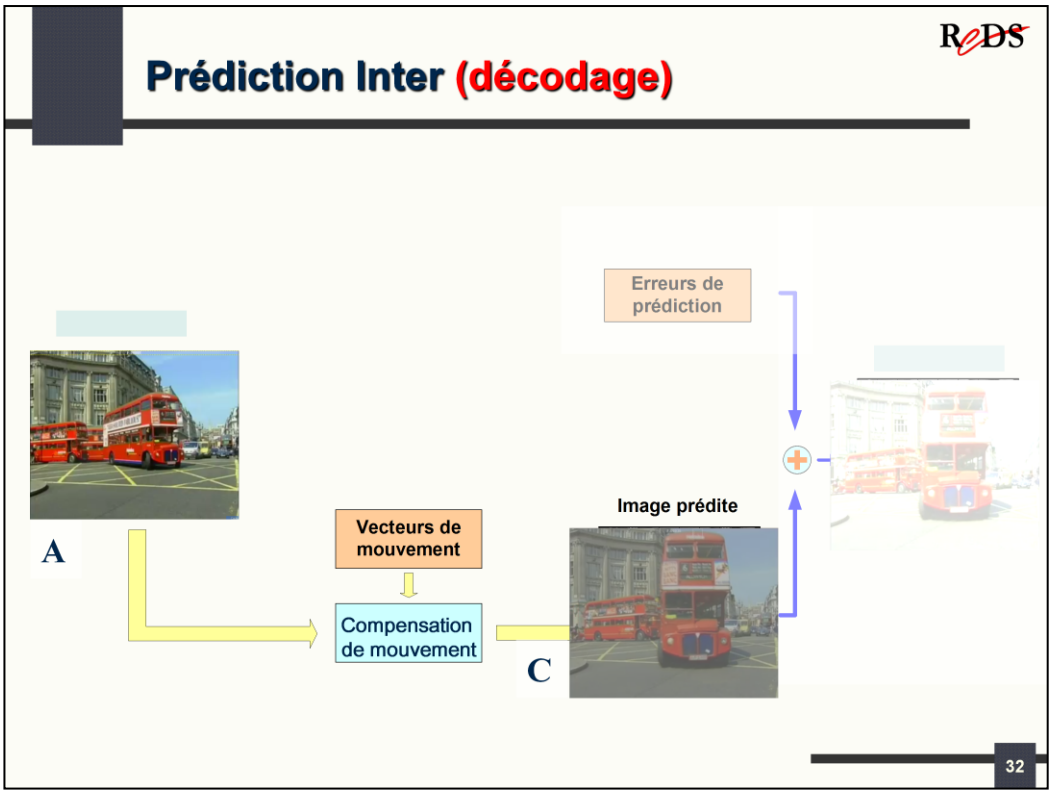


3- L'image C est calculée à partir de l'image A à laquelle on applique sur tous les blocs les vecteurs de mouvement. Evidemment, cette image calculée est une approximation de l'image réelle B.

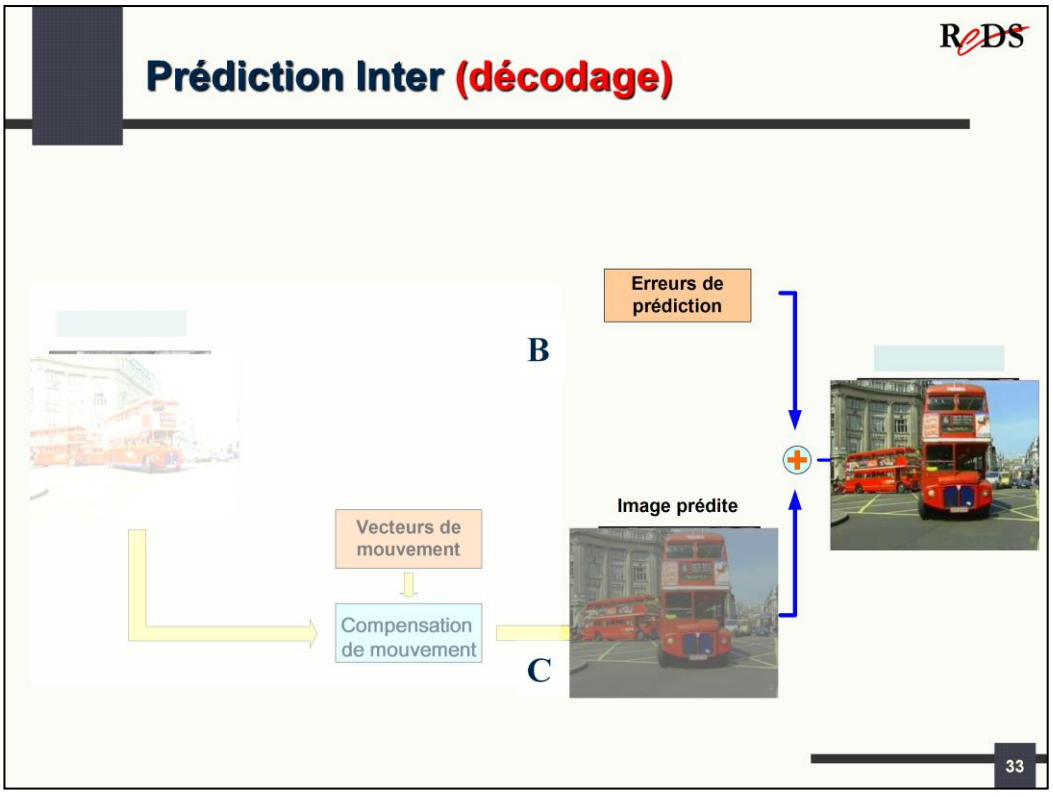


4- On compare (par soustraction) l'image C calculée avec l'image réelle B pour obtenir une table des erreurs de prédiction (B –C). L'image B est alors codée avec seulement la table des erreurs et les vecteurs de mouvement.

<u>On ne transmet pas l'image B, on transmet uniquement les informations nécessaires pour la calculer à partir de l'image A.</u>



5 - Au décodage, l'image C est calculée à partir de l'image A et des vecteurs de mouvement. Cette phase est appelée *compensation de mouvement*.

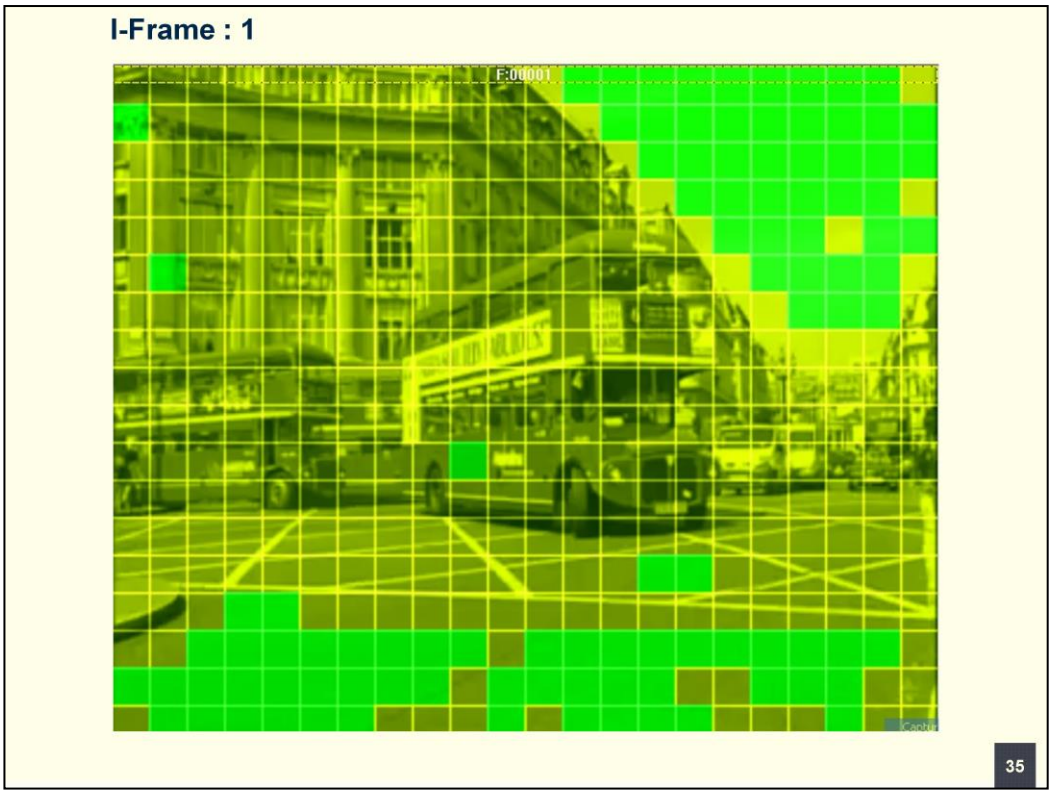


L'image B est ensuite calculée en ajoutant les erreurs de prédiction à l'image C.

I-Frame : 1

34

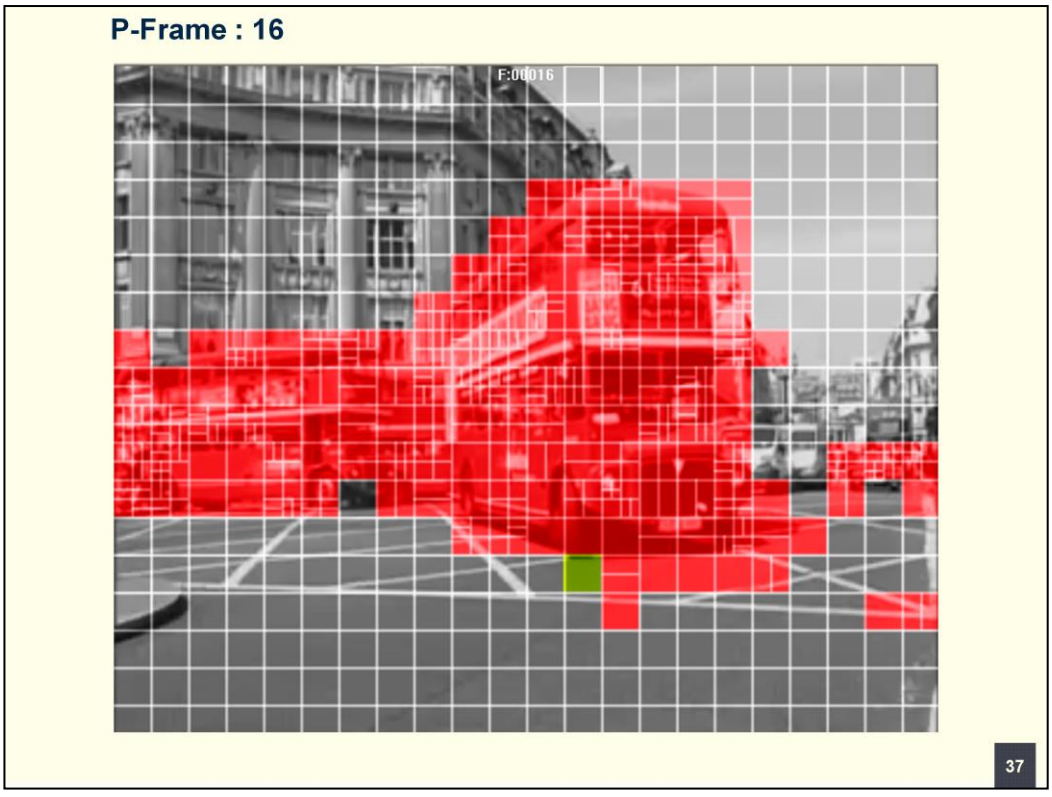
Cette image est l'I-Frame, image de référence transmise intégralement.



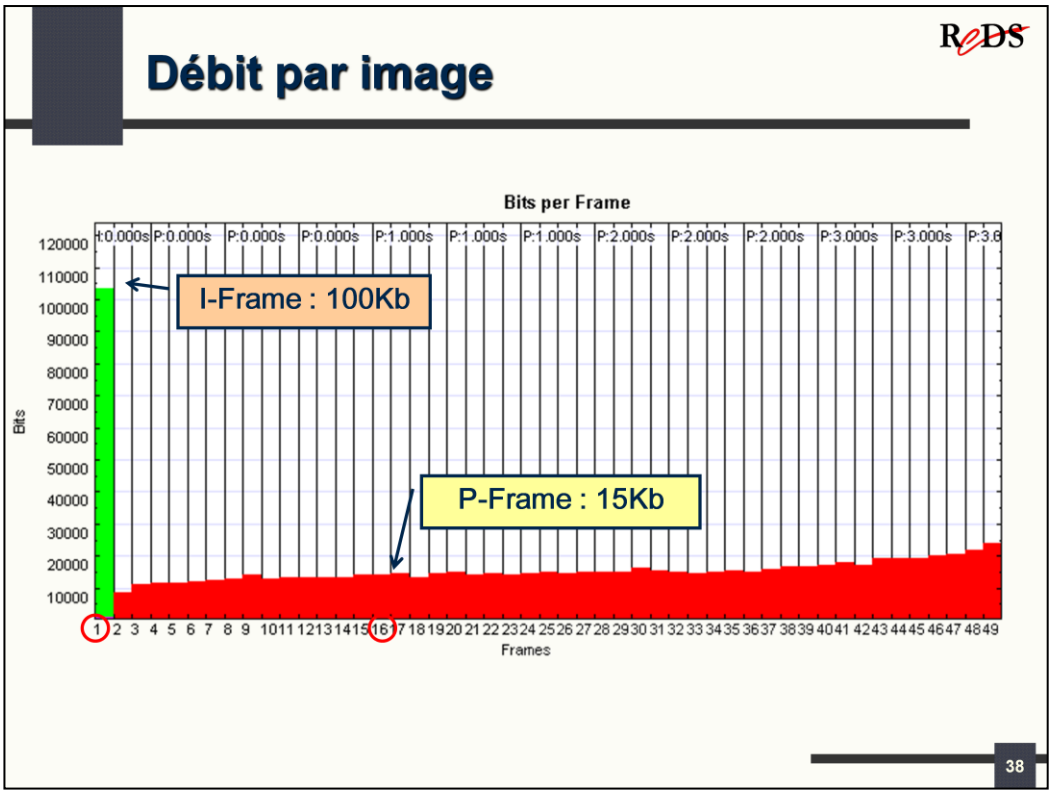
L’algorithme de compression repère les blocs quasi identiques, en vert uni sur l’image (prédiction inter).

P-Frame : 16

36



La P-Frame 16 est prédite à partir de l'I-Frame 1. Seuls les blocs rouges ont changé de place.

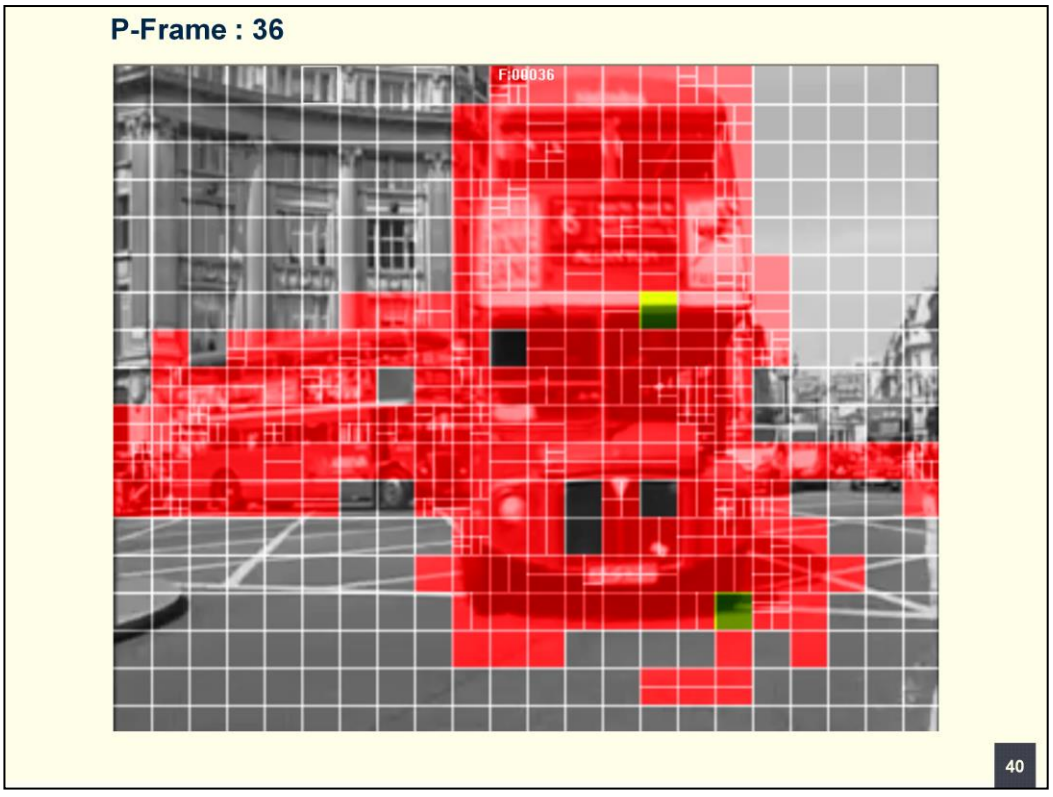


Noter la différence de débit (en Kbits) entre l'I-Frame 1 qui a été transmise intégralement et la P-Frame 16 qui a été transmise sous formes d'informations permettant de la calculer à partir de l'I-Frame.

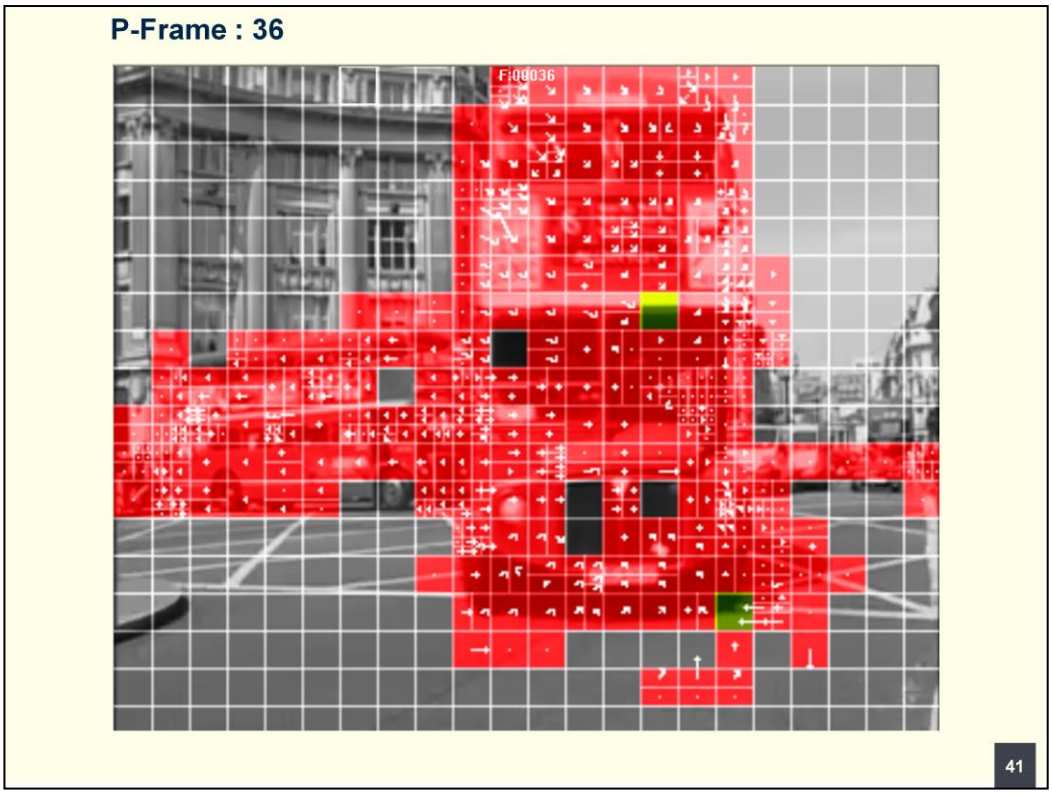
P-Frame : 36

39

Prenons maintenant l'exemple de la P-Frame 36 (image B des slides précédentes).



L’algorithme a identifié les blocs qui ont changés entre l’I-Frame 1 et cette image.

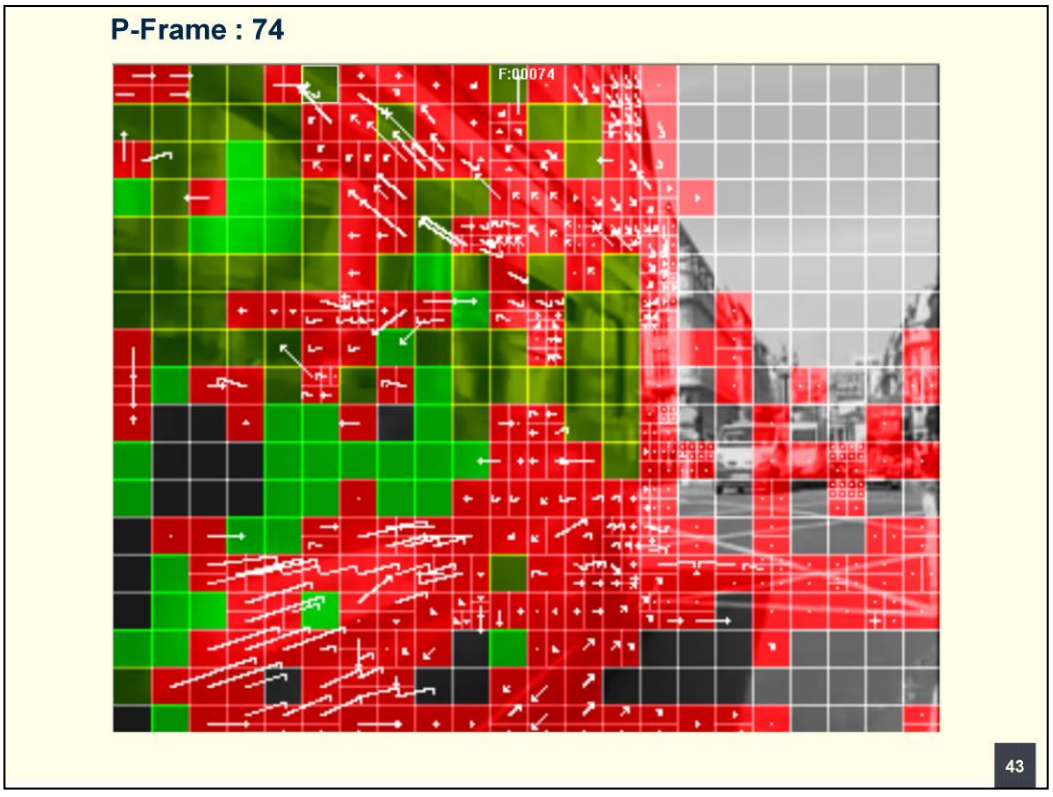


Les vecteurs de mouvement sont calculés.

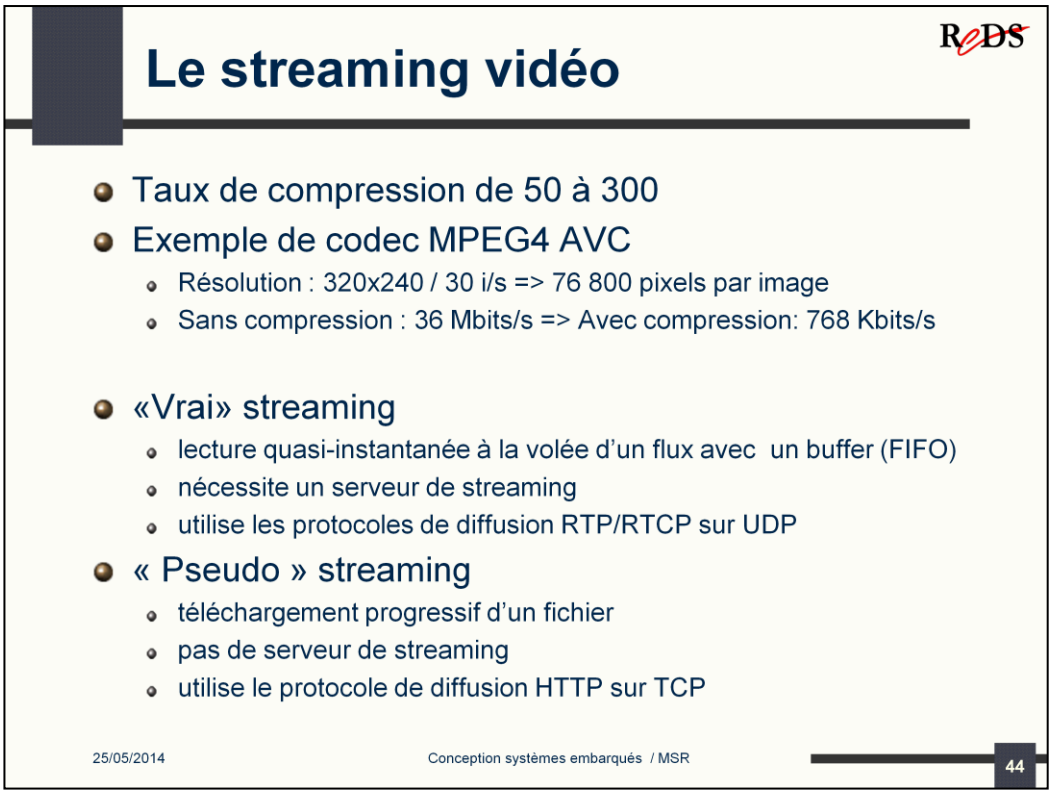
P-Frame : 74

42

Voici l’exemple d’une autre image avec prédiction inter et intra.



Les blocs rouges sont obtenus par prédiction intra et les blocs verts par prédiction inter. Les blocs gris n'ont pas changé.



Technologies de streaming :

- RealNetworks (RealMedia)
- Apple (QuickTimeMedia)
- Microsoft (Windows Media)
- Standard (ISMA)
- Cisco IP/TV
- VideoLAN (GNU Open Source)

Très bonne présentation :

Quelques mots sur la technologie de streaming

Nicolas.Meneceur

http://www.rap.prd.fr/pdf/technologie_streaming.pdf

Interconnections vidéo par câble


Analogique

- Composite (PAL) BNC / Cinch
- VGA 2048 × 1536 - 85p / DB 15pt

Numérique

- SDI : Vidéo professionelle / BNC
  - 143 Mbit/s (SD-SDI) à 2.970 Gbit/s, (HD-SDI /1080p)
- HDMI : High-Definition Multimedia Interface 19 pts
  - 4096 × 2160 - 60p audio + vidéo

- DVI : Digital Visual Interface 28pts
  - 3840 × 2400 – 33p vidéo

- Display Port : 20pts
  - 8192 × 4320 – 60p (V 1.3) audio + vidéo

Références

Video Demystified - A Handbook for the Digital Engineer

by Keith Jack

[1] P.Symes, *Video Compression*, McGraw-Hill, 1998

[2] R. Koenen, *Overview of the MPEG-4 Standard*, ISO/IEC JTC1/SC29/WG11 N2725, 1995

[3] ISO/IEC 11172, *MPEG-1, Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s*, 1993

[4] ISO/IEC 13818, *MPEG-2,* Generic coding of moving pictures and associated audio information, 1996

[4] Iain E G Richardson, *H.264 / MPEG-4 Part 10 White Paper, 2002*

[6] Tektronix, *MTS4EACompressed Video ES AnalyzerUser Tutorials*