目录

[一． 音视频基础 2](#_Toc12644)

[1. 基础概念介绍 2](#_Toc29759)

[1.1 GOP（Group of picture） 2](#_Toc22024)

[1.2 CABAC / CAVLC 3](#_Toc26581)

[1.3 Deblocking / FORCE\_IDR 3](#_Toc10851)

[1.4 frame / tff / bff 3](#_Toc25222)

[1.5 码流 / 码率 4](#_Toc17317)

[1.6 采样率 4](#_Toc10465)

[1.7 比特率 5](#_Toc12242)

[1.8 常见编码模式 5](#_Toc17859)

[1.9 帧速率 6](#_Toc22419)

[1.10 分辨率 6](#_Toc27497)

[1.11 高清视频 6](#_Toc18309)

[1.12 QP 7](#_Toc21001)

[1.13 PSNR 7](#_Toc25830)

[1.14 profile level 7](#_Toc10302)

[1.15 Reference 8](#_Toc28891)

[1.16 主码流/副码流 8](#_Toc3018)

[1.17 I P B帧 8](#_Toc5786)

[1.18 MPEG-4协议（ISO/IEC 14496标准） 9](#_Toc30080)

[1.19 PCR 12](#_Toc25666)

[2. 音视频入门 12](#_Toc6475)

[2.1 音视频通信流程 12](#_Toc19249)

[2.2视音频格式简介 13](#_Toc947)

[2.3 DTS&PTS 14](#_Toc27646)

[2.4 mpeg4视频中，I帧、p帧、B帧的判定 14](#_Toc18516)

[二．音视频格式与传输协议 15](#_Toc21404)

[1. MPEG2 PS&TS 16](#_Toc28949)

[1.1简介 16](#_Toc14887)

[1.2 PS流格式 19](#_Toc7212)

[1.3 TS流格式 47](#_Toc6495)

[1.4 PTS DTS PCR同步机制 61](#_Toc26415)

[2. H.264 66](#_Toc24561)

[2.1 简介 66](#_Toc30661)

[2.2 NALU H264网络传输的结构 80](#_Toc28657)

[2.3 H.264码流结构图 95](#_Toc13995)

[3. RTP 100](#_Toc18993)

[3.1 RTP介绍 100](#_Toc14387)

[3.2 RTP数据格式 102](#_Toc5473)

[3.3 RTP会话 107](#_Toc25957)

[3.4 RTP荷载H264码流 107](#_Toc26396)

[3.5 RTP荷载PS流 109](#_Toc21346)

[3.6 RTP的时间戳 同步 110](#_Toc9791)

[3.7 RTP打包方式 114](#_Toc25285)

[4. RTSP 114](#_Toc22865)

[4.1 RTSP概述 114](#_Toc20233)

[4.2 RTSP状态 116](#_Toc2234)

[4.3 资源预订协议RSVP协议 116](#_Toc22091)

[5. AAC音频格式 117](#_Toc23578)

[5.1 音频初步介绍 117](#_Toc9178)

[5.2 AAC 音频格式 118](#_Toc4870)

[5.3 AAC音频文件格式 120](#_Toc15467)

[5.4 AAC 与 MP3比较 127](#_Toc15598)

[6. AVI封装格式 128](#_Toc24828)

[6.1 AVI概述 128](#_Toc28929)

[6.2 AVI Layout(AVI文件层次划分) 131](#_Toc17820)

[7. MP4封装格式 137](#_Toc3643)

[7.1 MP4概述 137](#_Toc9146)

[7.2 MP4结构 138](#_Toc3349)

[7.3 MP4术语解释及知识补充 153](#_Toc11180)

[7.4 Mp4info工具各个参数介绍： 153](#_Toc30563)

[7.5MP4中如何seek 154](#_Toc14205)

# 音视频基础

## 基础概念介绍

### 1.1 GOP（Group of picture）

两个IDR帧之间的距离。增加关键帧个数可改善质量，但是同时增加带宽和网络负载。GOP=30，30FPS，即每个I帧之间相差1s。需要说明的是，通过提高GOP值来提高图像质量是有限度的，在遇到场景切换的情况时，H.264编码器会自动强制插入一个I帧，此时实际的GOP值被缩短了。另一方面，在一个GOP中，P、B帧是由I帧预测得到的，当I帧的图像质量比较差时，直到下一个GOP开始才有可能得以恢复，所以GOP值也不宜设置过大。过长的GOP还会影响Seek操作的响应速度，Seek操作需要先解码得到本GOP内的I帧及之前的N个预测帧才可以，GOP值越长，需要解码的预测帧就越多，seek响应的时间也越长。

### 1.2 CABAC / CAVLC

H.264/AVC标准中两种熵编码方法：CABAC叫自适应二进制算数编码，CAVLC叫前后自适应可变长度编码。其实就是帧编码和场编码。

CABAC：是一种无损编码方式，画质好，X264就会舍弃一些较小的DCT系数，码率降低，可以将码率再降低10-15%（特别是在高码率情况下），会降低编码和解码的速速。

CAVLC：占用更少的CPU资源，但会影响压缩性能。

帧：当采样视频信号时，如果是通过逐行扫描，那么得到的信号就是一帧图像，通常帧频为25帧每秒（PAL制）、30帧每秒（NTSC制）；

场：当采样视频信号时，如果是通过隔行扫描（奇、偶数行），那么一帧图像就被分成了两场，通常场频为50Hz（PAL制）、60Hz（NTSC制）；

帧频、场频的由来：最早由于抗干扰和滤波技术的限制，电视图像的场频通常与电网频率（交流电）相一致，于是根据各地交流电频率不同就有了欧洲和中国等PAL制的50Hz和北美等NTSC制的60Hz，但是现在并没有这样的限制了，帧频可以和场频一样，或者场频可以更高。

帧编码、场编码方式：逐行视频帧内邻近行空间相关性较强，因此当活动量非常小或者静止的图像比较适宜采用帧编码方式；而场内相邻行之间的时间相关性较强，对运动量较大的运动图像则适宜采用场编码方式。

### 1.3 Deblocking / FORCE\_IDR

Deblocking：开启会减少块效应

FORCE\_IDR：是否让每个I帧变成IDR帧，如果是IDR帧，支持随机访问。

### 1.4 frame / tff / bff

frame 将两场合并作为一帧进行编码

tff Enable interlaced mode (开启隔行编码并设置上半场在前)

bff Enable interlaced mode。

PAFF 和MBAFF：当对隔行扫描图像进行编码时，每帧包括两个场，由于两个场之间存在较大的扫描间隔，这样，对运动图像来说，帧中相邻两行之间的空间相关性相对于逐行扫描时就会减小，因此这时对两个场分别进行编码会更节省码流。

对帧来说，存在三种可选的编码方式：将两场合并作为一帧进行编码(frame 方式)或将两场分别编码(field 方式)或将两场合并起来作为一帧，但不同的是将帧中垂直相邻的两个宏块合并为宏块对进行编码；前两种称为PAFF 编码，对运动区域进行编码时field 方式有效，对非运区域编码时，由于相邻两行有较大的相关性，因而frame 方式会更有效。当图像同时存在运动区域和非运动区域时，在MB 层次上，对运动区域采取field 方式，对非运动区域采取frame 方式会更加有效，这种方式就称为MBAFF，预测的单位是宏块对。

### 1.5 码流 / 码率

码流(Data Rate)是指视频文件在单位时间内使用的数据流量，也叫码率或码流率，通俗就是取样率，画面质量控制中最重要的部分，一般是kb/s或者Mb/s。一般来说同样分辨率下，视频文件的码流越大，压缩比就越小，画面质量就越高，说明单位时间内取样率越大，精度就越高，处理出来的文件就越接近原始文件，图像质量越好，要求播放设备的解码能力也越高。

　　当然，码流越大，文件体积也越大，其计算公式是文件体积=时间X码率/8。例如，网络上常见的一部90分钟1Mbps码流的720P RMVB文件，其体积就=5400秒×1Mb/8=675MB。

　　通常来说，一个视频文件包括了画面及声音，而我们所说的一个视频文件码流率大小，一般是指视频文件中音频及视频信息码流率的总和。

以以国内最流行，大家最熟悉的RMVB视频文件为例，RMVB中的VB，指的是VBR，中文含义是可变比特率，表示RMVB采用的是动态编码的方式，把较高的采样率用于复杂的动态画面(歌舞、飞车、战争、动作等)，而把较低的采样率用于静态画面，达到画质与体积可兼得的效果。

### 1.6 采样率

采样率（也称为采样速度或者采样频率）定义了每秒从连续信号中提取并组成离散信号的采样个数，它用赫兹（Hz）来表示。采样率是指将模拟信号转换成数字信号时的采样频率，也就是单位时间内采样多少点。一个采样点数据有多少个比特。比特率是指每秒传送的比特(bit)数。单位为 bps(Bit Per Second)，比特率越高，传送的数据越大，音质越好.比特率 =采样率 x 采用位数 x声道数.

采样率类似于动态影像的帧数，比如电影的采样率是24赫兹，PAL制式的采样率是25赫兹，NTSC制式的采样率是30赫兹。当我们把采样到的一个个静止画面再以采样率同样的速度回放时，看到的就是连续的画面。同样的道理，把以44.1kHZ采样率记录的CD以同样的速率播放时，就能听到连续的声音。显然，这个采样率越高，听到的声音和看到的图像就越连贯。当然，人的听觉和视觉器官能分辨的采样率是有限的，基本上高于44.1kHZ采样的声音，绝大部分人已经觉察不到其中的分别了。

而声音的位数就相当于画面的颜色数，表示每个取样的数据量，当然数据量越大，回放的声音越准确，不过受人的器官的机能限制，16位的声音和24位的画面基本已经是普通人类的极限了，更高位数就只能靠仪器才能分辨出来了。比如电话就是3kHZ取样的7位声音，而CD是44.1kHZ取样的16位声音，所以CD就比电话更清楚。

当你理解了以上这两个概念，比特率就很容易理解了。以电话为例，每秒3000次取样，每个取样是7比特，那么电话的比特率是21000。 而CD是每秒 44100次取样，两个声道，每个取样是13位PCM编码，所以CD的比特率是44100\*2\*13=1146600，也就是说CD每秒的数据量大约是 144KB，而一张CD的容量是74分等于4440秒，就是639360KB＝640MB。

码率和取样率最根本的差别就是码率是针对源文件来讲的。

### 1.7 比特率

比特率是指每秒传送的比特(bit)数。单位为bps(Bit Per Second)，比特率越高，传送的数据越大。在视频领域,比特率常翻译为码率 !!!

比特率表示经过编码（压缩）后的音、视频数据每秒钟需要用多少个比特来表示，而比特就是二进制里面最小的单位，要么是0，要么是1。比特率就是比特率越高，音、视频的质量就越好，但编码后的文件就越大；如果比特率越少则情况刚好相反。

比特率是指将数字声音、视频由模拟格式转化成数字格式的采样率，采样率越高，还原后的音质、画质就越好。

### 1.8 常见编码模式

VBR（Variable Bitrate）动态比特率 也就是没有固定的比特率，压缩软件在压缩时根据音频数据即时确定使用什么比特率，这是以质量为前提兼顾文件大小的方式，推荐编码模式；

ABR（Average Bitrate）平均比特率 是VBR的一种插值参数。LAME针对CBR不佳的文件体积比和VBR生成文件大小不定的特点独创了这种编码模式。ABR在指定的文件大小内，以每50帧（30帧约1秒）为一段，低频和不敏感频率使用相对低的流量，高频和大动态表现时使用高流量，可以做为VBR和CBR的一种折衷选择。

CBR（Constant Bitrate），常数比特率 指文件从头到尾都是一种位速率。相对于VBR和ABR来讲，它压缩出来的文件体积很大，而且音质相对于VBR和ABR不会有明显的提高。

### 1.9 帧速率

帧速率也称为FPS(Frames PerSecond)的缩写——帧/秒。是指每秒钟刷新的图片的帧数，也可以理解为图形处理器每秒钟能够刷新几次。越高的帧速率可以得到更流畅、更逼真的动画。每秒钟帧数(FPS)越多，所显示的动作就会越流畅。

### 1.10 分辨率

帧大小每一帧就是一副图像。640\*480分辨率的视频，建议视频的码速率设置在700以上，音频采样率44100就行了。

一个音频编码率为128Kbps，视频编码率为800Kbps的文件，其总编码率为928Kbps，意思是经过编码后的数据每秒钟需要用928K比特来表示。

计算输出文件大小公式：（音频编码率（KBit为单位）/8 +视频编码率（KBit为单位）/8）×影片总长度（秒为单位）=文件大小（MB为单位）。不同的编码需要不同的系统资源，大概可以认为是H.264>VC-1>MPEG2。相对于MPEG2，VC-1的压缩比更高，但相对于H.264而言，编码解码的计算则要稍小一些，一般来说，VC-1多为 “.wmv”后缀，但这都不是绝对的，具体的编码格式还是要通过软件来查询。

总的来说，从压缩比上来看，H.264的压缩比率更高一些，也就是同样的视频，通过H.264编码算法压出来的视频容量要比VC-1的更小，但是VC-1 格式的视频在解码计算方面则更小一些，一般通过高性能的CPU就可以很流畅的观看高清视频。相信这也是目前NVIDIA Geforce 8系列显卡不能完全解码VC-1视频的主要原因。

### 1.11 高清视频

目前的高清视频编码格式主要有H.264、VC-1、MPEG-2、MPEG-4、DivX、XviD、WMA-HD以及X264。

事实上，现在网络上流传的高清视频主要以两类文件的方式存在：一类是经过MPEG-2标准压缩，以tp和ts为后缀的视频流文件;一类是经过WMV-HD(Windows Media Video HighDefinition)标准压缩过的wmv文件，还有少数文件后缀为avi或mpg，其性质与wmv是一样的。真正效果好的高清视频更多地以H.264与VC-1这两种主流的编码格式流传。一般来说，H.264格式以“.avi”、“.mkv”以及“.ts”封装比较常见。

### 1.12 QP

介于0~31之间，值越小，量化越精细，图像质量就越高，而产生的码流也越长。

### 1.13 PSNR

允许计算峰值信噪比(PSNR,Peak signal-to-noise ratio),编码结束后在屏幕上显示PSNR计算结果。开启与否与输出的视频质量无关，关闭后会带来微小的速度提升。

### 1.14 profile level

 MPEG-2按压缩比大小的不同分成五个档次（profile），每一个档次又按图像清晰度的不同分成四种图像格式，或称为级别（level）。五个档次四种级别共有20种组合，但实际应用中有些组合不太可能出现，较常用的是11种组合（下表中标识“√”的项）。这11种组合分别应用在不同的场合，如MP@ML。

分别是BP、EP、MP、HP：  
　　1、BP-Baseline Profile：基本画质。支持I/P 帧，只支持无交错（Progressive）和CAVLC；  
　　2、EP-Extended profile：进阶画质。支持I/P/B/SP/SI 帧，只支持无交错（Progressive）和CAVLC；  
　　3、MP-Main profile：主流画质。提供I/P/B 帧，支持无交错（Progressive）和交错（Interlaced），也支持CAVLC 和CABAC 的支持；  
　　4、HP-High profile：高级画质。在main Profile 的基础上增加了8x8内部预测、自定义量化、无损视频编码和更多的YUV 格式；

H.264规定了三种档次，每个档次支持一组特定的编码功能，并支持一类特定的应用。

1）基本档次：利用I片和P片支持帧内和帧间编码，支持利用基于上下文的自适应的变长编码进行的熵编码（CAVLC）。主要用于可视电话、会议电视、无线通信等实时视频通信；

2）主要档次：支持隔行视频，采用B片的帧间编码和采用加权预测的帧内编码；支持利用基于上下文的自适应的算术编码（CABAC）。主要用于数字广播电视与数字视频存储；

3）扩展档次：支持码流之间有效的切换（SP和SI片）、改进误码性能（数据分割），但不支持隔行视频和CABAC。主要用于网络的视频流，如视频点播。

### 1.15 Reference

Reference：指两个P帧之间的距离。

### 1.16 主码流/副码流

主码流/副码流：主码流位率高，图像质量高，便于本地存储；副码流位率低，图像质量低，便于网络传输。

### 1.17 I P B帧

I frame ：帧内编码帧 又称intra picture，I 帧通常是每个 GOP（MPEG 所使用的一种视频压缩技术）的第一个帧，经过适度地压缩，做为随机访问的参考点，可以当成图象。I帧可以看成是一个图像经过压缩后的产物。

P frame: 前向预测编码帧 又称predictive-frame，通过充分将低于图像序列中前面已编码帧的时间冗余信息来压缩传输数据量的编码图像，也叫预测帧；

B frame: 双向预测内插编码帧 又称bi-directional interpolated prediction frame，既考虑与源图像序列前面已编码帧，也顾及源图像序列后面已编码帧之间的时间冗余信息来压缩传输数据量的编码图像，也叫双向预测帧；

PTS：Presentation Time Stamp。PTS主要用于度量解码后的视频帧什么时候被显示出来

DTS：Decode Time Stamp。DTS主要是标识读入内存中的bit流在什么时候开始送入解码器中进行解码。

在没有B帧存在的情况下DTS的顺序和PTS的顺序应该是一样的。

I frame:自身可以通过视频解压算法解压成一张单独的完整的图片。

P frame：需要参考其前面的一个I frame 或者P frame来生成一张完整的图片。

B frame:则要参考其前一个I或者P帧及其后面的一个P帧来生成一张完整的图片。

如上图：I frame 的解码不依赖于任何的其它的帧.而p frame的解码则依赖于其前面的I frame或者P frame。B frame的解码则依赖于其前的最近的一个I frame或者P frame 及其后的最近的一个P frame。

### 1.18 MPEG-4协议（ISO/IEC 14496标准）

MPEG-4是个协议簇，包含了多个协议。



MPEG-4中只是定义了一部分AAC类型，AAC主要包括MPEG-2 AAC和MPEG-4 AAC, 详细如下图:



可以看出MPEG-4中只是定义了一部分类型的AAC.

MP4V在MPEG-4 的第2部分MPEG-4 Video中定义。

AVC在MPEG-4 的第10部分H264 or MPEG-4 AVC定义。

另: MP4A在MPEG-4的第三部分MPEG-4 Audio中定义。与MP4V相比，AVC具备更高的压缩效率。MP4V是原始视频数据。

MP4V-ES是RTP数据, 可通过RTP协议进行传输。

ISO/IEC 14496是MPEG专家组制定的MPEG-4标准，分为多个部分（仍在更新）。

参考：https://en.wikipedia.org/wiki/Category:ISO/IEC\_14496

第一部分（ISO/IEC 14496-1）：系统：描述视频和音频数据流的控制、同步以及混合方式（即混流Multiplexing，简写为MUX）。

第二部分（ISO/IEC 14496-2）：视频：定义一个对各种视觉信息（包括自然视频、静止纹理、计算机合成图形等等）的编解码器。（例如XviD编码就属于MPEG-4 Part 2）

第三部分（ISO/IEC 14496-3）：音频：定义一个对各种音频信号进行编码的编解码器的集合。包括高级音频编码（Advanced Audio Coding，缩写为AAC）的若干变形和其他一些音频/语音编码工具。

第四部分（ISO/IEC 14496-4）：一致性：定义对本标准其他的部分进行一致性测试的程序。

第五部分（ISO/IEC 14496-5）：参考软件：提供用于演示功能和说明本标准其他部分功能的软件。

第六部分（ISO/IEC 14496-6）：多媒体传输集成框架（DMIF for Delivery Multimedia Integration Framework）

第七部分（ISO/IEC 14496-7）：优化的参考软件：提供对实现进行优化的例子（这里的实现指的是第五部分）。

第八部分（ISO/IEC 14496-8）：在IP网络上传输：定义在IP网络上传输MPEG-4内容的方式。

第九部分（ISO/IEC 14496-9）：参考硬件：提供用于演示怎样在硬件上实现本标准其他部分功能的硬件设计方案。

第十部分（ISO/IEC 14496-10）：高级视频编码或称高级视频编码（Advanced Video Coding，缩写为AVC）：定义一个视频编解码器（codec）。AVC和XviD都属于MPEG-4编码，但由于AVC属于MPEG-4 Part 10，在技术特性上比属于MPEG-4 Part2的XviD要先进。另外，它和ITU-T H.264标准是一致的，故又称为H.264。

第十二部分（ISO/IEC 14496-12）：基于ISO的媒体文件格式：定义一个存储媒体内容的文件格式。

第十三部分（ISO/IEC 14496-13）：知识产权管理和保护（IPMP for Intellectual Property Management and Protection）拓展。

第十四部分（ISO/IEC 14496-14）：MPEG-4文件格式：定义基于第十二部分的用于存储MPEG-4内容的视频文件格式。

第十五部分（ISO/IEC 14496-15）：AVC文件格式：定义基于第十二部分的用于存储第十部分的视频内容的文件格式。

第十六部分（ISO/IEC 14496-16）：动画框架扩展（AFX : Animation Framework eXtension）。

第十七部分（ISO/IEC 14496-17）：同步文本字幕格式。

第十八部分（ISO/IEC 14496-18）：字体压缩和流式传输（针对开放字体格式Open Font Format）。

第十九部分（ISO/IEC 14496-19）：合成材质流（Synthesized Texture Stream）。

第二十部分（ISO/IEC 14496-20）：简单场景表示（LASeR for Lightweight Scene Representation。

第二十一部分（ISO/IEC 14496-21）：用于描绘（Rendering）的MPEG-J拓展。

第二十二部分（ISO/IEC 14496-22）：开放字体格式（Open Font Format）。

第二十三部分（ISO/IEC 14496-23）：符号化音乐表示（Symbolic Music Representation）。

第二十四部分（ISO/IEC 14496-24）：音频与系统交互作用（Audio and systems interaction）。

第二十五部分（ISO/IEC 14496-25）：3D图形压缩模型（3D Graphics Compression Model）。

第二十六部分（ISO/IEC 14496-26）：音频一致性检查：定义测试音频数据与ISO/IEC 14496-3是否一致的方法（Audio conformance）。

第二十七部分（ISO/IEC 14496-27）：3D图形一致性检查：定义测试3D图形数据与ISO/IEC 14496-11:2005, ISO/IEC 14496-16:2006, ISO/IEC 14496-21:2006,和ISO/IEC 14496-25:2009是否一致的方法（3D Graphics conformance）。

MP4是在“ISO/IEC 14496-14”标准文件中定义的一种多媒体容器格式，它是MPEG4 （ISO/IEC 14496）标准的一部分。是“ISO/IEC 14496-12(MPEG-4 Part 12 ISO base media file format)”标准中所定义的媒体格式的一种实现。

### 1.19 PCR

PCR 仅仅在实时播出流中才有意义（即前端编码器实时编码或视频服务器实时播出，后端机顶盒实时接收的情况），对DVD机这类从文件中取TS流的设备PCR毫无意义。

前端编码器工作原理：编码器中有一个27M的时钟，编码过程中不断读取硬件时钟放入TS中PCR字段，和PES的PTS，DTS字段。

后端机顶盒工作原理：机顶盒中也有一个27M的时钟，它根据PES中的PTS，DTS字段中指示的时间进行解码和显示。那么TS中的PCR还有什么用呢？首先如果前端编码器的时钟与后端机顶盒中时钟"绝对"同步，那么TS中的PCR就没有任何意义了。但是如果"绝对"变对"相对"以后会出先什么情况？它们之间的"微小"误差经过长时间的累积（1-2小时就足以）机顶盒中解码器会因为自己时钟"快"了而造成buffer中没有数据--停帧，会因为时钟"慢"了而造成buffer中数据溢出--丢帧。所对它需要TS中的PCR字段来不断修正自己时钟与编码器时钟间的"微小"误差。

TS流在经过复用器，加扰机这些中间设备时，PCR"必须"进行"非均匀"延迟修正（原因如上）。所以对于没有PCR修正功能的复用器和加扰机，建设大家一定不要买。

为什么说TS流中的PCR对DVD机毫无意义呢？DVD机是从光盘中读取TS流，读盘速度与TS流传送速率毫无关联，PCR还有指示时间的意义吗？但因为它不是"实时流"传送，所以不会出现buffer溢出的问题，也就不会产生停丢帧现象。

## 音视频入门

### 音视频通信流程



### 2.2视音频格式简介

视频、图像、音频的格式。这既包括MP4，FLV这样的封装格式，也包括H264，MPEG2这样的编码格式，还包括420 422 444这样的取样格式，甚至RGB，HSV，YUV这样的彩色空间也可以算作一种格式，将来在应用时还可能会遇到RTMP，HLS这样的流媒体传输协议或者其他的协议，都可以算作格式。

### 2.3 DTS&PTS

### 2.4 mpeg4视频中，I帧、p帧、B帧的判定

mpeg4的每一帧开头是固定的：00 00 01 b6，那么我们如何判断当前帧属于什么帧呢？在接下来的2bit，将会告诉我们答案。注意：是2bit，不是byte，下面是各类型帧与2bit的对应关系：

　　00: I Frame

　　01: P Frame

10: B Frame

为了更好地说明，我们举几个例子，以下是16进制显示的视频编码：

　　00 00 01 b6 10 34 78 97 09 87 06 57 87 …… I帧

　　00 00 01 b6 50 78 34 20 cc 66 b3 89 …… P帧

00 00 01 b6 96 88 99 06 54 34 78 90 98 …… B帧

下面我们来分析一下为什么他们分别是I、P、B帧

　　0x10 = 0001 0000

　　0x50 = 0101 0000

　　0x96 = 1001 0100

大家看红色的2bit，再对照开头说的帧与2bit的对应关系，是不是符合了呢？

下面给出一段c++代码供大家参考：

switch(buf[i] & (BYTE)0xc0)

{

case 0x00:

//I Frame

break;

case 0x40:

//P Frame

break;

case 0x80:

//B Frame

break;

default:

break;

}

## 图片入门

#### PNG

便携式网络图形（Portable Network Graphics）是一种无损压缩的位图图形格式。其设计目的是试图替代GIF和TIFF文件格式，同时增加一些GIF文件格式所不具备的特性。PNG使用从LZ77派生的无损数据压缩算法，一般应用于JAVA程序、网页或S60程序中，原因是它压缩比高，生成文件体积小。

PNG格式的图片具有如下特点：

体积小：网络通讯中因受带宽制约，保证图片清晰、逼真。

无损压缩：PNG文件采用LZ77算法获得高的压缩比，不损失数据。标记重复出现的数据，对图像的颜色没有影响，也不可能产生颜色的损失。

索引彩色模式：PNG-8格式与GIF图像类似，同样采用8位调色板将RGB彩色图像转换为索引彩色图像。图像中保存的不再是各个像素的彩色信息，而是从图像中挑选出来的具有代表性的颜色编号，每一编号对应一种颜色，图像的数据量也因此减少，这对彩色图像的传播非常有利。

更优化的网络传输显示： PNG图像在浏览器上采用流式浏览，即使经过交错处理的图像会在完全下载之前提供浏览者一个基本的图像内容，然后再逐渐清晰起来。它允许连续读出和写入图像数据，这个特性很适合于在通信过程中显示和生成图像。

支持透明效果：PNG可以为原图像定义256个透明层次，使得彩色图像的边缘能与任何背景平滑地融合，从而彻底地消除锯齿边缘。这种功能是GIF和JPEG没有的。

PNG同时还支持真彩和灰度级图像的Alpha通道透明度。

最高支持24位真彩色图像以及8位灰度图像。

支持Alpha通道的透明/半透明特性。

支持图像亮度的Gamma校准信息。

支持存储附加文本信息，以保留图像名称、作者、版权、创作时间、注释等信息。

渐近显示和流式读写，适合在网络传输中快速显示预览效果后再展示全貌。

使用CRC防止文件出错。

最新的PNG标准允许在一个文件内存储多幅图像。

# 二．音视频格式与传输协议

MPEG2系统任务包括：1. 规定以包传输数据的协议；2. 规定收发两端数据流同步的协议；3. 提供多个数据流的复用和解复用协议；3. 提供数据流加密的协议。以包形式存储和传送数据流是MPEG2系统之要点。

## 1. MPEG2 PS&TS

### 1.1简介

#### 1.1.1 PS TS ES介绍

1）ES ES--Elementary  Streams

原始流是直接从编码器出来的数据流，可以是编码过的视频数据流（H.264,MJPEG等），音频数据流（AAC），或其他编码数据流的统称。ES流经过PES打包器之后，被转换成PES包。每个ES都由若干个存取单元（AU）组成，每个视频AU或音频AU都是由头部和编码数据两部分组成，1个AU相当于编码的1幅视频图像或1个音频帧，也可以说，每个AU实际上是编码数据流的显示单元，即相当于解码的1幅视频图像或1个音频帧的取样。

1. PES PES--Packetized  Elementary Streams

ES形成的分组称为PES分组，是用来传递ES的一种数据结构。PES流是ES流经过PES打包器处理后形成的数据流，在这个过程中完成了将ES流分组、打包、加入包头信息等操作（对ES流的第一次打包）。PES流的基本单位是PES包。PES包由包头和payload组成。 3）

1. PS PS--Program Stream 节目流

PS流由PS包组成，而一个PS包又由若干个PES包组成（到这里，ES经过了两层的封装）。PS包的包头中包含了同步信息与时钟恢复信息。一个PS包最多可包含具有同一时钟基准的16个视频PES包和32个音频PES包。

1. TS TS--Transport Stream 传输流

由定长的TS包组成（188字节），而TS包是对PES包的一个重新封装（到这里，ES也经过了两层的封装）。PES包的包头信息依然存在于TS包中。TS流与PS流的区别在于TS流的包结构是固定长度的,而PS流的包结构是可变长度的。PS包由于长度是变化的,一旦丢失某一PS包的同步信息,接收机就会进入失步状态,从而导致严重的信息丢失事件。而TS码流由于采用了固定长度的包结构,当传输误码破坏了某一TS包的同步信息时,接收机可在固定的位置检测它后面包中的同步信息,从而恢复同步,避免了信息丢失。因此在信道环境较为恶劣、传输误码较高时一般采用TS码流,而在信环境较好、传输误码较低时一般采用PS码流。

PS&TS：两种视频或影片封装格式，常用于高清片。扩展名分别为VOB/EVO和TS等；其文件编码一般用MPEG2/VC-1/H.264。TS流与PS流的区别在于TS流的包结构是固定长度的，而PS流的包结构是可变长度的。

MPEG2(PS)就是常见dvd上那种vob格式的（PS是Program Stream程序流缩写），TS的全称则是Transport Stream（传输流），TS格式在网络上现在主要在一些高清视频，至pal和ntsc是两种不同的制式,二者在分辨率和帧数上有区别。MPEG2-PS主要应用于存储的具有固定时长的节目，如DVD电影，而MPEG-TS则主要应用于实时传送的节目，比如实时广播的电视节目。这两种格式的主要区别是什么呢？简单地打个比喻说，你将DVD上的VOB文件的前面一截cut掉（或者干脆就是数据损坏），那么就会导致整个文件无法解码了，而电视节目是你任何时候打开电视机都能解码（收看）的，所以，MPEG2-TS格式的特点就是要求从视频流的任一片段开始都是可以独立解码的。

PS包与TS包在结构上的这种差异，导致了它们对传输误码具有不同的抵抗能力，因而应用的环境也有所不同。TS码流由于采用了固定长度的包结构，当传输误码破坏了某一TS包的同步信息时，接收机可在固定的位置检测它后面包中的同步信息，从而恢复同步，避免了信息丢失。而PS包由于长度是变化的，一旦某一PS包的同步信息丢失，接收机无法确定下一包的同步位置，就会造成失步，导致严重的信息丢失。因此，在信道环境较为恶劣，传输误码较高时，一般采用TS码流；而在信道环境较好，传输误码较低时，一般采用PS码流。

ES是直接从编码器出来的数据流，可以是编码过的视频数据流，音频数据流，或其他编码数据流的统称。ES流经过PES打包器之后，被转换成PES包。PES包由包头和payload组成。PTS/DTS是打在PES包里面的，这两个parameters是解决视音频同步显示，防止解码器输入缓存上溢或下溢的关键。PTS表示显示单元出现在系统目标解码器(STD: system target

decoder)的时间，DTS表示将存取单元全部字节从STD的ES解码缓存器移走的时刻。每个I、P、B帧的包头都有一个PTS和DTS，但PTS与DTS对B帧都是一样的，无须标出B帧的DTS。对I帧和P帧，显示前一定要存储于视频解码器的重新排序缓存器中，经过延迟（重新排序）后再显示，一定要分别标明PTS和DTS。

ES首先需打包成PES流包，然后PES根据需要打包成PS或TS包进行存储或传输。其每路ES只包含一路信源的编码数据流，所以每路PES也只包含相对应信源的数据流。

对PS流而言，每个PES包头含有PTS和DTS，流识别码，用于区别不同性质ES。然后通过PS复用器将PES包复用成PS包。实际上是将PES包分解为更细小的PS包。在解码的时候，解复用器将PS分解成一个个PES包，拆包器然后将PES包拆成视频和音频的ES，最后输入至各自解码器进行解码。一个问题是：各个ES在解码时，如何保证视音频的同步呢？除了PTS和DTS的配合工作外，还有一个重要的参数是SCR(system clockreference)。在编码的时候，PTS，DTS和SCR都是由STC(system timeclock)生成的，在解码时，STC会再生，并通过锁相环路（PLL－phase lockloop），用本地SCR相位与输入的瞬时SCR相位锁相比较，以确定解码过程是否同步，若不同步，则用这个瞬时SCR调整27MHz的本地时钟频率。最后，PTS，DTS和SCR一起配合，解决视音频同步播放的问题。

TS流也是由一个或多个PES组合而来的，他们可以具有相同的时间基准，也可以不同。其基本的复用思想是，对具有相同时间基准的多个PES现进行节目复用，然后再对相互有独立时间基准的各个PS进行传输复用，最终产生出TS。

#### 1.1.2 参考链接：

1.PS流的格式和解析总结

https://blog.csdn.net/rankun1/article/details/69950873

2.MPEG-2 PS流

<https://blog.csdn.net/duanbeibei/article/details/1698183>

1. MPEG-TS 格式解析

https://blog.csdn.net/Kayson12345/article/details/81266587

1. 浅析mpeg-ts封装结构

<https://blog.csdn.net/michaeluo/article/details/75263462>

1. MPEG2-TS音视频同步原理(PCR dts pts)

<https://blog.csdn.net/charleslei/article/details/52299388>

### 1.2 PS流格式

将具有共同时间基准的一个或多个PES组合（复合）而成的单一的数据流称为节目流（Program Stream）。ES是直接从编码器出来的数据流，可以是编码过的视频数据流，音频数据流，或其他编码数据流的统称。ES流经过PES打包器之后，被转换成PES包。

针对H264 做PS 封装：每个IDR NALU 前一般都会包含SPS、PPS 等NALU，因此将SPS、PPS、IDR 的NALU 封装为一个PS 包，包括ps 头，然后加上PS system header，PS system map，PES header+h264 raw data。所以一个IDR NALU PS 包由外到内顺序是：

PSheader | PS system header | PS system Map | PES header | h264 raw data。

对于其它非关键帧的PS 包，就简单多了，直接加上PS头和PES 头就可以了。顺序为：

PS header | PES header | h264raw data。

以上是对只有视频video 的情况，如果要把音频Audio也打包进PS 封装，也可以。当有音频数据时，将数据加上PES header 放到视频PES 后就可以了。顺序如下：PS 包=PS头|PES(video)|PES(audio)，再用RTP 封装发送就可以了。

PES包长度域有两个字节，共16比特，因此PES包的最大长度是65535字节。

#### 1.2.1 PS包结构

PS包结构由包头、系统头、系统映射、PES包四部分构成。具体参考以上。

包头由PS包起始码、系统时钟基准（SCR-System Clock Reference）的基本部分、SCR的扩展部分和PS复用速率4部分组成。

结构如下：

#### IMG_2561.2.2 PS包头PSheader

位串，用来标志一个包的开始PACK identifier -- 0xBA，marker\_bit=1



SCR和SCR\_ext：是系统时钟参考，一个以27MHz驱动的计数器，用作同步流的参考。时钟除以300以上（以匹配PTS / DTS等90KHz时钟），商为SCR（33位），余数余数为SCR\_ext（9位）

Program\_Mux\_Rate：这是一个22位整数，指定程序流目标解码器在包含它的包中接收节目流的速率。 program\_mux\_rate的值以50字节/秒为单位进行测量。值0被禁止。该字段所表示的值用于在2.5.2中定义P-STD输入端的字节到达时间。该字段值在本标准中的节目多路复合流的不同包中取值可能不同。

pack\_stuffing\_length：一个3位整数，指定此字段后面的填充字节数。

stuffing\_byte：填充字节，这是一个固定的8位值，等于'1111 1111'，可以由编码器插入，例如，以满足通道的要求。它被解码器丢弃。

#### 1.2.3 PS系统头 PSsystem header：

位串，BB指出系统标题的开始

|  |  |  |
| --- | --- | --- |
| 语  法 | 位数 | 助记符 |
| system\_header() { |  |  |
| system\_header\_start\_code | 32 | bslbf |
| header\_length | 16 | uimsbf |
| marker\_bit | 1 | bslbf |
| rate\_bound | 22 | uimsbf |
| marker\_bit | 1 | bslbf |
| audio\_bound | 6 | uimsbf |
| fixed\_flag | 1 | bslbf |
| CSPS\_flag | 1 | bslbf |
| system\_audio\_lock\_flag | 1 | bslbf |
| system\_video\_lock\_flag | 1 | bslbf |
| marker\_bit | 1 | bslbf |
| vedio\_bound | 5 | uimsbf |
| packet\_rate\_restriction\_flag | 1 | bslbf |
| reserved\_bits | 7 | bslbf |
| while (nextbits()=='1') { |  |  |
| stream\_id | 8 | uimsbf |
| '11' | 2 | bslbf |
| P-STD\_buffer\_bound\_scale | 1 | bslbf |
| P-STD\_buffer\_size\_bound | 13 | uimsbf |
| } |  |  |
| } |  |  |

header\_length：标题长度字段，16位字段。指出该字段后的系统标题的字节长度。在本规范将来的扩充中可能扩展该字段。

rate\_bound ：速率界限字段，22位字段，取值不小于编码在节目流的任何包中的program\_mux\_rate字段的最大值。该字段可被解码器用于估计是否有能力对整个流解码。

audio\_bound：音频界限字段，6位字段，取值是在从0到32的闭区间中的整数，且不小于节目流中解码过程同时活动的GB/T XXXX.3和GB/T AAAA.3音频流的最大数目。在本小节中，若STD缓冲区非空或展现单元正在P-STD模型中展现，则GB/T XXXX.3和GB/T AAAA.3音频流的解码过程是活动的。

fixed\_flag：固定标志字段，1位标志位。置'1'时表示比特率恒定的操作；置'0'时，表示操作的比特率可变。在恒定比特率的操作期间，复合的GB/T XXXX.1流中的system\_clock\_reference字段值应遵从下面的线性公式:

SCR\_base(i)＝((c1×i＋c2) DIV 300) % 233 (2-22)

SCR\_ext(i)＝((c1×i＋c2) DIV 300) % 300 (2-23)

其中：

c1 对所有i均有效的实型常数；

c2 对所有i均有效的实型常数；

i 在GB/T XXXX.1复合流中包含任何system\_clock\_reference字段的最后一位的字节索引。

CSPS\_flag：CSPS标志字段，1位字段。置'1'时，节目流符合2.7.9中定义的限制

system\_audio\_lock\_flag：系统音频锁定标志字段，1位字段。表示在系统目标解码器的音频采样率和system\_clock\_frequency之间存在规定的比率。system\_clock\_frequency在2.5.2.1中定义而音频采样率由GB/T XXXX.3规定。如果对节目流中所有音频基本流的所有展现单元，system\_clock\_frequency和实际音频采样率的比例SCASR是恒定的，且对音频流中所指出的标准采样率和下表中数值相等，则该字段只能为'1'。

SCASR＝(system\_clock\_frequency) / audio\_sample\_rate\_in\_the\_P-STD (2-24)

记号X/Y表示实数除法。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 标准音频采  样频率(kHz) | 16 | 32 | 22.05 | 44.1 | 24 | 48 |
| SCASR | 27 000 000  --------  16 000 | 27 000 000  -------  32 000 | 27 000 000  ------  22 050 | 27 000 000  ------  44 100 | 27 000 000  ------  24 000 | 27 000 000  --------  48 000 |

system\_video\_lock\_flag：系统视频锁定标志字段，1位字段。表示在系统目标解码器的视频帧速率和system\_clock\_frequency之间存在规定的比率。system\_clock\_frequency在2.5.2.1中定义而视频帧速率由GB/T XXXX.2规定。如果对GB/T XXXX.1中所有视频基本流的所有展现单元，system\_clock\_frequency和实际视频帧速率的比例SCFR是恒定的，且对视频流中所指出的标准帧速率和下表中数值相等，则该字段只能为'1'。

SCFR＝system\_clock\_frequency / frame\_rate\_in\_the\_P-STD (2-25)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 标准帧速率(Hz) | 23.976 | 24 | 25 | 29.97 | 30 | 50 | 59.94 | 60 |
| SCFR | 1 126 125 | 1 125 000 | 1 080 000 | 900 900 | 900 000 | 540 000 | 450 450 | 450 000 |

比率SCFR的值是精确的。对于23.976，29.97或59.94帧/秒的标准速率，实际的帧速率与标准速率略有不同。

video\_bound：视频界限字段。5位字段，取值是在从0到16的闭区间中的整数且不小于节目流中解码过程同时活动的GB/T XXXX.2和GB/T AAAA.2流的最大数目。在本小节中，若P-STD缓冲区非空或展现单元正在P-STD模型中展现，则GB/T XXXX.2和GB/T AAAA.2视频流的解码过程是活动的。

packet\_rate\_restriction\_flag：分组速率限制标志字段，1位标志位。若CSPS标识为'1'，则该字段表示2.7.9中规定的哪个限制适用于分组速率。若CSPS标识为'0'，则该字段的含义未定义。

reserved\_bits：保留位字段，7位字段。被保留供ISO/IEC将来使用。它的值应为'111 1111'，除非ISO/IEC对它作出其它规定。

stream\_id：流标识字段，8位字段。指示其后的P-STD\_buffer\_bound\_scale和P-STD\_buffer\_size\_bound字段所涉及的流的编码和基本流号码。

若取值'1011 1000'，则其后的P-STD\_buffer\_bound\_scale和P-STD\_buffer\_size\_bound字段指节目流中所有的音频流。

若取值'1011 1001'，则其后的P-STD\_buffer\_bound\_scale和P-STD\_buffer\_size\_bound字段指节目流中所有的视频流。

若stream\_id取其它值，则应该是大于或等于'1011 1100'的一字节值且应根据表2-18解释为流的编码和基本流号码。

节目流中的每个基本流应在每个系统标题中通过这种机制精确地规定一次它的P-STD\_buffer\_bound\_scale和P-STD\_buffer\_size\_bound。

P-STD\_buffer\_bound\_scale：P-STD缓冲区界限比例字段。1位字段。表示用于解释后续P-STD\_buffer\_size\_bound字段的比例系数。若前面的stream\_id表示一个音频流，则该字段值为'0'。若表示一个视频流，则该字段值为'1'。对于所有其它的流类型，该字段值可以为'0'也可以为'1'。

P-STD\_buffer\_size\_bound：P-STD缓冲区大小界限字段。13位无符号整数，取值不小于节目流中流n的所有分组的P-STD缓冲区大小BSn的最大值。若P-STD\_buffer\_bound\_scale的值为'0'，则该字段以128字节为单位来度量缓冲区大小的边界。若P-STD\_buffer\_bound\_scale的值为'1'，则该字段以1024字节为单位来度量缓冲区大小的边界。因此：

if (P-STD\_buffer\_bound\_scale = = 0)

BSn≤P-STD\_buffer\_size\_bound×128

else

BSn≤P-STD\_buffer\_size\_bound×1024

所以对于系统头部的解析，我们一般只要先首先判断是否存在系统头（根据系统头的起始码0x000001BB），然后我们读取系统头的头部长度,即PS SYSTEM HEADER LENGTH部分，然后根据头部的长度，跳过PS系统头。进入下一个部分，即PS 节目流映射头。

#### 1.2.4 PS 映射头 PS system Map

PS流的节目映射流部分(节目流映射)定义

|  |  |  |
| --- | --- | --- |
| 语  法 | 位数 | 助记符 |
| program\_stream\_map() { |  |  |
| packet\_start\_code\_prefix | 24 | bslbf |
| map\_stream\_id | 8 | uimsbf |
| program\_stream\_map\_length | 16 | uimsbf |
| current\_next\_indicator | 1 | bslbf |
| reserved | 2 | bslbf |
| program\_stream\_map\_version | 5 | uimsbf |
| reserved | 7 | bslbf |
| marker\_bit | 1 | bslbf |
| program\_stream\_info\_length | 16 | uimsbf |
| for (i=0;i<N;i++){ |  |  |
| descriptor() |  |  |
| } |  |  |
| elementary\_stream\_map\_length | 16 | uimsbf |
| for (i=0;i<N1;i++){ |  |  |
| stream\_type | 8 | uimsbf |
| elementary\_stream\_id | 8 | uimsbf |
| elementary\_stream\_info\_length | 16 | uimsbf |
| for (i=0;i<N2;i++) { |  |  |
| descriptor() |  |  |
| } |  |  |
| } |  |  |
| CRC\_32 | 32 | rpchof |
| } |  |  |

packet\_start\_code\_prefix：分组起始码前缀字段。24位码。它和跟随其后的map\_stream\_id共同组成一个分组起始码以标志分组的开始。该字段是值为'0000 0000 0000 0000 0000 0001' (0x000001)的位串。

map\_stream\_id：映射流标识字段。8位字段，值为0xBC。

program\_stream\_map\_length：节目流映射长度字段。16位字段。指示紧跟在该字段后的program\_stream\_map中的字节数。该字段的最大值为1018(0x3FA)。

current\_next\_indicator：当前下一个指示符字段。1位字段。置'1'时表示传送的节目流映射当前是可用的。置'0'时表示传送的节目流映射还不可用，但它将是下一个生效的表。

program\_stream\_map\_version：节目流映射版本字段。5位字段，表示整个节目流映射的版本号。一旦节目流映射的定义发生变化，该字段将递增1，并对32取模。在current\_next\_indicator为'1'时，该字段应该是当前适用的节目流映射的版本号；在current\_next\_indicator为'0'时，该字段应该是下一个适用的节目流映射的版本号。

program\_stream\_info\_length：节目流信息长度字段。16位字段，指出紧跟在该字段后的描述符的总长度。

marker\_bit：标记位字段。1位字段，取值为'1'。

elementary\_stream\_map\_length：基本流映射长度字段。16位字段，指出在该节目流映射中的所有基本流信息的字节长度。它只包括stream\_type、elementary\_stream\_id和elementary\_stream\_info\_length字段。(这里注意一下，这里的基本流映射长度，他只包括他后面的指定的那几个定义字段的总和，即从从这个长度，我们可以知道后面他根了几种类型的流定义，因为一种流的这个定义字段：stream\_type(1BYTE)、elementary\_stream\_id(1byte)和elementary\_stream\_info\_length(2byte)字段总和为4个字节，所以用elementary\_stream\_map\_length/4可以得到后面定义了几个流类型信息。)

stream\_type：流类型字段 。8位字段，根据表2-29规定了流的类型。该字段只能标志包含在PES分组中的基本流且取值不能为0x05。

(这里我们暂时根据国标GB28181中的定义可以知道

1、MPEG-4 视频流： 0x10；2、H.264 视频流： 0x1B；3、SVAC 视频流： 0x80；4、G.711 音频流： 0x90；5、G.722.1 音频流： 0x92；6、G.723.1 音频流： 0x93；7、G.729 音频流： 0x99；8、SVAC音频流： 0x9B。

因为节目映射流字段只有在关键帧打包的时候，才会存在，所以如果要判断PS打包的流编码类型，就根据这个字段来判断。)

elementary\_stream\_id：基本流标识字段。8位字段，指出该基本流所在PES分组的PES分组标题中stream\_id字段的值。（这个字段的定义，其中0x(C0~DF)指音频，0x(E0~EF)为视频）

elmentary\_stream\_info\_length：基本流信息长度字段。16位字段，指出紧跟在该字段后的描述符的字节长度。(即这个类型的流描述长度。这个后面的字段后面的指定长度不在elementary\_stream\_map\_length指定的范围类。)

CRC\_32：CRC 32字段。32位字段，它包含CRC值以在处理完整个节目流映射后在附录A中定义的解码器寄存器产生0输出值。

对于这个字段的解析，我们需要取值0x000001BC的位串，指出节目流映射的开始，暂时不需要处理，读取Header Length直接跳过即可，如果需要解析流编码类型，必须详细解析这个字段。

#### 1.2.5 PES包 PES\_packet

PES包=PES header+code raw data。



这个是ps对pes包得包装。Ps自带标志位ba bb后面紧跟pes 包，pes包可以是eo bc等stream id标识。





|  |  |  |
| --- | --- | --- |
| 语  法 | 位数 | 助记符 |
| PES\_packet(){ |  |  |
| packet\_start\_code\_prefix | 24 | bslbf |
| stream\_id | 8 | uimsbf |
| PES\_packet\_length | 16 | uimsbf |
| if(stream\_id != program\_stream\_map |  |  |
| && stream\_id !=padding\_stream |  |  |
| && stream\_id !=private\_stream\_2 |  |  |
| && stream\_id !=ECM |  |  |
| && stream\_id !=EMM |  |  |
| && stream\_id !=program\_stream\_directory |  |  |
| && stream\_id !=DSMCC\_stream |  |  |
| && stream\_id !=ITU-T Rec.H.222.1 type E stream){ |  |  |
| '10' | 2 | bslbf |
| PES\_scrambling\_control | 2 | bslbf |
| PES\_priority | 1 | bslbf |
| data\_alignment\_indicator | 1 | bslbf |
| copyright | 1 | bslbf |
| original\_or\_copy | 1 | bslbf |
| PTS\_DTS\_flags | 2 | bslbf |
| ESCR\_flag | 1 | bslbf |
| ES\_rate\_flag | 1 | bslbf |
| DSM\_trick\_mode\_flag | 1 | bslbf |
| additional\_copy\_info\_flag | 1 | bslbf |
| PES\_CRC\_flag | 1 | bslbf |
| PES\_extension\_flag | 1 | bslbf |
| PES\_header\_data\_length | 8 | uimsbf |
| if(PTS\_DTS\_flags =='10'){ |  |  |
| '0010' | 4 | bslbf |
| PTS[32..30] | 3 | bslbf |
| marker\_bit | 1 | bslbf |
| PTS[29..15] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| PTS[14..0] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| } |  |  |
| if(PTS\_DTS\_flags =='11'){ |  |  |
| '0011' | 4 | bslbf |
| PTS[32..30] | 3 | bslbf |
| marker\_bit | 1 | bslbf |
| PTS[29..15] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| PTS[14..0] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| '0001' | 4 | bslbf |
| PTS[32..30] | 3 | bslbf |
| marker\_bit | 1 | bslbf |
| PTS[29..15] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| PTS[14..0] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| } |  |  |
| if(ESCR\_flag =='1'){ |  |  |
| reserved | 2 | bslbf |
| ESCR\_base[32..30] | 3 | bslbf |
| marker\_bit | 1 | bslbf |
| ESCR\_base[29..15] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| ESCR\_base[14..0] | 15 | bslbf |
| marker\_bit | 1 | bslbf |
| ESCR\_extension | 9 | uimsbf |
| marker\_bit | 1 | bslbf |
| } |  |  |
| if(ES\_rate\_flag =='1'){ |  |  |
| marker\_bit | 1 | bslbf |
| ES\_rate | 22 | uimsbf |
| marker\_bit | 1 | bslbf |
| } |  |  |
| if (DSM\_trick\_mode\_flag =='1'){ |  |  |
| trick\_mode\_control | 3 | uimsbf |
| if ( trick\_mode\_control = =fast\_forward ) { |  |  |
| field\_id | 2 | bslbf |
| intra\_slice\_refresh | 1 | bslbf |
| frequency\_truncation | 2 | bslbf |
| } |  |  |
| else if ( trick\_mode\_control = = slow\_motion ) { |  |  |
| rep\_cntrl | 5 | uimsbf |
| } |  |  |
| else if ( trick\_mode \_control = = freeze\_frame ) { |  |  |
| field\_id | 2 | uimsbf |
| reserved | 3 | bslbf |
| } |  |  |
| else if ( trick\_mode \_control = = fast\_reverse ) { |  |  |
| field\_id | 2 | bslbf |
| intra\_slice\_refresh | 1 | bslbf |
| frequency\_truncation | 2 | bslbf |
| else if ( trick\_mode\_control = = slow\_reverse ) { |  |  |
| rep\_cntrl | 5 | uimsbf |
| } |  |  |
| else |  |  |
| reserved | 5 | bslbf |
| } |  |  |
| if ( additional\_copy\_info\_flag = ='1'){ |  |  |
| marker\_bit | 1 | bslbf |
| additional\_copy\_info | 7 | bslbf |
| } |  |  |
| if (PES\_CRC\_flag==‘1’){ |  |  |
| previous\_PES\_packet\_CRC | 16 | bslbf |
| } |  |  |
| if ( PES\_extension\_flag =='1') { |  |  |
| PES\_private\_data\_flag | 1 | bslbf |
| pack\_header\_field\_flag | 1 | bslbf |
| program\_packet\_sequence\_counter\_flag | 1 | bslbf |
| P-STD\_buffer\_flag | 1 | bslbf |
| reserved | 3 | bslbf |
| PES\_extension\_flag\_2 | 1 | bslbf |
| if(PES\_private\_data\_flag =='1'){ |  |  |
| PES\_private\_data | 128 | bslbf |
| } |  |  |
| if (pack\_header\_field\_flag == '1'){ |  |  |
| pack\_field\_length | 8 | uimsbf |
| pack\_header() |  |  |
| } |  |  |
| if (program\_packer\_sequence\_counter\_flag == '1'){ |  |  |
| marker\_bit | 1 | bslbf |
| program\_packet\_sequence\_counter | 7 | uimsbf |
| marker-bit | 1 | bslbf |
| MPEG1\_MPEG2\_indentifier | 1 | bslbf |
| original\_stuff\_length | 6 | uimsbf |
| } |  |  |
| if (P-STD\_buffer\_flag = = '1'({ |  |  |
| '01' | 2 | bslbf |
| P-STD\_buffer\_scale | 1 | bslbf |
| P-STD\_buffer\_size | 13 | uimsbf |
| } |  |  |
| if (PES\_extension\_flag\_2 == '1'{ |  |  |
| marker\_bit | 1 | bslbf |
| PES\_extension\_field\_length | 7 | uimsbf |
| for(i=0;i<PES\_extension\_field\_length;i++){ |  |  |
| reserved | 8 | bslbf |
| } |  |  |
| } |  |  |
| } |  |  |
| for (i=0;i<N1;i++){ |  |  |
| stuffing\_byte | 8 | bslbf |
| } |  |  |
| for (i=0;i<N2;i++){ |  |  |
| PES\_packet\_data\_byte | 8 | bslbf |
| } |  |  |
| } |  |  |
| else if (stream\_id = = program\_stream\_map |  |  |
| || stream\_id = = private\_stream\_2 |  |  |
| || stream\_id = = ECM |  |  |
| || stream\_id = = EMM |  |  |
| || stream\_id = = program\_stream\_directory |  |  |
| || stream\_id = = DSMCC\_stream |  |  |
| || stream\_id = = ITU-T Rec. H.222.1 type E stream ){ |  |  |
| for (i=0;i<PES\_packet\_length;i++){ |  |  |
| PES\_packet\_data\_byte | 8 | bslbf |
| } |  |  |
| } |  |  |
| else if (steam\_id = = padding\_stream){ |  |  |
| for (i=0;i<PES\_packet\_length;i++){ |  |  |
| padding\_byte | 8 | bslbf |
| } |  |  |
| } |  |  |
| } |  |  |

以上都是根据steam\_id进行判断处理

表2-18 Stream\_id赋值

|  |  |  |
| --- | --- | --- |
| stream\_id | 注 | 流 编 码 |
| 1011 1100 | 1 | program\_stream\_map（0xBC） |
| 1011 1101 | 2 | private\_stream\_1(0xBD) |
| 1011 1110 |  | padding\_stream(0xBE) |
| 1011 1111 | 3 | private\_stream-2(0xBF) |
| 110x xxxx |  | GB/T XXXX.3或GB/T AAAA.3音频流编号xxxx(0xC0~0xDF) |
| 1110 xxxx |  | GB/T XXXX.2或GB/T AAAA.2视频流编号xxxx(0xE0~0xEF) |
| 1111 0000 | 3 | ECM\_stream(0xF0) |
| 1111 0001 | 3 | EMM\_stream(0xF1) |
| 1111 0010 | 5 | GB/T XXXX.1附录B或GB/T XXXX.6\_DSMCC\_stream(0xF2) |
| 1111 0011 | 2 | ISO/IEC\_13522\_stream(0xF3) |
| 1111 0100 | 6 | ITU-T Rec. H.222.1类型A |
| 1111 0101 | 6 | ITU-T Rec. H.222.1类型B |
| 1111 0110 | 6 | ITU-T Rec. H.222.1类型C |
| 1111 0111 | 6 | ITU-T Rec. H.222.1类型D |
| 1111 1000 | 6 | ITU-T Rec. H.222.1类型E |
| 1111 1001 | 7 | ancillary\_stream(0xF9) |
| 1111 1010…1111 1110 |  | 保留数据流 |
| 1111 1111 | 4 | program\_stream\_directory(0xFF) |

符号x表示值'0'或'1'均被允许且可产生相同的流类型。流号码由x的取值决定。

注

1 类型为program\_stream\_map的PES分组有唯一的语法，在2.5.4.1中作了规定。

2 类型为private\_stream\_1和ISO/IEC\_13352\_stream的PES分组与GB/T XXXX.2及GB/T XXXX.3音频流服从相同的PES分组语法。

3 类型为private\_stream\_2，ECM\_stream和EMM\_stream的PES分组与private\_stream\_1相似，除了在PES\_packet\_length字段后未规定语法。

4 类型为program\_stream\_directory的PES分组有唯一的语法，在2.5.5中作了规定。

5 类型为DSM\_CC\_stream的PES分组有唯一的语法，在GB/T XXXX.6中作了规定。

6 stream\_id与表2-29中的stream\_type 0x09相关联。

7 stream\_id仅用于PES分组。PES分组在传输流中携带了来源于节目流或GB/T AAAA.1系统流的数据(参见2.4.3.7)。

0x0000 01B9结束，即MPEG\_Program\_end\_code

PES分组中各字段的语义定义：

packet\_start\_code\_prefix：分组起始码前缀字段。24位代码，它和后面的stream\_id构成了标识分组开始的分组起始码。它是一个值为'0000 0000 0000 0000 0000 0001' (0x000001)的位串。

stream\_id：流标识字段。在节目流中，它规定了基本流的号码和类型。定义参见表2-18。在传输流中，它可以被设定为正确描述表2-18中定义的基本流类型的任何有效值。在传输流中，基本流类型在2.4.4的节目特定信息中作了规定。

PES\_packet\_length：PES分组长度字段。16位字段，指出了PES分组中跟在该字段后的字节数目。值为0表示PES分组长度要么没有规定要么没有限制。这种情况只允许出现在有效负载包含来源于传输流分组中某个视频基本流的字节的PES分组中。

PES\_scrambling\_control：PES加扰控制字段。2位字段，表示PES分组有效负载的加扰方式。当加扰发生在PES层，PES分组标题，如果有可选字段的话也包括在内，不应被加扰(参见2-19)。

PES\_priority：PES优先级字段。1位字段，指示PES分组中有效负载的优先级。'1'表示PES分组中有效负载的优先级高于该字段为'0'的PES分组有效负载。多路复合器能使用该字段来区分安排基本流中数据的优先级。传输机制不应改动该字段。

data\_alignment\_indicator：数据对齐指示符字段。1位标志。置'1'时表示PES分组标题后紧跟着在2.6.10中的data\_alignment\_indicator所指出的视频起始码或音频同步字，如果有data\_alignment\_indicator描述符的话。若其值为'1'且无该描述符，则需要在表2-47和2-48中alignment\_type '01'所表示的对齐。当值为'0'时，没有定义是否有任何此种的对齐。

Copyright：版权字段。1位字段。置'1'时表示相关PES分组有效负载的材料受到版权保护。当值为'0'时，没有定义该材料是否受到版权保护。2.6.24中描述的版权描述符与包含PES分组的基本流相关。若描述符作用于包含PES分组的材料，则版权标志被置为'1'。

original\_or\_copy：原始或拷贝字段。1位字段。置'1'时表示相关PES分组有效负载的内容是原始的；值为'0'表示相关PES分组有效负载的内容是一份拷贝。

PTS\_DTS\_flags：PTS DTS标志字段。2位字段。当值为'10'时，PTS字段应出现在PES分组标题中；当值为'11'时，PTS字段和DTS字段都应出现在PES分组标题中；当值为'00'时，PTS字段和DTS字段都不出现在PES分组标题中。值'01'是不允许的。

ESCR\_flag：ESCR标志字段。1位标志。置'1'时表示ESCR基础和扩展字段出现在PES分组标题中；值为'0'表示没有ESCR字段。

ES\_rate\_flag：ES速率标志字段。1位标志。置'1'时表示ES\_rate字段出现在PES分组标题中；值为'0'表示没有ES\_rate字段。

DSM\_trick\_mode\_flag：DSM特技方式标志字段。1位标志。置'1'时表示有8位特技方式字段；值为'0'表示没有该字段。

additional\_copy\_info\_flag：附加版权信息标志字段。1位标志。置'1'时表示有附加拷贝信息字段；值为'0'表示没有该字段。

PES\_CRC\_flag：PES CRC标志字段。1位标志。置'1'时表示CRC字段出现在PES分组标题中；值为'0'表示没有该字段。

PES\_extension\_flag ：PES扩展标志字段。1位标志。置'1'时表示PES分组标题中有扩展字段；值为'0'表示没有该字段。

PES\_header\_data\_length：PES标题数据长度字段。8位字段。指出包含在PES分组标题中的可选字段和任何填充字节所占用的总字节数。该字段之前的字节指出了有无可选字段。

marker\_bit：标记位字段。值为'1'的1位字段。

PTS：展现时间戳字段。展现时间与解码时间的关系如下：PTS是一个编码在三个分离字段中的33位数字。它指出了基本流n的第k个展现单元在系统目标解码器中的展现时间tpn(k)。PTS的值以系统时钟频率的1/300(即90 kHz)为单位。展现时间由PTS根据式2-11计算而来。对编码展现时间戳频率的约束参见2.7.4。

PTS(k)＝((system\_clock\_frequency×tpn(k)) DIV 300) % 233 (2-11)

其中，tpn(k)是展现单元Pn(k)的展现时间。对音频而言，若PES分组标题中有PTS，则它是指PES分组中开始的第一个存取单元。若PES分组中有音频存取单元的首字节，则有一个音频存取单元开始于该PES分组中。对视频而言，若PES分组标题中有PTS，则它是指包含PES分组中开始的第一个画面起始码的存取单元。若PES分组中有画面起始码的首字节，则有一个画面起始码开始于该PES分组中。对音频展现单元(PU)，low\_delay序列中的视频PU以及B画面，展现时间tpn(k)应等于tdn(k)。对于非low\_delay中的I画面和P画面，在存取单元(AU) k和k'之间无解码不连续时，展现时间tpn(k)应等于下一个传输的I画面或P画面的解码时间tdn(k) (参见2.7.5)。若有解码不连续或流终止，则tpn(k)和tdn(k)之间的差别应与初始流一直延续，没有不连续也没有终止时完全相同。

注1： low\_delay序列是low\_delay标志被设置的视频序列(参见GB/T XXXX.2中的6.2.2.3)。

若音频中有滤波，则系统模型假定滤波不会导致延迟。因此，编码时PTS所涉及的采样与解码时PTS所涉及的采样是相同的。对于可伸缩编码，参见2.7.6。

DTS：解码时间戳字段。DTS是一个编码在三个分离字段中的33位数字。它指出了基本流n的第j个展现单元在系统目标解码器中的解码时间tdn(j)。DTS的值以系统时钟频率的1/300 (即90 kHz)为单位。解码时间由DTS根据式2-12计算而来：

DTS(j)＝((system\_clock\_frequency×tdn(j)) DIV 300) % 233 (2-12)

其中，tdn(j)是存取单元An(j)的解码时间。

对视频而言，若PES分组标题中有DTS，则它是指包含PES分组中开始的第一个画面起始码的存取单元。若PES分组中有画面起始码的首字节，则该画面起始码开始于该PES分组中。

对于可伸缩编码，参见2.7.6。

ESCR\_base，ESCR\_extension：ESCR字段。42位字段，分两部分编码。第一部分是一个长度为33位的字段，其值ESCR\_base(i)由式2-14给出；第二部分是一个长度为9位的字段，其值ESCR\_ext(i)由式2-15给出。ESCR字段指出了基本流中包含ESCR\_base最后一个比特的字节到达PES-STD输出端的期望时间(参见2.5.2.4)。

特别地

ESCR(i)＝ESCR\_base(i)×300＋ESCR\_ext(i) (2-13)

其中：

ESCR\_base(i)＝((system\_clock\_frequency×t(i)) DIV 300) % 233 (2-14)

ESCR \_ext(i)＝((system\_clock\_frequency×t(i)) DIV 1) % 300 (2-15)

ESCR和ES\_rate字段(参见下面紧接的语义)包含与PES流序列相关的时间信息。这些字段应满足2.7.3中定义的约束。

ES\_rate：基本流速率字段 。22位无符号整数。对于PES流而言，它指出了系统目标解码器接收PES分组的速率。该字段在它所属的PES分组以及同一个PES流的后续PES分组中一直有效，直到遇到一个新的ES\_rate字段。该字段的值以50字节/秒为单位，且不能为0。该字段用于定义PES流的字节到达P-STD输入端的时间(参见2.5.2.4中的定义)。在各个PES分组中，编码在该字段中的值可能不同。

trick\_mode\_control：特技方式控制字段。3位字段。它表示作用于相关视频流的特技方式。对其它类型的基本流，该字段及其后5位的含义没有定义。trick\_mode状态的定义参见2.4.2.3的特技模式部分。当trick\_mode状态为假时，对GB/T XXXX.2视频而言，解码过程输出渐进序列中每幅画面的次数N由repeat\_first\_field和top\_field\_first字段来规定。对GB/T AAAA.2视频而言，由序列标题决定。对于隔行序列，当trick\_mode状态为假时，对GB/T XXXX.2视频而言，次数N由repeat\_first\_field和progressive\_frame字段来规定。

当trick\_mode状态为真时，画面的播放次数依赖于值N。

表2-20 特技方式控制值

|  |  |
| --- | --- |
| 值 | 描  述 |
| '000' | 快进 |
| '001' | 慢动作 |
| '010' | 冻结帧 |
| '011' | 快倒 |
| '100' | 慢倒 |
| '101'－'111' | 保留 |

当该字段值发生变化或特技模式操作停止时，可能会出现下列情况的任意组合：

l 时基不连续；

l 解码不连续；

l 连续性计数器不连续；

在特技模式的情况下，解码和展现的非标准速度可能会导致视频基本流数据中定义的某些字段值不正确。同样，片段结构的语义约束也可能无效。这些例外所涉及的视频语法元素为：

l bit\_rate;

l vbv\_delay;

l repeat\_first\_field;

l v\_axis\_positive;

l field\_sequence;

l subcarrier;

l burst\_amplitude;

l subcarrier\_phase;

在特技模式中，解码器不应该依赖于编码在这些字段中的值。

标准并不要求解码器能解码trick\_mode\_control字段。但是，能解码该字段的解码器应能满足以下标准要求。

fast forword：快进。trick\_mode\_control字段中的值'000'。当该值出现时，它表示一个快进视频流并定义了PES分组标题中后续5位的含义。intra\_slice\_refresh位可以被设定为'1'以指出可能有丢失的宏块。解码器可以用前一个解码画面中相同位置的宏块来代替。表2-21中定义的field\_id字段，表示应该显示哪个或哪些字段。frequency\_truncation字段指出了可能包括的一个系数受限集合。该字段值的含义如表2-22所示。

slow motion：慢动作。trick\_mode\_control字段中的值'001'。当该值出现时，它表示一个慢动作视频流，并定义了PES分组标题中后续5个比特的含义。对渐进序列而言，该画面应被显示N×rep\_cntrl时间，其中N定义如上。

对GB/T AAAA.2视频和GB/T XXXX.2视频渐进序列而言，该画面应被显示N×rep\_cntrl时间，其中N定义如上。

对GB/T XXXX.2隔行序列而言，该画面应被显示N×rep\_cntrl时间。若该画面是一个帧画面，则待显示的第一个字段在top\_field\_first为1时应该是顶字段，在top\_field\_first为0时，应该是底字段(参见GB/T XXXX.2)。该字段被显示N×rep\_cntrl / 2时间。该画面的其它字段被显示N－N×rep\_cntrl / 2时间。

freeze frame：冻结帧。trick\_mode\_control字段中的值'010'。当该值出现时，它表示冻结帧视频流，并定义了PES分组标题中后续5位的含义。表2-21中定义的field\_id字段，表示应该显示哪个(些)字段。field\_id字段指出了包含该字段的PES分组中开始的第一个视频存取单元，除非该PES分组包含0个有效负载字节。在后一种情况下，field\_id字段指出了最近的前一个视频存取单元。

fast reverse：快倒。trick\_mode\_control字段中的值'011'。当该值出现时，它表示一个快倒视频流并定义了PES分组标题中后续5位的含义。intra\_slice\_refresh位可以被设定为'1'以指出可能有丢失的宏块。解码器可以用前一个解码画面中相同位置的宏块来代替。表2-21中定义的field\_id字段，表示应该显示哪个或哪些字段。frequency\_truncation字段指出了可能包括的一个系数受限集合。该字段值的含义如表2-22“系数选择值”所示。

slow reverse：慢倒。trick\_mode\_control字段中的值'100'。当该值出现时，它表示一个慢倒视频流并定义了PES分组标题中后续5位的含义。对GB/T AAAA.2视频和GB/T XXXX.2视频渐进序列而言，该画面应被显示N×rep\_cntrl时间，其中N定义如上。

对GB/T XXXX.2隔行序列而言，该画面应被显示N×rep\_cntrl时间。若该画面是一个帧画面，则待显示的第一个字段在top\_field\_first为1时应该是底字段，在top\_field\_first为'0'时，应该是顶字段(参见GB/T XXXX.2)。该字段被显示N×rep\_cntrl/2时间。该画面的其它字段被显示N－N×rep\_cntrl / 2时间。

field\_id：字段标识字段。2位字段，表示应该显示哪个(些)字段。根据表2-21对其进行编码。

表2-21 field\_id字段控制值

|  |  |
| --- | --- |
| 值 | 描  述 |
| '00' | 仅自顶向下播放 |
| '01' | 仅自底向上播放 |
| '10' | 播放所有帧 |
| '11' | 保留 |

intra\_slice\_refresh：片内参考字段。1位标志。置'1'时表示PES分组的视频数据编码片中可能有丢失的宏块；置'0'时，表示上述情况可能不出现。更多的信息可参见GB/T XXXX.2。解码器可以用前一个解码画面中同一个位置的宏块来代替丢失的宏块。

frequency\_truncation：频率截断字段。2位字段。指出在对PES分组中数据进行编码时可能用到受限系数集合。其值定义于表2-22。

表2-22 系数选择值

|  |  |
| --- | --- |
| 值 | 描述 |
| '00' | 仅DC系数非0 |
| '01' | 仅前三个系数非0 |
| '10' | 仅前六个系数非0 |
| '11' | 所有系数均可能非0 |

rep\_cntrl：显示次数控制字段。5位字段，指出隔行画面中每一字段的显示次数或渐进画面显示次数。对隔行画面而言，顶字段或底字段是否应首先显示是视频序列标题中trick\_mode\_control字段和top\_field\_first字段的功能。该字段值不能为'0'。

additional\_copy\_info：附加版权信息字段。7位字段，包含与版权信息有关的专用数据。

previous\_PES\_packet\_CRC：前PES分组CRC字段。16位字段。在对前一个PES分组(不包括该PES分组的标题)进行处理后，该字段包含一个在解码器的16个寄存器中生成0输出的CRC值。该CRC值与附录A中所定义的相类似，但具有以下多项式：

x16＋x12＋x5＋1

注2： 该CRC值是为了用于网络维护，例如将有间隙性错误的源隔离开来，而不是为了供基本流解码器使用。它仅用于计算数据字节，因为在传输过程中PES分组标题数据可能被修改。

PES\_private\_data\_flag：PES专用数据标志字段。1位标志。置'1'时表示PES分组标题中包含专用数据；置'0'时表示PES分组标题中无专用数据。

pack\_header\_field\_flag：包标题字段标志字段。1位标志。置'1'时表示PES分组标题中有GB/T AAAA.2包标题或节目流包标题。若该字段在包含于节目流中的PES分组中，其值应为'0'。在传输流中，当值为'0'时表示PES标题中无包标题。

program\_packet\_sequence\_counter\_flag：节目分组序列计数标志字段。1位标志。值为'1'时表示PES分组有program\_packet\_sequence\_counter，MPEG1\_MPEG2\_identifier和original\_stuff\_length字段。值为'0'时表示PES分组标题中无这些字段。

P-STD\_buffer\_flag：P-STD缓冲区标志字段。1位标志。置'1'时表示PES分组标题中有P-STD\_buffer\_scale和P-STD\_buffer\_size字段。值为'0'时表示PES标题中无这些字段。

PES\_extension\_flag\_2：PES扩展标志字。1位标志，置'1'时表示有PES\_extension\_field\_length及相关字段。

PES\_private\_data：PES专用数据字段。16位字段。包含专用数据。这些数据与其前后的字段组合在一起时，不能与packet\_start\_code\_prefix (0x000001)冲突。

pack\_field\_length：包字段长度字段。8位字段。表示pack\_header\_field()以字节为单位时的长度。

program\_packet\_sequence\_counter：节目分组序列计数字段。7位字段。它是一个可选的计数器，随着来自于节目流或GB/T AAAA.1流的每一个后续的PES分组或传输流中具有单个节目定义的PES分组而递增，以提供与连续性计数器(参见2.4.3.2)相似的功能。它能用于检索节目流或原始GB/T AAAA.1流中的初始PES分组序列。该计数器在达到最大值后回卷为0。PES分组不能出现重复。因此，复合节目中任何两个连续的PES分组不应具有相同的program\_packet\_sequence\_counter值。

MPEG1\_MPEG2\_identifier：MPEG1 MPEG2标识符字段。1位标志。置'1'时表示PES分组携带的信息来自于GB/T AAAA.1流；置'0'时表示PES分组携带的信息来自于节目流。

original\_stuff\_length：初始填充长度字段。6位字段。指定用于初始GB/T XXXX.1分组标题或初始GB/T AAAA.1分组标题中的填充字节数。

P-STD\_ buffer\_scale：P-STD缓冲区比例字段。1位字段。仅当该PES分组包含于节目流中时才有意义。它指出了用来解释后续P-STD\_buffer\_size字段的比例因子。若前面的stream\_id表示一个音频流，该字段值应为'0'；若前面的stream\_id表示一个视频流，该字段值应为'1'。对于所有的其它流类型，其值可以为'0'或'1'。

P-STD\_buffer\_size：P-STD缓冲区大小字段 。13位无符号整数。仅当该PES分组包含于节目流中时才有意义。它定义了P-STD输入缓冲区的大小BSn。若P-STD\_ buffer\_scale的值为'0'，那么P-STD\_buffer\_size以128字节为单位来度量缓冲区的大小。若P-STD\_buffer\_scale的值为'1'，那么P-STD\_buffer\_size以1024字节为单位来度量缓冲区的大小。因此：

if ( P-STD\_buffer\_scale == 0)

BSn＝P-STD\_buffer\_size×128

else

BSn＝P-STD\_buffer\_size×1024

当该字段被GB/T XXXX.1系统目标解码器收到后，其编码值立即生效。

PES\_extension\_field\_length：PES扩展字段长度字段。7位字段。指出了跟在该字段之后在PES扩展字段中直到且包括任何保留字节的数据的字节长度。

stuffing\_byte ：填充字节字段。8位字段，其值恒定为'1111 1111'。可以由编码器插入以满足通道的需求等。解码器丢弃该字段。一个PES分组标题中只能出现32个填充字节。

PES\_packet\_data\_byte ：PES分组数据字节字段。该字段应该是来自于由分组的stream\_id或PID所指定的基本流的连续数据字节。当基本流数据符合GB/T XXXX.2或GB/T XXXX.3时，该字段应该是与本标准的字节相对齐的字节。基本流的字节序应得到保持。该字段的字节数N由PES\_packet\_length字段规定。N应等于PES\_packet\_length减去在PES\_packet\_length字段的最后一个字节与第一个 PES\_packet\_data\_byte间的字节数。

padding\_byte：填料字节字段 ，8位字段，其值恒定为'1111 1111'。该字段被解码器丢弃。

#### 1.2.6 PES中DTS&PTS

可以看到PTS/DTS（流识别码，用于区别不同性质ES）是打在PES包里面的，这两个参数是解决视音频同步显示，防止解码器输入缓存上溢或下溢的关键。PTS表示显示单元出现在系统目标解码器(STD: system target decoder)的时间，DTS表示将存取单元全部字节从STD的ES解码缓存器移走的时刻。每个I、P、B帧的包头都有一个PTS和DTS，但PTS与DTS对B帧都是一样的，无须标出B帧的DTS。对I帧和P帧，显示前一定要存储于视频解码器的重新排序缓存器中，经过延迟（重新排序）后再显示，一定要分别标明PTS和DTS。

关于音视频的同步：

除了PTS和DTS的配合工作外，还有一个重要的参数是SCR(system clock reference)。在编码的时候，PTS，DTS和SCR都是由STC(system time clock)生成的，在解码时，STC会再生，并通过锁相环路（PLL－phase lock loop），用本地SCR相位与输入的瞬时SCR相位锁相比较，以确定解码过程是否同步，若不同步，则用这个瞬时SCR调整27MHz的本地时钟频率。最后，PTS，DTS和SCR一起配合，解决视音频同步播放的问题。

### 1.3 TS流格式

#### 1.3.1 简介

因为Apple公司提出的HLS(http live streaming)格式的流行，mpeg-ts封装的文件在互联网上已经随处可见。这套体系的强大之处就在于它的简单，只要你有工具可以把一个或多个视频文件切割成一堆的小ts文件，并且生成一个内容非常简单的m3u8文件，然后只要有一个标准的web server（apache, nginx）就能对外提供点播视频流媒体服务了。假如可以对直播数据流进行切片存储成小文件，那就又具备了直播流媒体服务能力。

介绍一个非常好用的解析库bitstream，并提供据此库实现的ts解析工具与大家共同学习探讨。在MPEG-2系统中编码端实现单一码流的详细过程如下：

（1）. A/D转换后，通过MPEG-2压缩编码得到ES基本流。该流数据量很大，并且只是I P B帧的取样信息。

（2）.通过PES打包器，打包并在每个帧中插入PTS/DTS标识，编程PES。原来是流的格式，现在成了数据包的分割形式。

（3）.PES根据需要打包成PS或者TS包进行存储

MPEG-TS一种标准数据容器格式，传输与存储音视频、节目与系统信息协议数据，应用于数字广播系统，譬如DVB,ATSC与IPTV。MPEG2/DVB是一种多媒体传输、复用技术，在数字电视广播中可提供数百个节目频道。注意，DVB全称为Digital Video Broadcasting，包括不同的系统，如卫星数字电视广播系统，有线数字电视广播系统，地面开路数字电视广播系统，交互式数字电视广播系统以及数字电视加扰系统。DVB系统标准是一种全球数字电视技术的标准。MPEG-2标准是定义比特流的语法与句法的一个ISO/IEC标准，即13818-1标准。DVB系统的核心技术是采用MPEG-2技术进行视频、音频的编码，使用统一的MPEG-2传输流（TS流）。

MPEG-2标准中，有两种不同的码流输出到信道，一种是节目码流（PS: Program Stream），适用于没有传输误差的场景；一种是传送流（TS：Transport Stream)，适用于有信道噪声的传输场景。节目流设计用于合理可靠的媒体，如光盘（如DVD），而传输流设计用于不太可靠的传输，即地面或卫星广播。此外，传输流可以携带多个节目。

MPEG-2 system（编号13818-1）是MPEG-2标准的其中一部分，该部分描述了多个视频，音频和数据多种基本流（ES）合成传输流（TS）和节目流（PS）的方式。一路TS比特流通常由连续的固定字节的TS包组成，所包含的内容有：

（1）.一路或多路视频流（多个PES包组成，每个PES包的PID是一致的，一个PES包可能由若干个TS包组成）

（2）.一路或多路音频流（通常为杜比的音频格式）

（3）.一路或多路字幕

（4）.PSI表格信息（Program Specific Information，包括PAT与PMT表，即节目关联表与节目映射表）

（5）.PES: Packetized Elementary Stream，一路基本码流（如MEPG2视频流）会在编码器端被打包成PES流，由多个PES包组成，打包的过程中主要加入了PTS/DTS信息。

TS过一段时间要有一个pat，否则可能被认为是错误的流。

PAT描述有多少路节目，每路节目的PMT表的PID是多少，PMT则描述了本节目有多少流，每一路流的类型与PID是多少，举个例子，你找个一个TS包，它的PID是0，说明它的负载内容是PAT信息，解析PAT信息，你发现节目1的PMT表的PID是0x10，接着，你在比特流中寻找一个PID为0x10的TS包，它的负载内容是节目1的PMT表信息，解析该PMT信息，你可以发现第一路流是MPEG2音频流，PID号0x21，第二路流是MPEG2视频流，PID号是0x22，第三路流是DVB字幕流，PID号是0x23，解析完毕，凡是比特流中PID号为0x22的TS包，所负载的内容为MPEG2视频流，把这些包一个一个找出来，把其中的有效码流一部分一部分的拼接起来，然后送给解码器去解码。PAT介绍有多少路节目和节目对应PMT,PMT介绍有多少流和对应流。

注意，就一般的视频流而言，只要拼接成一个完整的PES包，就可以送出去给解码器，然后再继续拼接下一个PES包。

TS PS是在pes流包得基础上插入了所需要得数据，如上图所示。TS是将pes所有内容进行分组切割成188字节，每个字节都分成ts包头和ts内容。

PSI：节目特定信息，该表格信息用来描述传送流的组成结构。PSI信息由四种类型的表组成，包括节目关联表（PAT），节目映射表（PMT），条件接收表（CAT），网络信息表（NIT）。PAT与PMT两张表帮助我们找到该传送流中的所有节目与流。



#### 1.3.2 TS结构

**TS是将pes所有内容进行分组切割成188字节，每个字节都分成ts包头和ts内容。**

**TS头一共32位 。TS头4字节，TS内容184字节。**

TS流的形成过程：

1) 将原始音视频数据压缩之后，压缩结果组成一个基本码流（ES）。

2) 对ES（基本码流）进行打包形成PES。

3) 在PES包中加入时间戳信息(PTS/DTS)。

4) 将PES包内容分配到一系列固定长度的传输包（TS Packet）中。

5) 在传输包中加入定时信息(PCR)。

6) 在传输包中加入节目专用信息(PSI) 。

7) 连续输出传输包形成具有恒定比特率的MPEG-TS流。

TS流的解析过程，可以说是生成的逆过程：

1) 从复用的MPEG-TS流中解析出TS包；

2) 从TS包中获取PAT及对应的PMT（PSI中的表格）；

3) 从而获取特定节目的音视频PID；

4) 通过PID筛选出特定音视频相关的TS包，并解析出PES；

5) 从PES中读取到PTS/DTS，并从PES中解析出基本码流ES；

6) 将ES交给解码器，获得压缩前的原始音视频数据

主要的字段解析如下：

**Sync byte:** 同步字节，值为0x47；可以理解为开始标志位。

**Transport error indicator:** 传输错误指示位，置1时，表示传送包中至少有一个不可纠正的错误位。

**Payload unit start indicator:** 负载单元起始指标位，表示TS包的有效净荷以PES/PSI包的第一个字节开始，举个例子，一个PES包可能由多个TS包构成，第一个TS包的负载单元起始指标位才会被置位。

**Transport priority:** 传输优先级，表明该包比同个PID的但未置位的TS包有更高的优先级。

**PID:** 该TS包的ID号，如果净荷是PAT包，则PID固定为0x00。

**Transport scrambling control:** 传输加扰控制位

**Adaption field control:** 自适应调整域控制位，置位则表明该TS包存在自适应调整字段。

**Continuity counter:** 连续计数器，随着具有相同PID的TS包的增加而增加，达到最大时恢复为0，如果两个连续相同PID的TS包具有相同的计数，则表明这两个包是一样的，只取一个解析即可。 可以理解为相同pid不同得ts包id

**Payload:** 负载内容，可能为PAT/PMT/PES。data\_byte为1B长度的数据，为负载字节。

#### IMG_256IMG_2561.3.3 Adaptation field解析

Adaptation field是自适应数据区，用以指示接下来的数据是否为逻辑上独立的一段内容，并且存放了PCR信息。是否存在由TS头中**Adaption field control**控制。

#### IMG_2561.3.4 PAT包解析

根据从头部取得的PID接下来要做就是确定这个PID所属的类型。而类型可分为TS固有类型与音视频流类型等。TS流的特点是没有严格意义上的起始或结束部分，所以也不能像mp4或者avi格式那样从meta域取得整体信息，这样我们势必要有一个办法知道当前的TS流的结构是怎样的，包含哪些音视频流等。所有的一切都要从一个叫做PAT的包开始。

TS固有类型的第一个要知道的就是PAT包，它的全称是Program Association Table（节目关联表）。PID的值等于0就判断是否PAT包。PAT包采用section格式封装，section结构简单说来可分为头部与负载区，头部区由若通用干字段组成。





**table\_id:** 标识一个TS PSI 分段的内容是节目关联分段，条件访问分段还是节目映射分段。对于PAT，置为0x00。

**section\_syntax\_indicator:** 对于PAT，置为0x01。

**section\_length**: 分段长度字段，其值为从section\_length（包括在内）到CRC\_32字段的字节数，其值不超过1021。

**transport\_stream\_id**: 区别与其他复用流的标识。

**version\_number:** PAT的版本号，如果PAT有变，则版本号加1。

**current\_next\_indicator**:置0时，表明该传送的表分段不能使用，下一个表分段才有效。

**section\_number**: 表明该TS包属于该PAT的第几个分段，分段号从0开始。

**last\_section\_number**: 表明最后一个分段号，同时表明该PAT的最大分段数目。一般，一个PAT表由一个TS包传送。

**program\_number**: 节目的编号。

**network\_PID:** NIT表的PID值。

**program\_map\_PID**: PMT表的PID值。 指定节目pmt对应pid值

**CRC\_32:** CRC校验。

#### IMG_256**1.3.5 PMT包解析**

PMT包括了一个节目相关的所有流信息。流这个概念对应的是一路音频或一路视频。我们知道，一个节目播放时有图像也有声音，但在编码层面它们往往是分开并且独立的，因为它们在最终呈现时是由不同的硬件设备来渲染的。音频与视频的同步则要借助于PCR，它是同步参考时钟，决定着哪一帧画面与哪一段音频相对应。





**table\_id**: 标识一个TS PSI 分段的内容是节目关联分段，条件访问分段还是节目映射分段。对于PMT，置为0x02。

**section\_syntax\_indicator**: 对于PMT，置为0x01。

**section\_length**: 分段长度字段，其值为从section\_length（包括在内）到CRC\_32字段的字节数，其值不超过1021。

**program\_number:** 表明一共有多少个节目。

**version\_number**: PMT的版本号，如果字段中有关信息有变，则版本号以32为模加1。版本号是对一个节目的定义。

**current\_next\_indicator**:置0时，表明该传送的表分段不能使用，下一个表分段才有效。

section\_number: 总为0x00。

**last\_section\_number**: 总为0x00。

**PCR\_PID:** 指示含有该节目的PCR字段的TS包的PID。

**program\_info\_length**: 表明跟随其后的对节目信息描述的字节数，也就是第一个N loop descriptors的字节数。

**stream\_type:** 表明PES流的类型。譬如，0x01表明是MPEG-1视频，0X03表明是MPEG-1音频。

**elementary\_PID**: 表明该负载有该PES流的TS包的PID值。

**ES\_info\_length**: 表明跟随其后的描述相关节目元素的字节数，也就是第二个N loop descriptors的字节数。

**CRC\_32:** 在CEDARX代码中仅对DVB的场景下作校验。



Streamtype值的列表：

0x00 ITU-T | ISO/IEC Reserved

0x01 ISO/IEC 11172 Video

0x02 ITU-T Rec. H.262 | ISO/IEC 13818-2 Video or ISO/IEC 11172-2 constrainedparameter video stream

0x03 ISO/IEC 11172 Audio

0x04 ISO/IEC 13818-3 Audio

0x05 ITU-T Rec. H.222.0 | ISO/IEC 13818-1 private\_sections

0x06 ITU-T Rec. H.222.0 | ISO/IEC 13818-1 PES packets containing private data

0x07 ISO/IEC 13522 MHEG

0x08 ITU-T Rec. H.222.0 | ISO/IEC 13818-1 Annex A DSM-CC

0x09 ITU-T Rec. H.222.1

0x0A ISO/IEC 13818-6 type A

0x0B ISO/IEC 13818-6 type B

0x0C ISO/IEC 13818-6 type C

0x0D ISO/IEC 13818-6 type D

0x0E ITU-T Rec. H.222.0 | ISO/IEC 13818-1 auxiliary

0x0F ISO/IEC 13818-7 Audio with ADTS transport syntax

0x10 ISO/IEC 14496-2 Visual

0x11 ISO/IEC 14496-3 Audio with the LATM transport syntax as defined in ISO/IEC14496-3 / AMD 1

0x12 ISO/IEC 14496-1 SL-packetized stream or FlexMux stream carried in PES packets

0x13 ISO/IEC 14496-1 SL-packetized stream or FlexMux stream carried inISO/IEC14496\_sections.

0x14 ISO/IEC 13818-6 Synchronized Download Protocol

0x15-0x7F ITU-T Rec. H.222.0 | ISO/IEC 13818-1 Reserved

0x80-0xFF User Private.

#### 1.3.6 SDT包解析

SDT是流信息描述表，它的作用是传输节目的内容名称。例如我们在制作节目单时就可以这个表中取出cctv 1, cctv2这样的内容名称。



### 1.4 PTS DTS PCR同步机制

#### 1.4.1 PTS DTS PCR介绍

当新节目的PCR到达解码器时，需要更新时间基点，STC就被置位。通常第一个从解复用器中 解出的PCR被直接装入到STC计数器，其后PLL闭环操作。每当一个新节目的PCR到达解码器时，此值被认为是锁相环的参考频率，用来与STC的当前值 比较，产生的差值e经过脉宽调制后被输入低通滤波器并经放大，输出控制信号f，用来控制振荡器（VCO）的瞬时频率，VCO输出的频率是在27MHz左右 振荡的信号，作为解码器的系统时钟。27MHz时钟经过波形整理后输入到计数器中，产生当前的STC值，其33bits的90kHz部分用于和 PTS/DTS比较，产生解码和显示的同步信号。其实pcr用于生成27mz的系统时钟与PTS/DTS进行比较进行同步。

MPEG-2系统用于视音频同步以及系统时钟恢复的时间标签分别在ES，PES和TS这3个层次中。在ES层，与同步有关的主要是视频缓冲验证VBV（Video Buffer Verifier），用以防止解码器的缓冲器出现上溢或者下溢；在PES层，主要是在PES头信息里出现的显示时间标签PTS（Presentation Time Stamp）和解码时间标签DTS（Decoding Time Stamp）；在TS层中，TS头信息包含了节目时钟参考PCR（Program Clock Reference），用于恢复出与编码端一致的系统时序时钟STC（System Time Clock）。在节目流PS包头中加入SCR,它的作用与PCR域相似。标准规定在原始音频和视频流中，PTS的间隔不能超过0.7s，而出现在TS包头的PCR间隔不能超过0.1s。

编码器中有一个系统时钟STC（其频率是27MHz）（是编码器中自带的），此时钟用来产生指示音视频的正确显示和解码的时间标签，同时可用来指示在采样过程中系统时钟本身的瞬时值。指示音视频显示时间的时间标签称为显示时间标签（PTS），指示音视频的解码时间标签称为解码时间标签（DTS），指示系统时钟本身的瞬时值的时间标签称为节目参考时钟标签（PCR）。解码器通过VBV\_delay（视频流延时值，在解码时利用视频流缓冲区把视频流缓存到相应的vbv\_delay时间后，再启动解码器解码、显示、实现音视频的同步。vbv\_delay存在于视频ES的头部，长度为16bit）的数值来确定解码的开始，用解码时间标签（DTS）和显示时间标签（PTS）来确定解码和显示的次序，用PCR来获得系统时钟的同步。解码器中的系统时钟及其同步就是依靠这些时间标签来进行恢复和修正的。

PCR的插入必须在PCR字段的最后离开复用器的那一时刻，同时把27MHz系统时钟的采样瞬时值作为PCR字段插入到相应的PCR域。是放在TS包头的自适应区中传送。27 MHz时钟经波形整理后分两路，一路是由27MHz脉冲直接触发计数器生成扩展域PCR\_ext，长度为9bits。另一路经一个300分频器后的90 kHz脉冲送入一个33位计数器生成90KHZ基值，列入PCR\_base（基值域），长度33bits，用于和PTS/DTS比较，产生解码和显示所需要的同步信号。这两部分被置入PCR域，共同组成了42位的PCR。

PCR-Base(i)=90kHz×t(i), mod(233)

PCR-Ext(i)=27MHz×t(i), mod(300)

PCR(i)=PCR-Base(i)×300＋PCR-Ext(i)

PTS域为33bits，它也是一个33b的计数值，也是对系统时钟的300分频的时钟的计数值。被编码成为3个独立的字段，表示此分组中第一个访问单元在系统目标解码器中的预定显示时间。

DTS域也为33bits，编码成为3个独立的字段，表示此分组中第一个访问单元在系统目标解码器中的预定解码时间。DTS就视频来说，因为视频编码的时候用到了双向预测，一个图像单元被解出，并非马上就被显示，可能在存储器中留一段时间，作为其余图像单元的解码参考，在被参考完毕后，才被显示。针对视频的显示，Mpeg还提出了一个视频PTS。针对音频和视频的同步显示，又提出了一个音频PTS。由于声音没有用到双向预测，它的解码次序就是它的显示次序，故对它只提出PTS的概念。

解码器在解码时，首先利用PCR重建和编码器同步的27 MHz系统时钟，恢复27 MHz系统时钟后，再利用PES流中的DTS，PTS进行音频与视频间的同步，同时利用VBV\_Delay设置解码器的缓冲时长，后启动初始解码。下图是利用PCR恢复STC的原理图。

PCR-Base的作用是在解码器切换节目时，提供对解码器PCR计数器的初始值，以让该PCR值与PTS、DTS最大可能地达到相同的时间起点。PCR-ext的作用是通过解码器端的锁相环电路修正解码器的系统时钟，使其达到和编码器一致的27MHz。

在PCR域的33 bits中的90 kHz部分（即PCR-Base域）用于与PTS和DTS作比较，当二者相同时，相应的单元被显示或者解码。但按照MPEG-2标准的规定，PTS/DTS位于PES包的包头中，而解复用器所解出的PES包头的字节不被送入到任何基本流解码器的输入缓冲区，仅用于控制各解码器工作。每当解复用器解到一个PTS/DTS时，PTS/DTS就作为其后送入各个基本流解码器的输入缓冲器中的数据的显示/解码时间，但此时送入信道缓冲器中的基本流并不一定立即被解码，因此基本流解码器必须记录当前这个PTS/DTS及所对应的码流位置，以利于后面的解码和显示同步。

针对以上分析，我们采用如下方法来达到解码与显示的同步：在每一个基本流解码器中建立一个33 bits的PTS/DTS先入先出存储器（FIFO），用以存放PTS/DTS值，同时在解复用器中加入一个检测电路，每当码流开始新的一帧图像时，则产生一个信号，用来通知基本流解码器，并将该帧图像的PTS/DTS写入到FIFO中。每当解码器解到一帧图像时，就从FIFO中读出相应的PTS/DTS，此PTS/DTS与STC进行比较，当两者相等时，当前图像就开始显示/解码。

Mpeg-2规定的系统时钟频率为27MHz，传输流中的PCR，PTS/DTS等均为对该共同系统时钟的采样值。解码端捕获PCR，恢复出本地的STC，作为音视频同步控制的基准，并依据PTS（DTS）时间标签来安排解码和显示时间表，使音视频分别同步于STC，以实现音视频之间的同步。标准规定在原始音频和视频流中，PTS的间隔不能超过0.7s，而出现在TS包头的PCR间隔不能超过0.1s。

在MPEG-2标准中，并非每一个PES包或每一帧图像均有PTS和DTS，一般PS流中每两个PTS之间的间隔是0.7 s，而TS流中每两个PTS之间的间隔是0.1 s，所以在没有PTS和DTS的一帧图像出现时，我们可以在前一PTS/DTS的基础上加一增量得到对应该帧图像的PTS/DTS，并将计算出来的新的PTS/DTS插入到存放PTS/DTS的FIFO中去，

即：PTSn=PTSn－1+ΔPTS（n为第n帧图像）。

解码器同步算法总结如下：

（1）解码器从输入码流的包头中解出时间信息PCR送入到系统时间时钟恢复电路；系统时间时钟恢复电路在接收到每一个新的PCR时，进行本地系统时间时钟恢复和锁相。

（2）解复用器后，从PES包头中解出显示时间标签PTS和解码时间标签DTS，并送入到基本流解码器中。

（3）基本流解码器在接收到新的PTS/DTS后，存入对应的FIFO(先进先处存储器)中进行管理；对于没有PTS/DTS的显示单元，需要对其时间标签进行插值，并送入到FIFO中管理。

（4）每一显示单元开始解码前，用其对应的DTS与STC进行比较，当STC与DTS相等时开始解码；

（5）每一显示单元开始显示前，用其对应的PTS与STC进行比较，当STC与PTS相等时开始显示。

27 MHz系统时钟经过300分频后，得到本地的33 bits PCR\_Base，该时钟与寄存器中当前图像的PTS/DTS进行比较，系统软件根据比较结果做出相应的处理：

（1）若当前的PTS/DTS比PCR计数器的值小于半帧以上，即PTS\_Base≤-ΔPTS/2，此时说明系统解码过慢，解码器处于失步状态，应根据该帧的结构做出相应的同步调整；

（2）若当前的PTS/DTS比PCR计数器的值在半帧时间以内，我们认为此时系统解码正常，立即显示/解码当前帧；

（3）若当前的PTS/DTS大于PCR计数器的值，则此时解码器稍快，在这种情况下，只需等到PCR与PTS/DTS相等时，就可显示/解码。

#### 1.4.2 PCR测量

㈠PCR 精度(PCR\_AC)：接收PCR中所含27MHz时钟的不准确度，但不包含任何传输定时损伤。测量时传输码流中PCR字节位置作为起点，计算出PCR 到达时间。

㈡PCR漂移率(PCR\_DR)：PCR漂移率测量指的是PCR中的低频误差并计入了由PCR发生和再生引起的误差以及由传输损伤所引起的到达时间误差。即PCR内所含27MHz时钟的低频变化速率，测量时以某一稳定外部基准作为参考，因此包含了任何传输定时变化。下面的例子可用来说明PCR 漂移率的重要性：我们来看一下由PLL控制的本地时钟再生器。再生器输出端的信号频率应当跟踪输入信号频率，随着输入信号频率的变化，PLL也随之作出响应，本地再生器将跟踪输入信号，随着输入信号变化速率的增加，PLL的跟踪能力会有所降低，最后将导致本地发生信号失锁而丢失节目。PCR 漂移率的测量给出了一种测定方法，即如何更好地使去复用器/解码器或再复用器的节目时钟锁定于输入TS中所包含的PCR(注意这里指的是漂移率的测量而不是绝对频率的测量，接收机端的PLL电路可以锁定于系统时

钟，该时钟有着固定的频率误差－如果是快速变化或漂移率大，则会出现问题)。

㈢PCR总抖动(PCR\_OJ)：PCR内所含27MHz时钟的高频变化，测量时以某一稳定外部基准作为参考，因此包含了任何传输定时变化。它是PCR测量中最重要的一项，是PCR中高频误差的总体测量并且还包括来自PCR发生和PCR再生的误差以及由传输损伤引起的到达时间误差。例如该项传输损伤可能会引入到接收机中的RF解调器中。接收机内再生的TS时钟以及去复用器/解码器所使用的TS时钟可能包含有基本高频变化，它会对PCR的恢复带来直接影响。在这种情形下，解码器在复原PCR中所发现的PCR到达时间误差可能是±500ns容限的许多倍，给PCR精度带来不利。任何解码器均靠接收PCR以准确再生其节目时钟并提供稳定的视频输出。与此相同，任何再复用器均靠接收PCR以在其输出端精确地重新作出PCR标记。PCR的到达时间是严格的；因此，测量包含有到达时间误差和精度误差的总抖动（PCR\_OJ）就是非常必要的。在理想运行系统中，传输损伤应当是可以设计的。然而在目前典型的分配网络中，包含有再复用器、ATM 层等各级链路，它们均会引入误差。PCR\_AC是在复用器输出端直接测量的数值，不过它在网络中的运行监视常会使人迷惑不解，因为按照定义，它应当消除一切传输损伤。

㈣PCR频率偏置(PCR\_FO)：即PCR中所含27MHz 时钟的频率偏置，测量时以某一稳定的外部基准作为参考

备注：

1. SCR

SCR是存在于PS中的，即PS的pack里面的一个field。他用来指定这个PS的pack期望的到达decoder的时间。

2. ESCR

ESCR是位于PES里面的，即PES的header里面可能会有，当PES的packet要表示一个PES stream的时候，PES packet里面的ESCR就表示这个PES packet的所期望的到达decoder的时间。

3．PCR

PCR是TS里面的，即TS packet的header里面可能会有，他用来指定所期望的该ts packet到达decoder的时间，他的作用于SCR类似。

4．DTS, PTS

对于一个ES来说，比如视频，他又许多I,P,B帧，而P, B帧都是以I，P帧作为参考。由于B帧是前向后向参考，因此要对B帧作decode的话，就必须先decode该B帧后面的P,或者I帧，于是，decode的时间与帧的真正的present的时间就不一致了，按照DTS一次对各个帧进行decode，然后再按照PTS对各个帧进行展现。有时候PES包头里面也会有DTS，PTS，对于PTS来说，他代表了这个PES包得payload里面的第一个完整地audio access unit或者video access unit的PTS时间（并不是每个audio/video access unit都带有PTS/DTS，因此，你可以在PES里面指定一个，作为开始）。PES包头的DTS也是这个原理，只不过注意的是：对于video来说他的DTS和PTS是可以不一样的，因为B帧的存在使其顺序可以倒置。而对于audio来说，audio没有双向的预测，他的DTS和PTS可以看成是一个顺序的，因此可一直采用一个，即可只采用PTS。

#### 1.4.5 ES PES TS

MPEG2系统用于视音频同步以及系统时钟恢复的时间标签分别在ES，PES和TS这3个层次中。

在TS 层, TS头信息包含了节目时钟参考PCR(Program Clock Reference),用于恢复出与编码端一致的系统时序时钟STC(System Time Clock)。

在PES层, 在PES头信息里包含有表示时间戳PTS(Presentation Time Stamp)和解码时间戳DTS(Decoding Time Stamp);

在ES 层, 与同步有关的主要是视频缓冲验证VBV(Video Buffer Verifier),用以防止解码器的缓冲器出现上溢或者下溢;

标准规定在原始音频和视频流中,PTS的间隔不能超过0.7s，出现在TS包头的PCR间隔不能超过0.1s。



从ES到PES的示意图，插入pts dts如图如上。

在插入PTS/DTS标志时，

对于B帧, 由于在B帧PTS和DTS是相等的，所以无须在B帧插入DTS(参见图1).

对于I帧和P帧, 由于经过复用后, 数据包的顺序会发生变化,显示前一定要存储于视频解码器的排序缓存器中，经过从新排序后再显示，所以一定要同时插入PTS和DTS作为从新排序的依据.

编码器

系统时钟STC:

编码器中有一个系统时钟(其频率是27MHz),

此时钟用来产生指示音视频的正确显示和解码的时间戳,

同时可用来指示在采样过程中系统时钟本身的瞬时值。

PCR(Program Clock Reference):

指示系统时钟本身的瞬时值的时间标签称为节目参考时钟标签(PCR)。

PCR的插入必须在PCR字段的最后离开复用器的那一时刻,

同时把27MHz系统时钟的采样瞬时值作为PCR字段插入到相应的PCR域。

它是放在TS包头的自适应区中传送.

27MHz的系统时钟STC经波形整理后分成两路:

PCR\_ext (9bits ), 由27MHz脉冲直接触发计数器生成扩展域.

PCR\_base(33bits), 经300分频器分频成90kHz脉冲送入一个33位计数器生成90kHz基值,用于和PTS/DTS比较，产生解码和显示所需要的同步信号.这两部分被置入PCR域，共同组成42位的PCR.

输入到T-STD解码器的第i个字节的PCR值:

PCR(i) = PCR\_base(i)\*300 + PCR\_ext(i) i: 包含program\_clock\_reference\_base域的最后一个比特的字节号.

PCR\_base(i) = ((system\_clock\_frequency \* t(i)) / 300) % 2^33

PCR\_ext(i) = ((system\_clock\_frequency \* t(i)) / 1 ) % 300 t(i): 字节i的编码时间.

例如:

时间"03:02:29.012"的PCR计算如下:

03:02:29.012 = ((3 \* 60) + 2) \* 60 + 29.012 = 10949.012s

PCR\_base = ((27 000 000 \* 10949.012) / 300) % 2^33 = 98 541 080

PCR\_ext = ((27 000 000 \* 10949.012) / 1 ) % 300 = 0

PCR = 98 541 080 \* 300 + 0 = 295 623 324 000

PCR-base的作用:

a. 与PTS和DTS作比较, 当二者相同时, 相应的单元被显示或者解码.

b. 在解码器切换节目时,提供对解码器PCR计数器的初始值,

以让该PCR值与PTS、DTS最大可能地达到相同的时间起点.

PCR-ext的作用:

通过解码器端的锁相环路修正解码器的系统时钟, 使其达到和编码器一致的27MHz.

PTS(Presentation Time Stamp):  
  指示音视频显示时间的时间戳称为显示时间戳(PTS);  
  
  PTS域为33bits, 是对系统时钟的300分频的时钟的计数值.  
  它被编码成为3个独立的字段:  
     PTS[32..30][29..15][14..0]  
  表示此分组中第一个访问单元在系统目标解码器中的预定显示时间.  
  
PTS值为:  
  PTS(k) = ((system\_clock\_frequency \* TPn(k)) / 300) % 2^33  
TPn(k): 表示单元Pn(k)的表示时间.  
  
DTS(Decoding Time Stamp):  
  指示音视频的解码时间戳称为解码时间戳(DTS),  
    
  DTS域为33bits,编码成为3个独立的字段:  
     DTS[32..30][29..15][14..0]  
  表示此分组中第一个访问单元在系统目标解码器中的预定解码时间.  
  
DTS值为:  
  DTS(j) = ((system\_clock\_frequency \* TDn(j)) / 300) % 2^33  
TDn(j): 第n个ES流的第j个存取单元An(j)的解码时间.  
     
  DTS就视频来说,因为视频编码的时候用到了双向预测,  
  一个图像单元被解出,并非马上就被显示,可能在存储器中留一段时间,作为其余图像单元的解码参考,  
  在被参考完毕后,才被显示.  
  
音频PTS:  
  针对音频和视频的同步显示，MPEG提出了一个音频PTS.  
  由于声音没有用到双向预测，它的解码次序就是它的显示次序，故它只有PTS.

VBV\_delay:  
  视频流延时值,  
  在解码时利用视频流缓冲区把视频流缓存到相应的vbv\_delay时间后,  
  再启动解码器解码、显示、实现音视频的同步.  
  VBV\_delay存在于视频ES的头部，长度为16bit.  
  
解码器  
首先,解析PCR, 重建和编码器同步的27MHz系统时钟, 恢复27MHz系统时钟后;再通过VBV\_delay(视频流延时值)的数值来确定解码的开始;之后,利用PES流中解码时间戳(DTS)和显示时间戳(PTS)来确定解码和显示的次序.用PCR来对系统时钟进行修正.  
解码器同步算法总结如下:  
(1).解码器从输入码流的包头中解出时间信息PCR送入到系统时间时钟恢复电路;系统时间时钟恢复电路在接收到每一个新的PCR时，进行本地系统时间时钟恢复和锁相。  
(2).解复用器后，从PES包头中解出显示时间标签PTS和解码时间标签DTS，并送入到基本流解码器中。  
(3).基本流解码器在接收到新的PTS/DTS后，存入对应的FIFO(先进先处存储器)中进行管理；对于没有PTS/DTS的显示单元，需要对其时间标签进行插值，并送入到FIFO中管理。  
(4).每一显示单元开始解码前，用其对应的DTS与STC进行比较，当STC与DTS相等时开始解码；  
(5).每一显示单元开始显示前，用其对应的PTS与STC进行比较，当STC与PTS相等时开始显示。

27 MHz系统时钟经过300分频后，得到本地的33 bits PCR\_Base, 该时钟与寄存器中当前图像的PTS/DTS进行比较,  
系统软件根据比较结果做出相应的处理:  
(1).  若当前的PTS/DTS比PCR计数器的值小于半帧以上，即PTS\_Base≤-ΔPTS/2,  
      此时说明系统解码过慢，解码器处于失步状态，应根据该帧的结构做出相应的同步调整；  
(2).  若当前的PTS/DTS比PCR计数器的值在半帧时间以内，  
      我们认为此时系统解码正常，立即显示/解码当前帧；  
(3).  若当前的PTS/DTS大于PCR计数器的值，则此时解码器稍快，  
      在这种情况下，只需等到PCR与PTS/DTS相等时，就可显示/解码。

#### 1.4.4 参考链接

1.Mpeg-2的同步及时间恢复--STC，PCR，DTS，PTS

<https://www.bbsmax.com/A/x9J2omDW56/>

2.PTS PCR DTS 详解

https://www.cnblogs.com/jiangzhaowei/p/4317755.html

## H.264

### 2.1 简介

#### 2.1.1 H264和MPEG-4标准

视频编解码技术有两套标准，国际电联（ITU-T）的标准H.261、H.263、H.263+等；还有ISO 的MPEG标准Mpeg1、Mpeg2、Mpeg4等等。H.264/AVC是两大组织集合H.263+和Mpeg4的优点联合推出的最新标准，最具价值的部分无疑是更高的数据压缩比。在同等的图像质量条件下，H.264的数据压缩比能比H.263高2倍，比MPEG-4高1.5倍。H.264/AVC 是ITU-T 视频编码专家组（VCEG）和ISO/IEC 动态图像专家组（MPEG ）联合组成的联合视频组（JVT）共同努力制订的新一代视频编码标准，它最大的优势是具有很高的数据压缩比率。

H.264/AVC标准是由ITU-T和ISO/IEC联合开发的，定位于覆盖整个视频应用领域，包括：低码率的无线应用、标准清晰度和高清晰度的电视广播应用、Internet上的视频流应用，传输高清晰度的DVD视频以及应用于数码相机的高质量视频应用等等。

ITU-T给这个标准命名为H.264（以前叫做H.26L），而ISO/IEC称它为MPEG-4 高级视频编码（Advanced Video Coding，AVC）,并且它将成为MPEG-4标准的第10部分。既然AVC是当前MPEG-4标准的拓展，那么它必然将受益于MPEG-4开发良好的基础结构（比如系统分层和音频等）。

H.264最具价值的部分无疑是更高的数据压缩比。在同等的图像质量条件下，H.264的数据压缩比能比当前DVD系统中使用的MPEG-2高2-3倍，比MPEG-4高1.5-2倍。在网络传输过程中所需要的带宽更少，也更加经济。在MPEG-4需要6Mbps的传输速率匹配时，H.264只需要3Mbps-4Mbps的传输速率。

H.264获得优越性能的代价是计算复杂度的大幅增加，例如分层设计、多帧参论、多模式运动估计、改进的帧内预测等，这些都显著提高了预测精度，从而获得比其他标准好得多的压缩性能。

H.264 与MPEG-4的比较：在极低码率（32-128Kbps）的情况下，H.264与MPEG-4相比具有性能倍增效应，即：相同码率的H.26L媒体流和MPEG-4媒体流相比，H.26L拥有大约3个分贝的增益（画质水平倍增）。 32Kbps的H.26L媒体流，其信躁比与128K的MPEG-4媒体流相近。即在同样的画面质量下，H.264的码率仅仅为MPEG-4的四分之一。

#### 2.1.2 H.264组成

H.264 是一次概念的革新，它打破常规，完全没有 I 帧、P帧、B 帧的概念，也没有 IDR帧的概念。对于 H.264中出现的一些概念从大到小排序依次是：序列、图像、片组、片、NALU、宏块、亚宏块、块、像素。这里有几点值得说明：

（1）、在 H.264协议中图像是个集合概念，顶场、底场、帧都可以称为图像（本文图像概念时都是集合概念）。因此我们可以知道，对于H.264 协议来说，我们平常所熟悉的那些称呼，例如：I 帧、P 帧、B帧等等，实际上都是我们把图像这个概念具体化和细小化了。我们在 H.264里提到的“帧”通常就是指不分场的图像；

（2）、如果不采用FMO（灵活宏块排序） 机制，则一幅图像只有一个片组；

（3）、如果不使用多个片，则一个片组只有一个片；

（4）、如果不采用DP（数据分割）机制，则一个片就是一个NALU，一个 NALU 也就是一个片。

否则，一个片由 三个 NALU 组成（即标准“表7-1”中 nal\_unit\_type 值为2、3、4 的三个 NALU 属于 一个片）；

一幅图像根据组成它的片类型来分，可以分为标准“表7-5”中的 8种类型。我们平常应用中所最常见到的其实是这些类型的特例。例如：我们平常所谓的“I帧”和“IDR 帧”，其实是 primary\_pic\_type 值为 0的图像，我们平常所谓的“P帧”其实是 primary\_pic\_type 值为 1的图像的特例，我们平常所谓的“B帧”其实是 primary\_pic\_type 值为 2的图像的特例。



一幅图像根据概念来分可以分为两种：IDR 图像和非 IDR图像。一幅图像是否是 IDR 图像是由组成该图像的 NALU决定的，如果组成该图像的 NALU 为标准“表7-1”中 nal\_unit\_type 值为 5 的NALU，则该图像为 IDR 图像，否则为非 IDR图像。这里也有几点值得说明：

（1）、nal\_unit\_type 值为 5 的 NALU 只会出现在 IDR 图像中，而 IDR图像中的所有 NALU 都是nal\_unit\_type 值为 5 的NALU；

（2）、我们以组成一幅图像的片的类型来区分该图像是否是IDR 图像是错误的。

例如：一幅图像中的所有片都是 I 片并不代表这幅图像就是 IDR 图像。因为 I片也可以从属于 nal\_unit\_type 值为 1 的 NALU 也即非IDR图像有可能全部包含I片。只不过我们平常最常见到的形式是：所有片都是 I片的图像就是 IDR 图像。其实这个时候 IDR 图像的概念也被我们具体化和细小化了。但IDR图像必定全部包含I片或SI片，不过只有用NALU的类型才能判断是不是IDR图像

一幅图像由 1～N个片组组成，而每一个片组又由一个或若干个片组成一个片由一个NALU或三个NALU（假如有数据分割）组成。图像解码过程中总是按照片进行解码，然后按照片组将解码宏块重组成图像。从这种意义上讲，片实际是最大的解码单元。

1 frame的数据可以分为多个slice。每个slice中的数据，在帧内预测只用到自己slice的数据，与其他slice数据没有依赖关系。 NAL 是用来将编码的数据进行大包的。 比如，每一个slice 数据可以放在NAL 包中。I frame 是自己独立编码，不依赖于其他frame 数据。P frame 依赖 I frame 数据。B frame 依赖 I frame, P frame 或其他 B frame 数据。

Slice是片的意思，264中把图像分成一帧（frame）或两场（field），而帧又可以分成一个或几个片（Slilce）；片由宏块（MB）组成。宏块是编码处理的基本单元。

**场和帧** ：视频的一场或一帧可用来产生一个编码图像。在电视中，为减少大面积闪烁现象，把一帧分成两个隔行的场。

**片**： 每个图象中，若干宏块被排列成片的形式。片分为I片、B片、P片和其他一些片。

I片只包含I宏块，P片可包含P和I宏块，而B片可包含B和I宏块。片的主要作用是用作宏块（Macroblock）的载体（ps：下面会介绍到宏块的概念）。片之所以被创造出来，主要目的是为限制误码的扩散和传输。 如何限制误码的扩散和传输？ 每个片（slice）都应该是互相独立被传输的，某片的预测（片（slice）内预测和片（slice）间预测）不能以其它片中的宏块（Macroblock）为参考图像。

I宏块利用从当前片中已解码的像素作为参考进行帧内预测。

P宏块利用前面已编码图象作为参考图象进行帧内预测。

B宏块则利用双向的参考图象（前一帧和后一帧）进行帧内预测。

片的目的是为了限制误码的扩散和传输，使编码片相互间是独立的。

某片的预测不能以其它片中的宏块为参考图像，这样某一片中的预测误差才不会传播到其它片中去。

对于切片（slice）来讲，分为以下几种类型：

I片：只包 I宏块，I 宏块利用从当前片中已解码的像素作为参考进行帧内预测(不能取其它片中的已解码像素作为参考进行帧内预测)。

P片：可包 P和I宏块，P 宏块利用前面已编码图象作为参考图象进行帧内预测，一个帧内编码的宏块可进一步作宏块的分割:即 16×16、16×8、8×16 或 8×8 亮度像素块(以及附带的彩色像素);如果选了 8×8 的子宏块，则可再分成各种子宏块的分割，其尺寸为 8×8、8×4、4×8 或 4×4 亮度像素块(以及附带的彩色像素)。

B片：可包 B和I宏块，B 宏块则利用双向的参考图象(当前和 来的已编码图象帧)进行帧内预测。

SP片(切换P)：用于不同编码流之间的切换，包含 P 和/或 I 宏块

SI片：扩展档次中必须具有的切换，它包 了一种特殊类型的编码宏块，叫做 SI 宏块，SI 也是扩展档次中的必备功能。

**宏块** ：一个编码图像通常划分成若干宏块组成，一个宏块由一个16×16亮度像素和附加的一个8×8 Cb和一个8×8 Cr彩色像素块组成。

数据之间的关系：H264结构中，一个视频图像编码后的数据叫做一帧，一帧由一个片（slice）或多个片组成，一个片由一个或多个宏块（MB）组成，一个宏块由16x16的yuv数据组成。宏块作为H264编码的基本单位。

H264编码过程中的三种不同的数据形式：

SODB：数据比特串 —-＞最原始的编码数据，即VCL数据；

RBSP：原始字节序列载荷 —-＞在SODB的后面填加了结尾比特（RBSP trailing bits　一个bit“1”）若干比特“0”,以便字节对齐；

EBSP：扩展字节序列载荷 —- > 在RBSP基础上填加了仿校验字节（0X03）它的原因是：在NALU加到Annexb上时，需要添加每组NALU之前的开始码StartCodePrefix,如果该NALU对应的slice为一帧的开始则用4位字节表示，ox00000001,否则用3位字节表示ox000001（是一帧的一部分）。另外，为了使NALU主体中不包括与开始码相冲突的，在编码时，每遇到两个字节连续为0，就插入一个字节的0x03。解码时将0x03去掉。也称为脱壳操作。

H.264的主要目标是：1．高的视频压缩比；2．良好的网络亲和性；

为了完成这些目标H264的解决方案是：

1. VCL video coding layer 视频编码层；
2. NAL network abstraction layer 网络提取层；

其实vcl就是编码成sodb成rbsp送给nal层，nal就是进行nalu包装。

H.264/AVC核心技术概览

这个新的标准是由下面几个处理步骤组成的：1）帧间和帧内预测；2）变换（和反变换）；3）量化（和反量化）；4）环路滤波；5）熵编码。

单张的图片流组成了视频，它能分成16X16像素的“宏块”，这种分块方法简化了在视频压缩算法中每个步骤的处理过程。举例来说，从标准清晰度标准视频流解决方案（720X480）中截取的一幅图片被分成1350（45X30）个宏块，然后在宏块的层次进行进一步的处理。

#### 2.1.3 H.264打包

IP网络的RTP打包封装分组打包的规则：

(1)额外开销要少，使MTU尺寸在100～64k字节范围都可以；

(2)不用对分组内的数据解码就可以判别该分组的重要性；

(3)载荷规范应当保证不用解码就可识别由于其他的比特丢失而造成的分组不可解码；

(4)支持将NALU分割成多个RTP分组；

(5)支持将多个NALU汇集在一个RTP分组中。

RTP的头标可以是NALU的头标，并可以实现以上的打包规则。

一个RTP分组里放入一个NALU，将NALU(包括同时作为载荷头标的NALU头)放入RTP的载荷中，设置RTP头标值。为了避免IP层对大分组的再 一次分割，片分组的大小一般都要小于MTU尺寸。由于包传送的路径不同，解码端要重新对片分组排序，RTP包含的次序信息可以用来解决这一问题。

NALU分割

对于预先已经编码的内容，NALU可能大于MTU尺寸的限制。虽然IP层的分割可以使数据块小于64千字节，但无法在应用层实现保护，从而降低了非等重保 护方案的效果。由于UDP数据包小于64千字节，而且一个片的长度对某些应用场合来说太小，所以应用层打包是RTP打包方案的一部分。

新的讨论方案(IETF)应当符合以下特征：

(1)NALU的分块以按RTP次序号升序传输；

(2)能够标记第一个和最后一个NALU分块；

(3)可以检测丢失的分块。

NALU合并

一些NALU如SEI、参数集等非常小，将它们合并在一起有利于减少头标开销。已有两种集合分组：

(1)单一时间集合分组(STAP)，按时间戳进行组合；

(2)多时间集合分组(MTAP)，不同时间戳也可以组合。

#### 2.1.4 H264压缩

H264的基本原理其实非常简单，下我们就简单的描述一下H264压缩数据的过程。通过摄像头采集到的视频帧（按每秒 30 帧算），被送到 H264 编码器的缓冲区中。编码器先要为每一幅图片划分宏块。

以下面这张图为例:



划分宏块

H264默认是使用 16X16 大小的区域作为一个宏块，也可以划分成 8X8 大小。



划分好宏块后，计算宏块的象素值。



以此类推，计算一幅图像中每个宏块的像素值，所有宏块都处理完后如下面的样子。



划分子块

H264对比较平坦的图像使用 16X16 大小的宏块。但为了更高的压缩率，还可以在 16X16 的宏块上更划分出更小的子块。子块的大小可以是 8X16､ 16X8､ 8X8､ 4X8､ 8X4､ 4X4非常的灵活。



上幅图中，红框内的 16X16 宏块中大部分是蓝色背景，而三只鹰的部分图像被划在了该宏块内，为了更好的处理三只鹰的部分图像，H264就在 16X16 的宏块内又划分出了多个子块。

这样再经过帧内压缩，可以得到更高效的数据。下图是分别使用mpeg-2和H264对上面宏块进行压缩后的结果。其中左半部分为MPEG-2子块划分后压缩的结果，右半部分为H264的子块划压缩后的结果，可以看出H264的划分方法更具优势。



宏块划分好后，就可以对H264编码器缓存中的所有图片进行分组了。

帧分组

对于视频数据主要有两类数据冗余，一类是时间上的数据冗余，另一类是空间上的数据冗余。其中时间上的数据冗余是最大的。

对于这些关联特别密切的帧，其实我们只需要保存一帧的数据，其它帧都可以通过这一帧再按某种规则预测出来，所以说视频数据在时间上的冗余是最多的。

为了达到相关帧通过预测的方法来压缩数据，就需要将视频帧进行分组。那么如何判定某些帧关系密切，可以划为一组呢？我们来看一下例子，下面是捕获的一组运动的台球的视频帧，台球从右上角滚到了左下角。

H264编码器会按顺序，每次取出两幅相邻的帧进行宏块比较，计算两帧的相似度。如下图：

通过宏块扫描与宏块搜索可以发现这两个帧的关联度是非常高的。进而发现这一组帧的关联度都是非常高的。因此，上面这几帧就可以划分为一组。其算法是：在相邻几幅图像画面中，一般有差别的像素只有10%以内的点,亮度差值变化不超过2%，而色度差值的变化只有1%以内，我们认为这样的图可以分到一组。

在这样一组帧中，经过编码后，我们只保留第一帖的完整数据，其它帧都通过参考上一帧计算出来。我们称第一帧为IDR／I帧，其它帧我们称为P／B帧，这样编码后的数据帧组我们称为GOP。

运动估计与补偿

在H264编码器中将帧分组后，就要计算帧组内物体的运动矢量了。还以上面运动的台球视频帧为例，我们来看一下它是如何计算运动矢量的。

H264编码器首先按顺序从缓冲区头部取出两帧视频数据，然后进行宏块扫描。当发现其中一幅图片中有物体时，就在另一幅图的邻近位置（搜索窗口中）进行搜索。如果此时在另一幅图中找到该物体，那么就可以计算出物体的运动矢量了。下面这幅图就是搜索后的台球移动的位置。通过上图中台球位置相差，就可以计算出台图运行的方向和距离。H264依次把每一帧中球移动的距离和方向都记录下来就成了下面的样子。我们把运动矢量与补偿称为帧间压缩技术，它解决的是视频帧在时间上的数据冗余。除了帧间压缩，帧内也要进行数据压缩，帧内数据压缩解决的是空间上的数据冗余。下面我们就来介绍一下帧内压缩技术。

帧内预测

人眼对图象都有一个识别度，对低频的亮度很敏感，对高频的亮度不太敏感。所以基于一些研究，可以将一幅图像中人眼不敏感的数据去除掉。这样就提出了帧内预测技术。

H264的帧内压缩与JPEG很相似。一幅图像被划分好宏块后，对每个宏块可以进行 9 种模式的预测。找出与原图最接近的一种预测模式。



下面这幅图是对整幅图中的每个宏块进行预测的过程。



然后，将原始图像与帧内预测后的图像相减得残差值。再将我们之前得到的预测模式信息一起保存起来，这样我们就可以在解码时恢复原图了。经过帧内与帧间的压缩后，虽然数据有大幅减少，但还有优化的空间。

对残差数据做DCT

可以将残差数据做整数离散余弦变换，去掉数据的相关性，进一步压缩数据。如下图所示，左侧为原数据的宏块，右侧为计算出的残差数据的宏块。将残差数据宏块数字化后如下图所示：



将残差数据宏块进行 DCT 转换。去掉相关联的数据后，我们可以看出数据被进一步压缩了。做完 DCT 后，还不够，还要进行 CABAC 进行无损压缩。

CABAC

上面的帧内压缩是属于有损压缩技术。也就是说图像被压缩后，无法完全复原。而CABAC属于无损压缩技术。

无损压缩技术大家最熟悉的可能就是哈夫曼编码了，给高频的词一个短码，给低频词一个长码从而达到数据压缩的目的。MPEG-2中使用的VLC就是这种算法，我们以 A-Z 作为例子，A属于高频数据，Z属于低频数据。看看它是如何做的。

CABAC也是给高频数据短码，给低频数据长码。同时还会根据上下文相关性进行压缩，这种方式又比VLC高效很多。其效果如下：



现在将 A-Z 换成视频帧，它就成了下面的样子。



从上面这张图中明显可以看出采用 CACBA 的无损压缩方案要比 VLC 高效的多。

H264编码的基本原理，包括：

宏块的划分

图像分组

帧内压缩技术原理

帧间压缩技术原理。

DCT

CABAC压缩原理。

#### 2.1.5 H.264序列的说明

在H264中图像以序列为单位进行组织，一个序列是一段图像编码后的数据流，以 I 帧开始，到下一个 I 帧结束。一个序列的第一个图像叫做 IDR 图像（立即刷新图像），IDR 图像都是 I 帧图像。H.264 引入 IDR 图像是为了解码的重同步，当解码器解码到 IDR 图像时，立即将参考帧队列清空，将已解码的数据全部输出或抛弃，重新查找参数集，开始一个新的序列。这样，如果前一个序列出现重大错误，在这里可以获得重新同步的机会。IDR 图像之后的图像永远不会使用 IDR 之前的图像的数据来解码。

I 帧：帧内编码帧 ，I 帧表示关键帧，你可以理解为这一帧画面的完整保留；解码时只需要本帧数据就可以完成（因为包含完整画面）。

P 帧：前向预测编码帧。P 帧表示的是这一帧跟之前的一个关键帧（或P 帧）的差别，解码时需要用之前缓存的画面叠加上本帧定义的差别，生成最终画面。（也就是差别帧，P 帧没有完整画面数据，只有与前一帧的画面差别的数据）。P帧的预测与重构：P 帧是以 I 帧为参考帧,在 I 帧中找出 P 帧“某点”的预测值和运动矢量，取预测差值和运动矢量一起传送。在接收端根据运动矢量从 I 帧中找出 P 帧“某点”的预测值并与差值相加以得到 P 帧“某点”样值,从而可得到完整的 P 帧。P 帧属于前向预测的帧间编码。它只参考前面最靠近它的 I 帧或 P 帧;P 帧可以是其后面 P 帧的参考帧,也可以是其前后的 B 帧的参考帧;由于 P 帧是参考帧,它可能造成解码错误的扩散;由于是差值传送，P 帧的压缩比较高。

B 帧：双向预测内插编码帧。B 帧是双向差别帧，也就是 B 帧记录的是本帧与前后帧的差别（具体比较复杂，有 4 种情况，但我这样说简单些），换言之，要解码 B 帧，不仅要取得之前的缓存画面，还要解码之后的画面，通过前后画面的与本帧数据的叠加取得最终的画面。B 帧压缩率高，但是解码时 CPU 会比较累。B 帧的预测与重构：B 帧以前面的 I 或 P 帧和后面的 P 帧为参考帧,“找出”B 帧“某点”的预测值和两个运动矢量,并取预测差值和运动矢量传送。接收端根据运动矢量在两个参考帧中“找出(算出)”预测值并与差值求和,得到 B帧“某点”样值,从而可得到完整的 B 帧。

举例说明：



在如上图中，GOP (Group of Pictures) 长度为 13，S0~S7 表示 8 个视点，T0~T12 为 GOP 的 13 个时刻。每个 GOP 包含帧数为视点数 GOP 长度的乘积。在该图中一个 GOP 中，包含 94 个 B 帧。B 帧占一个 GOP 总帧数的 90.38%。GOP 越长，B 帧所占比例更高，编码的率失真性能越高。下图测试序列 Race1 在不同 GOP 下的率失真性能对比。

#### 2.1.6 H.264 压缩方法

（1）分组：把几帧图像分为一组(GOP，也就是一个序列)，为防止运动变化,帧数不宜取多。

（2）定义帧：将每组内各帧图像定义为三种类型,即 I 帧、B 帧和 P 帧;

（3）预测帧：以I帧做为基础帧，以 I 帧预测 P 帧,再由 I 帧和 P 帧预测 B 帧;

（4）数据传输：最后将 I 帧数据与预测的差值信息进行存储和传输。

#### 2.1.7 H.264 I 帧与 IDR 帧的区别

IDR（Instantaneous Decoding Refresh）--即时解码刷新。

    I 和 IDR 帧都是使用帧内预测的。它们都是同一个东西而已，在编码和解码中为了方便，要首个 I 帧和其他 I 帧区别开，所以才把第一个首个 I 帧叫 IDR，这样就方便控制编码和解码流程。IDR 帧的作用是立刻刷新，使错误不致传播,从IDR帧开始，重新算一个新的序列开始编码。而 I 帧不具有随机访问的能力，这个功能是由 IDR 承担。IDR 会导致DPB（参考帧列表——这是关键所在）清空，而 I 不会。IDR 图像一定是 I 图像，但I图像不一定是 IDR 图像。一个序列中可以有很多的I图像，I 图像之后的图像可以引用 I 图像之间的图像做运动参考。一个序列中可以有很多的 I 图像，I 图像之后的图象可以引用I图像之间的图像做运动参考。

对于 IDR 帧来说，在 IDR 帧之后的所有帧都不能引用任何 IDR 帧之前的帧的内容，与此相反，对于普通的 I 帧来说，位于其之后的 B- 和 P- 帧可以引用位于普通 I- 帧之前的 I- 帧。从随机存取的视频流中，播放器永远可以从一个 IDR 帧播放，因为在它之后没有任何帧引用之前的帧。但是，不能在一个没有 IDR 帧的视频中从任意点开始播放，因为后面的帧总是会引用前面的帧。

举个例子，在一段视频中，存在以下帧：I P B P B P B B P I P B…

如果这段视频应用了多重参照帧，那么蓝色的 P 帧在参照他前面的 I 帧（红色）的同时，还可能会参照 I 帧之前的 P （绿色），由于 I 帧前后的场景可能会有很大的反差甚至根本不同，所以此时 P 帧参考I帧之前的帧不但会没有意义，反而会造成很多问题。所以一种新型的帧被引入，那就是 IDR 帧。如果这段视频应用了多重参考帧的同时采用了 IDR 帧，那么帧的顺序就会变成这样：I P B P B P B B P IDR P B…由于 IDR 帧禁止后面的帧向自己前面的帧参照，所以这回那个蓝色的 P 帧就不会参照绿色的 P 帧了。使用 ffmpeg 获取 I 帧 ：ffmpeg -i wkll.mp4 -ss 00:00:00 -t 1 -r 1 -q:v 2 -f image2 pic-%03d.jpeg

### 2.2 NALU H264网络传输的结构

H264在网络传输的是NALU，NALU的结构是：NAL头+RBSP



NALU头用来标识后面的RBSP是什么类型的数据，他是否会被其他帧参考以及网络传输是否有错误。

每个NAL前有一个起始码 0x00 00 01（或者0x00 00 00 01），解码器检测每个起始码，作为一个NAL的起始标识，当检测到下一个起始码时，当前NAL结束。同时H.264规定，当检测到0x000000时，也可以表征当前NAL的结束。那么NAL中数据出现0x000001或0x000000时怎么办？H.264引入了防止竞争机制，如果编码器检测到NAL数据存在0x000001或0x000000时，编码器会在最后个字节前插入一个新的字节0x03，这样：

0x000003－>0x00000303。解码器检测到0x000003时，把03抛弃，恢复原始数据（脱壳操作）。解码器在解码时，首先逐个字节读取NAL的数据，统计NAL的长度，然后再开始解码。

#### 2.2.1 NALU头结构

NALU头结构长度：1byte

forbidden\_bit(1bit) + nal\_reference\_bit(2bit) + nal\_unit\_type(5bit)

1.forbidden\_bit：禁止位，初始为0，当网络发现NAL单元有比特错误时可设置该比特为1，以便接收方纠错或丢掉该单元。

2.nal\_reference\_bit：nal重要性指示，标志该NAL单元的重要性，值越大，越重要，解码器在解码处理不过来的时候，可以丢掉重要性为0的NALU。

3.nal\_unit\_type：不同类型的NALU的重要性指示如下表所示。

|  |  |  |
| --- | --- | --- |
| nal\_unit\_type | NAL类型 | nal\_reference\_bit |
| 0 | 未使用 | 0 |
| 1 | 非IDR的片 | 此片属于参考帧，则不等于0，  不属于参考帧，则等与0 |
| 2 | 片数据A分区 | 同上 |
| 3 | 片数据B分区 | 同上 |
| 4 | 片数据C分区 | 同上 |
| 5 | IDR图像的片 | 5 |
| 6 | 补充增强信息单元（SEI） | 0 |
| 7 | 序列参数集sps | 非0 |
| 8 | 图像参数集pps | 非0 |
| 9 | 分界符 | 0 |
| 10 | 序列结束 | 0 |
| 11 | 码流结束 | 0 |
| 12 | 填充 | 0 |
| 13..23 | 保留 | 0 |
| 24..31 | 不保留（RTP打包时会用到） | 0 |

RTP 打包时的扩展类型

|  |  |  |
| --- | --- | --- |
| 24 | STAP-A | Single-time aggregation packet |
| 25 | STAP-B | Single-time aggregation packet |
| 26 | MTAP16 | Multi-time aggregation packet |
| 27 | MTAP24 | Multi-time aggregation packet |
| 28 | FU-A | Fragmentation unit |
| 29 | FU-B | Fragmentation unit |
| 30-31 | undefined |  |

#### 2.2.2 RBSP数据

RBSP数据是下表中的一种，rbsp是根据nalu头类型决定。

|  |  |  |
| --- | --- | --- |
| RBSP类型 | 所写 | 描述 |
| 参数集 | PS | 序列的全局信息，如图像尺寸，视频格式等 |
| 增强信息 | SEI | 视频序列解码的增强信息 |
| 图像界定符 | PD | 视频图像的边界 |
| 编码片 | SLICE | 编码片的头信息和数据 |
| 数据分割 |  | DP片层的数据，用于错误恢复解码 |
| 序列结束符 |  | 表明一个序列的结束，下一个图像为IDR图像 |
| 流结束符 |  | 表明该流中已没有图像 |
| 填充数据 |  | 亚元数据，用于填充字节 |

要完成视频的解码，不仅需要传输VCL层编码出来的视频帧数据，还需要传输序列参数集、图像参数集等数据。

参数集:包括序列参数集 SPS 和图像参数集 PPS

**SPS：**包含的是针对一连续编码视频序列的参数，如标识符 seq\_parameter\_set\_id、帧数及 POC 的约束、参考帧数目、解码图像尺寸和帧场编码模式选择标识等等。

**PPS：**对应的是序列中某一幅图像或者某几幅图像，其参数如标识符pic\_parameter\_set\_id、可选的 seq\_parameter\_set\_id、熵编码模式选择标识、片组数目、初始量化参数和去方块滤波系数调整标识等等。

**SEi:**自定义信息，

数据分割：组成片的编码数据存放在 3 个独立的 DP（数据分割，A、B、C）中，各自包含一个编码片的子集。

分割Ａ包含片头和片中每个宏块头数据。

分割Ｂ包含帧内和 SI 片宏块的编码残差数据。

分割 C包含帧间宏块的编码残差数据。

每个分割可放在独立的 NAL 单元并独立传输。

**AUD**：一般文档没有对AUD进行描叙，其实这是一个帧开始的标志，字节顺序为：00 00 00 01 09 f0从结构上看，有start code, 所以确实是一个NALU，类型09在H264定义里就是AUD（分割器）。大部分播放器可以在没有AUD的情况下正常播放。紧随AUD，一般是SPS/PPS/SEI/IDR的组合或者简单就是一个SLICE，也就是一个帧的开始。像Flash这样的播放器，每次需要一个完整的帧数据，那么把2个AUD之间的数据按照格式打包给播放器就可以了。

h264有两种封装：一种是annexb模式，传统模式，有startcode，SPS和PPS是在ES中；一种是mp4模式，一般mp4 mkv会有，没有startcode，SPS和PPS以及其它信息被封装在container中，每一个frame前面是这个frame的长度。

很多解码器只支持annexb这种模式，因此需要将mp4做转换：

在ffmpeg中用h264\_mp4toannexb\_filter可以做转换

Ffmpeg中实现：

注册filter

avcbsfc = av\_bitstream\_filter\_init(“h264\_mp4toannexb”);

转换bitstream

av\_bitstream\_filter\_filter(AVBitStreamFilterContext \*bsfc,

AVCodecContext \*avctx, const char \*args,

uint8\_t \*\*poutbuf, int \*poutbuf\_size,

const uint8\_t \*buf, int buf\_size, int keyframe)

对于MP4文件，NAL单元之前没有同步码，却有若干字节的长度码，来表示NAL单元的长度，这个长度码所占用的字节数由MP4文件头给出；此外，从MP4读出来的视频帧不包含PPS和SPS，这些信息位于MP4的文件头中，解析器必须在打开文件的时候就获取它们。从MP4文件读出的一个H.264帧往往是下面的形式（假设长度码为2字节）：

00 19 06 [... 25 字节...] 24 aa 65 [... 9386 字节...]

SEI信息 IDR Slice

H264 over RTP基本上分三种类型：

(1)Single NAL unit packet 也就是实际的NAL类型，可以理解为一个包就是一帧H264数据，这个在实际中是比较多的。

(2)Aggregation packet 一包数据中含有多个H264帧。

STAP-A 包内的帧含有相同的NALU-Time，没有DON

STAP-B 包内的帧含有相同的NALU-Time，有DON

MTAP16 包内的帧含有不同的NALU-Time，timestamp offset = 16

MTAP24 包内的帧含有不同的NALU-Time，timestamp offset = 24

封装在Aggregation packet中的 NAL单元大小为65535字节

1. Fragmentation unit 一帧数据被分为多个RTP包，这也是很常见的，特别是对于关键帧。现存两个版本FU-A，FU-B。

h264包在传输的时候，如果包太大，会被分成多个片。NALU头会被如下的2个自己代替。

The FU indicator octet has the following format:

      +---------------+

      |0|1|2|3|4|5|6|7|

      +-+-+-+-+-+-+-+-+

      |F|NRI|  Type   |

      +---------------+

   别被名字吓到这个格式就是上面提到的RTP h264负载类型，Type为FU-A

下面这个应该就是rtp中h264格式The FU header has the following format:

      +---------------+

      |0|1|2|3|4|5|6|7|

      +-+-+-+-+-+-+-+-+

      |S|E|R|  Type   |

      +---------------+

   S bit为1表示分片的NAL开始，当它为1时，E不能为1

   E bit为1表示结束，当它为1，S不能为1

   R bit保留位

   Type就是NALU头中的Type,取1-23的那个值

#### 2.2.3 SPS语法元素及其含义

SPS即Sequence Paramater Set，又称作序列参数集。类型7表示该NAL Unit内保存的数据为Sequence ParamaterSet。丢失或出现错误，那么解码过程很可能会失败。SPS及后续将要讲述的图像参数集PPS在某些平台的视频处理框架（比如iOS的VideoToolBox等）还通常作为解码器实例的初始化信息使用。SPS中保存了一组编码视频序列(Coded

videosequence)的全局参数。所谓的编码视频序列即原始视频的一帧一帧的像素数据经过编码之后的结构组成的序列。而每一帧的编码后数据所依赖的参数保存于图像参数集中。一般情况SPS和PPS的NALUnit通常位于整个码流的起始位置。但在某些特殊情况下，在码流中间也可能出现这两种结构，主要原因可能为：

解码器需要在码流中间开始解码；

编码器在编码的过程中改变了码流的参数（如图像分辨率等）；

在做视频播放器时，为了让后续的解码过程可以使用SPS中包含的参数，必须对其中的数据进行解析。其中H.264标准协议中规定的SPS格式位于文档的7.3.2.1.1部分，如下图所示：



其中的每一个语法元素及其含义如下：

**(1) profile\_idc**：

标识当前H.264码流的profile。我们知道，H.264中定义了三种常用的档次profile：

基准档次：baseline profile;

主要档次：main profile;

扩展档次：extended profile;

在H.264的SPS中，第一个字节表示profile\_idc，根据profile\_idc的值可以确定码流符合哪一种档次。判断规律为：

profile\_idc = 66 → baseline profile;

profile\_idc = 77 → main profile;

profile\_idc = 88 → extended profile;

在新版的标准中，还包括了High、High 10、High 4:2:2、High 4:4:4、High 10 Intra、High

4:2:2 Intra、High 4:4:4 Intra、CAVLC 4:4:4 Intra等，每一种都由不同的profile\_idc表示。

另外，constraint\_set0\_flag ~ constraint\_set5\_flag是在编码的档次方面对码流增加的其他一些额外限制性条件。

在我们实验码流中，profile\_idc = 0x42 = 66，因此码流的档次为baseline profile。

**(2) level\_idc**

标识当前码流的Level。编码的Level定义了某种条件下的最大视频分辨率、最大视频帧率等参数，码流所遵从的level由level\_idc指定。

当前码流中，level\_idc = 0x1e = 30，因此码流的级别为3。

**(3) seq\_parameter\_set\_id**

表示当前的序列参数集的id。通过该id值，图像参数集pps可以引用其代表的sps中的参数。

**(4) log2\_max\_frame\_num\_minus4**

用于计算MaxFrameNum的值。计算公式为MaxFrameNum = 2^(log2\_max\_frame\_num\_minus4 +

4)。MaxFrameNum是frame\_num的上限值，frame\_num是图像序号的一种表示方法，在帧间编码中常用作一种参考帧标记的手段。

**(5) pic\_order\_cnt\_type**

表示解码picture order count(POC)的方法。POC是另一种计量图像序号的方式，与frame\_num有着不同的计算方法。该语法元素的取值为0、1或2。

**(6) log2\_max\_pic\_order\_cnt\_lsb\_minus4**

用于计算MaxPicOrderCntLsb的值，该值表示POC的上限。计算方法为MaxPicOrderCntLsb = 2^(log2\_max\_pic\_order\_cnt\_lsb\_minus4 + 4)。

**(7) max\_num\_ref\_frames**

用于表示参考帧的最大数目。

**(8) gaps\_in\_frame\_num\_value\_allowed\_flag**

标识位，说明frame\_num中是否允许不连续的值。

**(9) pic\_width\_in\_mbs\_minus1**

用于计算图像的宽度。单位为宏块个数，因此图像的实际宽度为:

frame\_width = 16 × (pic\_width\_in\_mbs\_minus1 + 1);

**(10) pic\_height\_in\_map\_units\_minus1**

使用PicHeightInMapUnits来度量视频中一帧图像的高度。PicHeightInMapUnits并非图像明确的以像素或宏块为单位的高度，而需要考虑该宏块是帧编码或场编码。PicHeightInMapUnits的计算方式为：

PicHeightInMapUnits = pic\_height\_in\_map\_units\_minus1 + 1;

**(11) frame\_mbs\_only\_flag**

标识位，说明宏块的编码方式。当该标识位为0时，宏块可能为帧编码或场编码；该标识位为1时，所有宏块都采用帧编码。根据该标识位取值不同，PicHeightInMapUnits的含义也不同，为0时表示一场数据按宏块计算的高度，为1时表示一帧数据按宏块计算的高度。

按照宏块计算的图像实际高度FrameHeightInMbs的计算方法为：

FrameHeightInMbs = ( 2 − frame\_mbs\_only\_flag ) \* PicHeightInMapUnits

**(12) mb\_adaptive\_frame\_field\_flag**

标识位，说明是否采用了宏块级的帧场自适应编码。当该标识位为0时，不存在帧编码和场编码之间的切换；当标识位为1时，宏块可能在帧编码和场编码模式之间进行选择。

**(13) direct\_8x8\_inference\_flag**

标识位，用于B\_Skip、B\_Direct模式运动矢量的推导计算。

**(14) frame\_cropping\_flag**

标识位，说明是否需要对输出的图像帧进行裁剪。

**(15) vui\_parameters\_present\_flag**

标识位，说明SPS中是否存在VUI信息。

#### 2.2.4 PPS语法元素及其含义

除了序列参数集SPS之外，H.264中另一重要的参数集合为图像参数集Picture Paramater

Set(PPS)。通常情况下，PPS类似于SPS，在H.264的裸码流中单独保存在一个NAL Unit中，只是PPS NALUnit的nal\_unit\_type值为8；而在封装格式中，PPS通常与SPS一起，保存在视频文件的文件头中。在H.264的协议文档中，PPS的结构定义在7.3.2.2节中，具体的结构如下表所示：



其中的每一个语法元素及其含义如下：

**(1) pic\_parameter\_set\_id**

表示当前PPS的id。某个PPS在码流中会被相应的slice引用，slice引用PPS的方式就是在Slice header中保存PPS的id值。该值的取值范围为[0,255]。

**(2) seq\_parameter\_set\_id**

表示当前PPS所引用的激活的SPS的id。通过这种方式，PPS中也可以取到对应SPS中的参数。该值的取值范围为[0,31]。

**(3) entropy\_coding\_mode\_flag**

熵编码模式标识，该标识位表示码流中熵编码/解码选择的算法。对于部分语法元素，在不同的编码配置下，选择的熵编码方式不同。例如在一个宏块语法元素中，宏块类型mb\_type的语法元素描述符为“ue(v)| ae(v)”，在baseline profile等设置下采用指数哥伦布编码，在main profile等设置下采用CABAC编码。

标识位entropy\_coding\_mode\_flag的作用就是控制这种算法选择。当该值为0时，选择左边的算法，通常为指数哥伦布编码或者CAVLC；当该值为1时，选择右边的算法，通常为CABAC。

**(4) bottom\_field\_pic\_order\_in\_frame\_present\_flag**

标识位，用于表示另外条带头中的两个语法元素delta\_pic\_order\_cnt\_bottom和delta\_pic\_order\_cn是否存在的标识。这两个语法元素表示了某一帧的底场的POC的计算方法。

**(5) num\_slice\_groups\_minus1**

表示某一帧中slice group的个数。当该值为0时，一帧中所有的slice都属于一个slice group。slice group是一帧中宏块的组合方式，定义在协议文档的3.141部分。

**(6) num\_ref\_idx\_l0\_default\_active\_minus1、num\_ref\_idx\_l0\_default\_active\_minus1**

表示当Slice Header中的num\_ref\_idx\_active\_override\_flag标识位为0时，P/SP/B

slice的语法元素num\_ref\_idx\_l0\_active\_minus1和num\_ref\_idx\_l1\_active\_minus1的默认值。

**(7) weighted\_pred\_flag**

标识位，表示在P/SP slice中是否开启加权预测。

**(8) weighted\_bipred\_idc**

表示在B Slice中加权预测的方法，取值范围为[0,2]。0表示默认加权预测，1表示显式加权预测，2表示隐式加权预测。

**(9) pic\_init\_qp\_minus26和pic\_init\_qs\_minus26**

表示初始的量化参数。实际的量化参数由该参数、slice header中的slice\_qp\_delta/slice\_qs\_delta计算得到。

**(10) chroma\_qp\_index\_offset**

用于计算色度分量的量化参数，取值范围为[-12,12]。

**(11) deblocking\_filter\_control\_present\_flag**

标识位，用于表示Slice header中是否存在用于去块滤波器控制的信息。当该标志位为1时，slice header中包含去块滤波相应的信息；当该标识位为0时，slice header中没有相应的信息。

**(12) constrained\_intra\_pred\_flag**

若该标识为1，表示I宏块在进行帧内预测时只能使用来自I和SI类型宏块的信息；若该标识位0，表示I宏块可以使用来自Inter类型宏块的信息。

**(13) redundant\_pic\_cnt\_present\_flag**

标识位，用于表示Slice header中是否存在redundant\_pic\_cnt语法元素。当该标志位为1时，slice header中包含redundant\_pic\_cnt；当该标识位为0时，slice header中没有相应的信息。

#### 2.2.5 SEI 语法元素及其含义

"0x06"SEI type后一个字节为“0x05”（淡黄底色）是SEI payload type，即表征SEI payload分析遵循user\_data\_unregistered()语法。





**SEI payload size**

“0x05”后一个字节为“0x2F”（淡蓝底色）是SEI payload size，此时整个payload是47个字节。SEI payload uuid"0x2F"随后的16个字节即为uuid，此时uuid为：dc45e9bde6d948b7962cd820d923eeef

**SEI payload content**

由于payload size是47个字节，除去16字节的uuid，剩下31个字节的content。由于content是字符串，所以有结束符"0x00"，有效的30个字符内容是：

Zencoder Video Encoding System

**rbsp trailing bits**

47个payload字节后的"0x80"（灰底色）即是rbsp trailing bits，在user\_data\_unregistered()里面都是按字节写入的，所以此时的NAL unit结尾写入的字节一定是0x80。



#### 2.2.6解析SDP中包含的H.264的SPS和PPS

用RTP传输H264的时候,需要用到sdp协议描述,其中有两项:Sequence Parameter Sets (SPS)和Picture Parameter Set(PPS)需要用到,那么这两项从哪里获取呢?答案是从H264码流中获取.在H264码流中,都是以"0x00 0x00 0x01"或者"0x000x00 0x00 0x01"为开始码的,找到开始码之后,使用开始码之后的第一个字节的低5位判断是否为7(sps)或者8(pps),

及data[4] & 0x1f == 7 ||data[4] & 0x1f ==8.然后对获取的nal去掉开始码之后进行base64编码,得到的信息就可以用于sdp.sps和pps需要用逗号分隔开来.SDP中的H.264的SPS和PPS串，包含了初始化H.264解码器所需要的信息参数，包括编码所用的profile，level，图像的宽和高，deblock滤波器等。

这里需要特别提一下这两个参数pic\_width\_in\_mbs\_minus1 = 21，pic\_height\_in\_mbs\_minus1 = 17分别表示图像的宽和高，以宏块（16x16）为单位的值减1

因此，实际的宽为 (21+1)\*16 = 352 高为 (17+1)\*16 = 288

#### 2.2.7 RTP打包H.264

H264的NALU打包成RTP包的模式（下面是用到的两种模式）

（1）、一个NALU打包成一个RTP包，只需要在一个12字节的RTP包头后添加去掉开始码的NALU即可 （这种模式在一个NALU的大小小于MTU时使用）。

（2）、一个NALU打包成几个RTP包（FU\_A模式），在12个字节的RTP头后面加上一个字节的 FU indicator和一个字节的FU header。FU indicator前3位是NALU头的前3位，后5位是28（十进制）， FU header第1位标记RTP包是否为NALU的第一片，第2位标记RTP包是否为NALU的最后一片。第3位是保 留位，后5位是NALU头的type位。

可能的结构类型分别有:

1. 单一 NAL 单元模式

即一个 RTP 包仅由一个完整的 NALU 组成. 这种情况下 RTP NAL 头类型字段和原始的 H.264的 NALU 头类型字段是一样的. 对于 NALU 的长度小于 MTU 大小的包, 一般采用单一 NAL 单元模式.

对于一个原始的 H.264 NALU 单元常由 [Start Code] [NALU Header] [NALU Payload] 三部分组成, 其中 Start Code 用于标示这是一个 NALU 单元的开始, 必须是 “00 00 00 01” 或 “00 00 01”, NALU 头仅一个字节, 其后都是 NALU 单元内容.

打包时去除 “00 00 01” 或 “00 00 00 01” 的开始码, 把其他数据封包的 RTP 包即可.

如有一个 H.264 的 NALU 是这样的:

[00 00 00 01 67 42 A0 1E 23 56 0E 2F … ]

这是一个序列参数集 NAL 单元. [00 00 00 01] 是四个字节的开始码, 67 是 NALU 头, 42 开始的数据是 NALU 内容.

封装成 RTP 包将如下:

[ RTP Header ] [ 67 42 A0 1E 23 56 0E 2F ]

即只要去掉 4 个字节的开始码就可以了.

2. 组合封包模式

即可能是由多个 NAL 单元组成一个 RTP 包. 分别有4种组合方式: STAP-A, STAP-B, MTAP16, MTAP24. 那么这里的类型值分别是 24, 25, 26 以及 27.

其次, 当 NALU 的长度特别小时, 可以把几个 NALU 单元封在一个 RTP 包中.

0 1 2 3

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| RTP Header |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

|STAP-A NAL HDR | NALU 1 Size | NALU 1 HDR |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| NALU 1 Data |

: :

+ +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| | NALU 2 Size | NALU 2 HDR |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| NALU 2 Data |

: :

| +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

| :...OPTIONAL RTP padding |

+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

3. 分片封包模式

用于把一个 NALU 单元封装成多个 RTP 包. 存在两种类型 FU-A 和 FU-B. 类型值分别是 28 和 29. 而当 NALU 的长度超过 MTU 时, 就必须对 NALU 单元进行分片封包. 也称为 Fragmentation Units (FUs).

### 2.3 H.264码流结构图

在H.264 中，句法元素共被组织成 序列、图像、片、宏块、子宏块五个层次。H.264中出现的一些概念从大到小排序依次是：序列、图像、片组、片、NALU、宏块、亚宏块、块、像素。



在这样的结构中，每一层的头部和它的数据部分形成管理与被管理的强依赖关系，头部的句法元素是该层数据的核心，而一旦头部丢失，数据部分的信息几乎不可能再被正确解码出来，尤其在序列层及图像层。



在 H.264 中，分层结构最大的不同是取消了序列层和图像层，并将原本属于序列和图像头部的大部分句法元素游离出来形成序列和图像两级参数集，其余的部分则放入片层。

参数集是一个独立的数据单位，不依赖于参数集外的其他句法元素。一个参数集不对应某一个特定的图像或序列，同一序列参数集可以被多个图像参数集引用，同理，同一个图像参数集也可以被多个图像引用。只在编码器认为需要更新参数集的内容时，才会发出新的参数集。





F

#### 参考链接：

1.H264格式 详细介绍

https://blog.csdn.net/Frakie\_Kwok/article/details/77989024

2 .介绍sps pps

https://blog.csdn.net/sdsszk/article/details/83509325

3.H264基本原理

https://blog.csdn.net/garrylea/article/details/78536775

## 3. RTP

### 3.1 RTP介绍

#### 3.1.1 RTP概述

RTP全名是Real-time Transport Protocol（实时传输协议）。它是IETF提出的一个标准，对应的RFC文档为RFC3550（RFC1889为其过期版本）。RTP为Internet上端到端的实时传输提供时间信息和流同步，但并不保证服务质量，服务质量由RTCP来提供。

RTP实时传输协议（Real-time Transport Protocol，RTP）是在Internet上处理多媒体数据流的一种网络协议，利用它能够在一对一（单播）或者一对多（multicast，多播） 的网络环境中实现传流媒体数据的实时传输RTP提供抖动补偿和数据无序到达检测的机制。由于IP网络的传输特性，数据的无序到达是很常见的。 RTP允许数据通过IP组播的方式传送到多个目的地。RTP被认为是在IP网络中传输音频和视频的基本标准。RTP通常配合模板和负载格式使用。

#### 3.1.2 RTP传输介绍

RTP通常使用UDP来进行多媒体数据的传输，但如果需要的话可以使用TCP或者ATM等其它协议，整个RTP 协议由两个密切相关的部分组成：RTP数据协议和RTP控制协议，RTP标准定义了两个子协议，RTP和RTCP。RTP 使用偶数端口号接收发送数据，相应的RTCP则使用相邻的下一位奇数端口号。每一个多媒体流会建立一个RTP会话。一个会话包含带有RTP和RTCP端口号的IP地址。例如，音频和视频流使用分开的RTP会话，这样用户可以选择其中一个媒体流。形成会话的端口由其他协议（例如RTSP和SIP）来协商。RTP和RTCP使用UDP端口1024 - 65535。简单的多播音频会议。语音通信通过一个多播地址（一对多得IP地址类似于远程游戏）和一对端口来实现。一个用于音频数据（RTP），另一个用于控制包（RTCP）。

如果在一次会议中同时使用了音频和视频会议，这两种媒体将分别在不同的RTP会话中传送，每一个会话使用不同的传输地址（IP地址＋端口）。如果一个用户同时使用了两个会话，则每个会话对应的RTCP包都使用规范化名字CNAME（Canonical Name）。与会者可以根据RTCP包中的CNAME来获取相关联的音频和视频，然后根据RTCP包中的计时信息(Network time protocol)来实现音频和视频的同步。

**翻译器和混合器：**翻译器和混合器都是RTP级的中继系统。翻译器用在通过IP多播不能直接到达的用户区，例如发送者和接收者之间存在防火墙。当与会者能接收的音频编码格式不一样，比如有一个与会者通过一条低速链路接入到高速会议，这时就要使用混合器。在进入音频数据格式需要变化的网络前，混合器将来自一个源或多个源的音频包进行重构，并把重构后的多个音频合并，采用另一种音频编码进行编码后，再转发这个新的RTP包。从一个混合器出来的所有数据包要用混合器作为它们的同步源（SSRC，见RTP的封装）来识别，可以通过贡献源列表（CSRC表，见RTP的封装）可以确认谈话者。

**数据传输协议RTP**：用于实时传输数据。该协议提供的信息包括：时间戳（用于同步）、序列号（用于丢包和重排序检测）、以及负载格式（用于说明数据的编码格式）。

**控制协议RTCP**：用于QoS反馈和同步媒体流。相对于RTP来说，RTCP所占的带宽非常小，通常只有5%。

#### 3.1.3 参考链接

1.RTP协议全解析（H264码流和PS流）

https://blog.csdn.net/chen495810242/article/details/39207305

### 3.2 RTP数据格式

#### 3.2.1 RTP解析

RTP数据协议。RTP数据协议负责对流媒体数据进行封包并实现媒体流的实时传输，每一个RTP数据报都由头部（Header）和负载（Payload）两个部分组成，其中头部前12个字节的含义是固定的，而负载则可以是音频或者视频数据。

RTP数据报的头部格式如图所示：



**(1) V**：RTP协议的版本号，占2位，当前协议版本号为2

**(2) P**：填充标志，占1位，如果P=1，则在该报文的尾部填充一个或多个额外的八位组，它们不是有效载荷的一部分。

**(3) X**：扩展标志，占1位，如果X=1，则在RTP报头后跟有一个扩展报头

**(4)  CC**：CSRC计数器，占4位，指示CSRC 标识符的个数

**(5)  M**: 标记，占1位，不同的有效载荷有不同的含义，对于视频，标记一帧的结束；对于音频，标记会话的开始。

**(6)  PT**: 有效荷载类型，占7位，用于说明RTP报文中有效载荷的类型，如GSM音频、JPEM图像等,在流媒体中大部分是用来区分音频流和视频流的，这样便于客户端进行解析。

**(7) 序列号**：占16位，用于标识发送者所发送的RTP报文的序列号，每发送一个报文，序列号增1。这个字段当下层的承载协议用UDP的时候，网络状况不好的时候可以用来检查丢包。同时出现网络抖动的情况可以用来对数据进行重新排序，序列号的初始值是随机的，同时音频包和视频包的sequence是分别记数的。

**(8)时戳(Timestamp)**：占32位，必须使用90 kHz 时钟频率。时戳反映了该RTP报文的第一个八位组的采样时刻。接收者使用时戳来计算延迟和延迟抖动，并进行同步控制。

**(9)  同步信源(SSRC)标识符**：占32位，用于标识同步信源。该标识符是随机选择的，参加同一视频会议的两个同步信源不能有相同的SSRC。同步源就是指RTP包流的来源。在同一个RTP会话中不能有两个相同的SSRC值。该标识符是随机选取的 RFC1889推荐了MD5随机算法。

**(10) 特约信源(CSRC)标识符**：每个CSRC标识符占32位，可以有0～15个。每个CSRC标识了包含在该RTP报文有效载荷中的所有特约信源。CSRC标识紧跟在RTP固定头部之后，用来表示RTP数据报的来源，RTP协议允许在同一个会话中存在多个数据源，它们可以通过RTP混合器合并为一个数据源。例如，可以产生一个CSRC列表来表示一个电话会议，该会议通过一个RTP混合器将所有 讲话者的语音数据组合为一个RTP数据源。

注：基本的RTP说明并不定义任何头扩展本身，如果遇到X=1，需要特殊处理

#### 3.2.2 RTCP解析

RTP需要RTCP为其服务质量提供保证，因此下面介绍一下RTCP的相关知识。

RTCP的主要功能是：服务质量的监视与反馈、媒体间的同步，以及多播组中成员的标识。在RTP会话期 间，各参与者周期性地传送RTCP包。RTCP包中含有已发送的数据包的数量、丢失的数据包的数量等统计资料，因此，各参与者可以利用这些信息动态地改变传输速率，甚至改变有效载荷类型。RTP和RTCP配合使用，它们能以有效的反馈和最小的开销使传输效率最佳化，因而特别适合传送网上的实时数据。

RTCP也是用UDP来传送的，但RTCP封装的仅仅是一些控制信息，因而分组很短，所以可以将多个RTCP分组封装在一个UDP包中。RTCP有如下五种分组类型:

**SR：**发送端报告，所谓发送端是指发出RTP数据报的应用程序或者终端，发送端同时也可以是接收端。

**RR：**接收端报告，所谓接收端是指仅接收但不发送RTP数据报的应用程序或者终端。

**SDES：**源描述，主要功能是作为会话成员有关标识信息的载体，如用户名、邮件地址、电话号码等，此外还具有向会话成员传达会话控制信息的功能。

**BYE：**通知离开，主要功能是指示某一个或者几个源不再有效，即通知会话中的其他成员自己将退出会话。

**APP：**由应用程序自己定义，解决了RTCP的扩展性问题，并且为协议的实现者提供了很大的灵活性。



上述五种分组的封装大同小异，下面只讲述SR类型，而其它类型请参考RFC3550。发送端报告分组SR（Sender Report）用来使发送端以多播方式向所有接收端报告发送情况。SR分组的主要内容有：相应的RTP流的SSRC，RTP流中最新产生的RTP分组的时间戳和NTP，RTP流包含的分组数，RTP流包含的字节数。SR包的封装如图3所示。



**版本（V）**：同RTP包头域。

**填充（P）**：同RTP包头域。

**接收报告计数器（RC）**：5比特，该SR包中的接收报告块的数目，可以为零。

**包类型（PT）**：8比特，SR包是200。

**长度域（Length）**：16比特，其中存放的是该SR包以32比特为单位的总长度减一。

**同步源（SSRC）**：SR包发送者的同步源标识符。与对应RTP包中的SSRC一样。NTP Timestamp（Network time protocol）SR包发送时的绝对时间值。NTP的作用是同步不同的RTP媒体流。

**RTP Timestamp**：与NTP时间戳对应，与RTP数据包中的RTP时间戳具有相同的单位和随机初始值。

**Sender’s packet count**：从开始发送包到产生这个SR包这段时间里，发送者发送的RTP数据包的总数. SSRC改变时，这个域清零。

**Sender`s octet count**：从开始发送包到产生这个SR包这段时间里，发送者发送的净荷数据的总字节数（不包括头部和填充）。发送者改变其SSRC时，这个域要清零。

**同步源n的SSRC标识符**：该报告块中包含的是从该源接收到的包的统计信息。

**丢失率（Fraction Lost）：**表明从上一个SR或RR包发出以来从同步源n(SSRC\_n)来的RTP数据包的丢失率。

**累计的包丢失数目**：从开始接收到SSRC\_n的包到发送SR,从SSRC\_n传过来的RTP数据包的丢失总数。

收到的扩展最大序列号：从SSRC\_n收到的RTP数据包中最大的序列号，

**接收抖动（Interarrival jitter）**：RTP数据包接受时间的统计方差估计

**上次SR时间戳（Last SR,LSR）**：取最近从SSRC\_n收到的SR包中的NTP时间戳的中间32比特。如果目前还没收到SR包，则该域清零。

上次SR以来的延时（Delay since last SR,DLSR）：上次从SSRC\_n收到SR包到发送本报告的延时。

RTP设计成允许应用自动扩展，连接数可从几个到上千个。例如，音频会议中，数据流量是内在限制的，因为同一时刻只有一两个人说话；对组播，给定连接数据率仍是常数，独立于连接数，但控制流量不是内在限制的。如每个参加者以固定速率发送接收报告，控制流量将随参加者数量线性增长，因此，速率必须按比例下降。

一旦确认地址有效，如后来标记成未活动，地址的状态应仍保留，地址应继续计入共享RTCP带宽地址的总数中，时间要保证能扫描典型网络分区，建议为30分钟。注意，这仍大于RTCP报告间隔最大值的五倍。

这个规范定义了除必需的CNAME外的几个源描述项，如NAME（人名）和EMAIL（电子邮件地址）。它也为定义新特定应用RTCP包类型的途径。给附加信息分配控制带宽应引起注意，因为它将降低接收报告和CNAME发送的速率而损害协议的性能。建议分配给单个参加者用于携带附加信息的RTCP带宽不要超过20%。而且并没有有意让所有SDES项包含在每个应用中。

### 3.3 RTP会话

#### 3.3.1 RTP会话过程

当应用程序建立一个RTP会话时，应用程序将确定一对目的传输地址。目的传输地址由一个网络地址和一对端口组成，有两个端口：一个给RTP包，一个给RTCP包，使得RTP/RTCP数据能够正确发送。RTP数据发向偶数的UDP端口，而对应的控制信号RTCP数据发向相邻的奇数UDP端口（偶数的UDP端口＋1），这样就构成一个UDP端口对。 RTP的发送过程如下，接收过程则相反。

1)RTP协议从上层接收流媒体信息码流（如H.263），封装成RTP数据包；RTCP从上层接收控制信息，封装成RTCP控制包。

2)RTP将RTP 数据包发往UDP端口对中偶数端口；RTCP将RTCP控制包发往UDP端口对中的接收端口。

### 3.4 RTP荷载H264码流



荷载格式定义三个不同的基本荷载结构，接收者可以通过RTP荷载的第一个字节后5位（如图2）识别荷载结构。

1)单个NAL单元包：荷载中只包含一个NAL单元。NAL头类型域等于原始 NAL单元类型,即在范围1到23之间

2)聚合包：本类型用于聚合多个NAL单元到单个RTP荷载中。本包有四种版本,单时间聚合包类型A (STAP-A)，单时间聚合包类型B (STAP-B)，多时间聚合包类型(MTAP)16位位移(MTAP16), 多时间聚合包类型(MTAP)24位位移(MTAP24)。赋予STAP-A, STAP-B, MTAP16, MTAP24的NAL单元类型号分别是 24,25, 26, 27

3)分片单元：用于分片单个NAL单元到多个RTP包。现存两个版本FU-A，FU-B,用NAL单元类型 28，29标识

常用的打包时的分包规则是：如果小于MTU采用单个NAL单元包，如果大于MTU就采用FUs分片方式。

因为常用的打包方式就是单个NAL包和FU-A方式，所以我们只解析这两种。

#### 3.4.12.1、单个NAL单元包

定义在此的NAL单元包必须只包含一个。这意味聚合包和分片单元不可以用在单个NAL 单元包中。并且RTP序号必须符合NAL单元的解码顺序。NAL单元的第一字节和RTP荷载头第一个字节重合。打包H264码流时，只需在帧前面加上12字节的RTP头即可。



#### IMG_2563.4.2分片单元（FU-A）

分片只定义于单个NAL单元不用于任何聚合包。NAL单元的一个分片由整数个连续NAL单元字节组成。每个NAL单元字节必须正好是该NAL单元一个分片的一部分。相同NAL单元的分片必须使用递增的RTP序号连续顺序发送(第一和最后分片之间没有其他的RTP包）。相似，NAL单元必须按照RTP顺序号的顺序装配。

当一个NAL单元被分片运送在分片单元(FUs)中时，被引用为分片NAL单元。STAPs,MTAPs不可以被分片。 FUs不可以嵌套。 即, 一个FU 不可以包含另一个FU。运送FU的RTP时戳被设置成分片NAL单元的NALU时刻。

图 4 表示FU-A的RTP荷载格式。FU-A由1字节的分片单元指示（如图5），1字节的分片单元头（如图6），和分片单元荷载组成。



S: 1 bit 当设置成1,开始位指示分片NAL单元的开始。当跟随的FU荷载不是分片NAL单元荷载的开始，开始位设为0。

E: 1 bit 当设置成1, 结束位指示分片NAL单元的结束，即, 荷载的最后字节也是分片NAL单元的最后一个字节。当跟随的 FU荷载不是分片NAL单元的最后分片,结束位设置为0。

R: 1 bit 保留位必须设置为0，接收者必须忽略该位

打包时，原始的NAL头的前三位为FU indicator的前三位，原始的NAL头的后五位为FU header的后五位。

取一段码流分析如下：

80 60 01 0f 00 0e 10 00 00 0000 00 7c 85 88 82 €`..........|???

00 0a 7f ca 94 05 3b 7f 3e 7f fe 14 2b 27 26 f8 ...??.;.>.?.+'&?

89 88 dd 85 62 e1 6d fc 33 01 38 1a 10 35 f2 14 ????b?m?3.8..5?.

84 6e 21 24 8f 72 62f0 51 7e 10 5f 0d 42 71 12 ?n!$?rb?Q~.\_.Bq.

17 65 62 a1 f1 44 dc df 4b 4a 38 aa 96 b7 dd 24 .eb??D??KJ8????$

前12字节是RTP Header

7c是FU indicator

85是FU Header

FU indicator（0x7C）和FU Header（0x85）换成二进制如下

0111 1100 1000 0101

按顺序解析如下：

0                            是F

11                          是NRI

11100                     是FU Type，这里是28，即FU-A

1                            是S，Start，说明是分片的第一包

0                            是E，End，如果是分片的最后一包，设置为1，这里不是

0                            是R，Remain，保留位，总是0

00101                     是NAl Type，这里是5，说明是关键帧

打包时，FUindicator的F、NRI是NAL Header中的F、NRI，Type是28；FU Header的S、E、R分别按照分片起始位置设置，Type是NAL Header中的Type。

解包时，取FU indicator的前三位和FU Header的后五位，即0110 0101（0x65）为NAL类型。

### 3.5 RTP荷载PS流

针对H264 做如下PS 封装：每个IDR NALU 前一般都会包含SPS、PPS 等NALU，因此将SPS、PPS、IDR 的NALU 封装为一个PS 包，包括ps 头，然后加上PS system header，PS system map，PES header+h264 raw data。所以一个IDR NALU PS 包由外到内顺序是：PSheader| PS system header | PS system Map | PES header | h264 raw data。对于其它非关键帧的PS 包，就简单多了，直接加上PS头和PES 头就可以了。顺序为：PS header | PES header | h264raw data。以上是对只有视频video 的情况，如果要把音频Audio也打包进PS 封装，也可以。当有音频数据时，将数据加上PES header 放到视频PES 后就可以了。顺序如下：PS 包=PS头|PES(video)|PES(audio)，再用RTP 封装发送就可以了。

        GB28181 对RTP 传输的数据负载类型有规定（参考GB28181 附录B），负载类型中96-127

        RFC2250 建议96 表示PS 封装，建议97 为MPEG-4，建议98 为H264

        即我们接收到的RTP 包首先需要判断负载类型，若负载类型为96，则采用PS 解复用，将音视频分开解码。若负载类型为98，直接按照H264 的解码类型解码。

        注：此方法不一定准确，取决于打包格式是否标准

PS 包中的流类型（stream type）的取值如下：

1)        MPEG-4 视频流： 0x10；

2)        H.264 视频流： 0x1B；

3)        SVAC 视频流： 0x80；

4)        G.711 音频流： 0x90；

5)        G.722.1 音频流： 0x92；

6)        G.723.1 音频流： 0x93；

7)        G.729 音频流： 0x99；

8)     SVAC音频流： 0x9B。

### 3.6 RTP的时间戳 同步

首先，了解几个基本概念：

时间戳单位：时间戳计算的单位不是秒之类的单位，而是由采样频率所代替的单位，这样做的目的就是 为了是时间戳单位更为精准。比如说一个音频的采样频率为8000Hz，那么我们可以把时间戳单位设为1 / 8000。

时间戳增量：相邻两个RTP包之间的时间差（以时间 戳单位为基准）。

采样频率：每秒钟抽取样本的次数，例如音频的采样率一般为8000Hz

帧率： 每秒传输或者显示帧数，例如25f/s

RTP包头的第2个32Bit即为RTP包的时间戳，Time Stamp ，占32位。

时间戳反映了RTP分组中的数据的第一个字节的采样时刻。在一次会话开始时的时间戳初值也是随机选择的。即使是没有信号发送时，时间戳的数值也要随时间不断的增加。接收端使用时间戳可准确知道应当在什么时间还原哪一个数据块，从而消除传输中的抖动。时间戳还可用来使视频应用中声音和图像同步。

在RTP协议中并没有规定时间戳的粒度，这取决于有效载荷的类型。因此RTP的时间戳又称为媒体时间戳，以强调这种时间戳的粒度取决于信号的类型。例如， 对于8kHz采样的话音信号，若每隔20ms构成一个数据块，则一个数据块中包含有160个样本（0.02×8000=160）。因此每发送一个RTP分 组，其时间戳的值就增加160。

官方的解释看懂没？没看懂？没关系，我刚开始也没看懂，那就听我的解释吧。

首先，时间戳就是一个值，用来反映某个数据块的产生（采集）时间点的， 后采集的数据块的时间戳肯定是大于先采集的数据块的。有了这样一个时间戳，就可以标记数据块的先后顺序。

第二，在实时流传输中，数据采集后立刻传递到RTP 模块进行发送，那么，其实，数据块的采集时间戳就直接作为RTP包的时间戳。

第三，如果用RTP来传输固定的文件，则这个时间戳 就是读文件的时间点，依次递增。这个不再我们当前的讨论范围内，暂时不考虑。

第四，时间戳的单位采用的是采样频率的倒数，例如采 样频率为8000Hz时，时间戳的单位为1 / 8000 ，在Jrtplib库中，有设置时间戳单位的函数接口，而ORTP库中根据负载类型直接给定了时间戳的单位（音频负载1/8000，视频负载1 /90000）

第五，时间戳增量是指两个RTP包之间的时间间隔， 详细点说，就是发送第二个RTP包相距发送第一个RTP包时的时间间隔（单位是时间戳单位）。

如果采样频率为90000Hz，则由上面讨论可知，时间戳单位为1/90000，我们就假设1s钟被划分了90000个时间块，那么，如果每秒发送25 帧，那么，每一个帧的发送占多少个时间块呢？当然是 90000/25 = 3600。因此，我们根据定义“时间戳增量是发送第二个RTP包相距发送第一个RTP包时的时间间隔”，故时间戳增量应该为3600。

下面说一下播放器如何对编码器发送的RTP包的时间戳做处理。这里的编码器是指发送RTP包的服务器，一般是指网络摄像机或RTSP服务器。

播放器本地建立了一个系统时钟，这个时钟一般根据系统的CPU时间来计算出来，当播放开始，时钟的时间为0。时间戳有分绝对时间和相对时间，收到的RTP包的时间戳为绝对时间，播放器需要把这个RTP时间戳转为播放器本地的相对时间戳，相对时间戳以0为开始，时间戳决定了一帧播放或渲染的时刻。当播放开始时，时钟开始运行，时间戳一直增加，播放器会用系统时钟的时间跟视频帧的时间戳（相对时间戳）进行对比，根据它们的大小关系决定是否马上渲染这帧还是要等待。但是，编码器发过来的时间戳只是一个参考值，有可能是不准的，为什么呢？编码器打时间戳基于的时钟跟播放器的本地时钟是可能不一样的，不同的编码器有不同的打时间戳算法，我只说一下常见的做法，一般地，编码器打时间戳是基于一个简单的规则：每发一个视频帧，PTS就递增一个固定的时间增量。根据TS流和Rtp协议的规定，视频的时钟频率是90000，也就是时间单位是1/90000秒。如果帧率是25帧，则相邻两帧的PTS增量就是3600，也就是每发一个帧，时间戳就增加3600，但是实际上两帧的时间增量应该根据它们实际经过的时间，而不是一个固定的值。实际上，编码器每秒采集图像的帧率并不是绝对固定的（取决于物理时钟的精度以及CPU处理的速度），有一定误差，平均采集帧率是25，但可能编码器某1秒采集了25帧，某一秒采集了24帧，假如都是按固定增量的方式递增时间戳，过一段时间后，编码器的时间戳跟播放器的时钟的时间就有很大差异，可能导致播放不流畅，或者缓冲了很多帧延时加大了。

   所以，解码器不能盲目相信编码器的时间戳，发现时间戳跟本地的时间差异太远需要尽快调整，应该实现一个动态适应的机制：当发现接收缓冲区里的帧数比较多时，要加快渲染速度；当缓冲区的帧数在一个比较低的水平（1-4），则稳定渲染速度

在Jrtplib中好像不需要自己管理时间戳的递增，由库内部管理。但在ORTP中每次数据的发送都需要自己传入时间戳的值，即自己 需要每次发完一个RTP包后，累加时间戳增量，不是很方便，这就需要自己对RTP的时间戳有比较深刻地理解，我刚开始就是因为没搞清楚，随时设置时间戳增 量导致传输一直有问题，困扰我好久。

多媒体通信同步方法，主要有时间戳同步法、同步标记法、多路复用同步法三种。下面主要讨论时间戳同步法，特别是RTP时间戳同步。内容包括RTP媒体间同步的实现，为什么需要RTCP的NTP时间来实现媒体间同步？没有RTCP，能实现RTP媒体间的同步吗？DirectShow时间戳和RTP时间 戳的区别，MPEG2-TS流的时间戳等。本文只简单讨论时间戳同步的原理，不涉及具体的实现方法，如音频帧和视频帧时间戳的计算方法，怎样根据时间戳去做音视频的呈现等。

根据RTP规范，不同的RTP媒体流是分开传输的，且使用各自独立的时间戳进行同步。假设在一次视频点播中，传输两路RTP媒体流，一路视频，一 路音频。根据视频帧时间戳，可以实现视频流内同步，这很好理解，通过视频帧时间戳可以计算出相邻视频帧的时间间隔，也就是视频帧之间的相对时间关系很容易 通过时间戳来确定，按照这个间隔去呈现视频，就可以获得较好的效果。同理，音频流也可以实现自身的同步。

那么音频和视频这两路媒体间如何实现同步呢？我们只使用音视频的RTP时间戳，看能否实现媒体间的同步。音视频的RTP时间戳的增长速率一般是不同的，但没关系，知道了具体的单位后，两者是可以通过单位换算联系起来的。如下图：

现在来看，这种方法好像可以实现同步，因为音视频被映射到同一个时间轴上了，音频和视频帧间的相对关系很清楚。慢着，RTP规范要求时间戳的初 始值应该是一个随机值，那么假设音频帧时间戳的初始值是随机值1234，视频帧时间戳的初始值是随机值5678，看起来应该是下面这样：

这么做合适吗？我们把音频帧时间戳1234和视频帧时间戳5678对应到绝对时间轴的0上，我们这么做的理由是什么？你可能会说，因为那是第一 个音频帧和第一个视频帧，所以可以对应到同一个点上，在第一幅图中我们就是这么做的，把音频帧时间戳0和视频帧时间戳0对应到绝对时间轴的0上。但是 RTP规范并没有规定第一个视频帧的时间戳和第一个音频帧的时间戳必须或者应该对应到绝对时间轴的同一个点上，从整个RTP规范中不能直接得出这样的结 论，也推导不出这样的结论。

我们上面两幅图所做的转换是不正确的，为什么呢？因为在做转换时，隐含了一个假设，我们想当然地认为这个假设是成立的，实际上它并不总是成立。这个假设就是第一个视频帧和第一个音频帧的时间戳应该对应到同一个点上，即无论它们时间戳是多少，都应该在同一时间播放。

仅仅使用RTP时间戳是无法实现媒体间同步的，根本的原因是音频时间轴和视频时间轴是完全独立的，通过音频帧和视频帧的时间戳，无法确定一个视频帧和一个音频帧的相对时间关系，也就是无法把它们都准确定位在绝对时间轴上，只能准确定位一个。

要实现RTP媒体间同步，需要借助于RTCP，在RTCP的SR包中，包含有<NTP时间，RTP时间戳>对，音频帧RTP时间戳和视 频帧RTP时间戳通过<NTP时间，RTP时间戳>对，都可以准确定位到绝对时间轴NTP上，音频帧和视频帧的相对时间关系就可以确定下来 了。

上面提到，我们的那个隐含的假设并不总是成立，那就是说它有成立的时候。那是不是说当它成立时，我们就可以不用RTCP来做媒体间同步了？答案是，基本上可以这么认为。

例如，对于RTP实时流，在发送端媒体间就同步的很好，在接收端只需做少许处理，不需要RTCP，就可以实现媒体间同步。当然，这只是少数例外。因为RTP规范并不包括这个假设，所以我们还是按照RTP规范来做吧。

下面说一下DirectShow和MPEG2-TS的时间戳。DirectShow中的时间戳和RTP中的时间戳，除了单位不一样，计算方法不一样 外，本质的区别就是DirectShow中的音频帧和视频帧时间戳使用的是同一个时间轴，所以不需要借助其他的东西，仅仅使用音频帧时间戳和视频帧时间戳 就可以实现媒体间同步。MPEG2-TS流中也有时间戳，它的时间戳和RTP及DirectShow的时间戳都不同，TS流中的音频帧和视频帧时间戳使用 的也是同一个时间轴，TS流中的音频和视频是复用的，这在一定程度上就起到了同步的作用，所以它并不是在每个帧上都打时间戳，比如它的PTS时间戳就是每 隔0.1秒一个，缺失的时间戳是通过其他时间戳插值计算出来的。

### 3.7 RTP打包方式

RFC3984 给出了3 中不同的RTP 打包方案：

（1）Single NALU Packet:在一个RTP 包中只封装一个NALU，在本文中对于小于 1400字节的NALU 便采用这种打包方案。  
       （2）Aggregation Packet:在一个RTP 包中封装多个NALU，对于较小的NALU 可以采用这种打包方案，从而提高传输效率。   
       （3）Fragmentation Unit:一个NALU 封装在多个RTP包中，在本文中，对于大于1400字节的NALU 便采用这种方案进行拆包处理。

## 4. RTSP

### 4.1 RTSP概述

实时流协议RTSP（Real Time Streaming Protocol）最早由Real Networks和Netscape公司共同提出，它位于RTP和RTCP之上，其目的是希望通过IP网络有效地传输多媒体数据。RTSP在体系结构上位于RTP和RTCP之上，它使用TCP或RTP完成数据传输。HTTP与RTSP相比，HTTP传送HTML，而RTP传送的是多媒体数据。HTTP请求由客户机发出，服务器作出响应；使用RTSP时，客户机和服务器都可以发出请求，即RTSP可以是双向的。

作为一个应用层协议，RTSP提供了一个可供扩展的框架，它的意义在于使得实时流媒体数据的受控和点播变得可能。总的说来，RTSP是一个流媒体表示协议，主要用来控制具有实时特性的数据发送，但它本身并不传输数据，而是必须依赖于下层传输协议所提供的某些服务。RTSP可以对流媒体提供诸如播放、 暂停、快进等操作，它负责定义具体的控制消息、操作方法、状态码等，此外还描述了与RTP间的交互操作。

RTSP在制定时较多地参考了HTTP/1.1协议，甚至许多描述与HTTP/1.1完全相同。RTSP之所以特意使用与HTTP/1.1类似的语 法和操作，在很大程度上是为了兼容现有的Web基础结构，正因如此，HTTP/1.1的扩展机制大都可以直接引入到RTSP中。

由RTSP控制的媒体流集合可以用表示描述（Presentation Description）来定义，所谓表示是指流媒体服务器提供给客户机的一个或者多个媒体流的集合，而表示描述则包含了一个表示中各个媒体流的相关信 息，如数据编码/解码算法、网络地址、媒体流的内容等。

虽然RTSP服务器同样也使用标识符来区别每一流连接会话（Session），但RTSP连接并没有被绑定到传输层连接（如TCP等），也就是说在 整个RTSP连接期间，RTSP用户可打开或者关闭多个对RTSP服务器的可靠传输连接以发出RTSP 请求。此外，RTSP连接也可以基于面向无连接的传输协议（如UDP等）。

RTSP协议目前支持以下操作：

检索媒体：允许用户通过HTTP或者其它方法向媒体服务器提交一个表示描述。如表示是组播的，则表示描述就包含用于该媒体流的组播地址和端口号；如果表示是单播的， 为了安全在表示描述中应该只提供目的地址。

邀请加入：媒体服务器可以被邀请参加正在进行的会议，或者在表示中回放媒体，或者在表示中录制全部媒体或其子集，非常适合于分布式教学。，会议中几方可轮流按远程控制按钮。

添加媒体：通知用户新加入的可利用媒体流，这对现场讲座来讲显得尤其有用。与HTTP/1.1类似，RTSP请求也可以交由代理、通道或者缓存来进行处理。　如服务器告诉用户可获得附加媒体内容，对现场讲座显得尤其有用。

### 4.2 RTSP状态

RTSP控制通过单独协议发送的流，与控制通道无关。例如，RTSP控制可通过TCP连接，而数据流通过UDP。因此，即使媒体服务器没有收到请求，数据也会继续发送。在连接生命期，单个媒体流可通过不同TCP连接顺序发出请求来控制。所以，服务器需要维持能联系流与RTSP请求的连接状态。RTSP中很多方法与状态无关，但下列方法在定义服务器流资源的分配与应用上起着重要的作用：

SETUP：让服务器给流分配资源，启动RTSP连接。

PLAY与RECORD：启动SETUP 分配流的数据传输。

PAUSE：临时停止流，而不释放服务器资源。

TEARDOWN：释放流的资源，RTSP连接停止。

标识状态的RTSP方法使用连接头段识别RTSP连接，为响应SETUP请求，服务器连接产生连接标识。

### 4.3 资源预订协议RSVP协议

由于音频和视频数据流比传统数据对网络的延时更敏感，要在网络中传输高质量的音频、视频信息，除带宽要求之外，还需其他更多的条件。RSVP(ResourceReserveProtocol)是正在开发的Internet上的资源预订协议，使用RSVP预留一部分网络资源（即带宽），能在一定程度上为流媒体的传输提供QoS。在某些试验性的系统如网络视频会议工具vic中就集成了RSVP。

RSVP运行在传输层，在IP上层。与ICMP和IGMP相比，它是一个控制协议。 RSVP协议的两个重要概念是流与预定。流是从发送者到一个或多个接收者的连接特征，通过IP包中"流标记"来认证。发送一个流前，发送者传输一个路径信息到目的接收方，这个信息包括源IP地址、目的IP地址和一个流规格。这个流规格是由流的速率和延迟组成的，这是流的QoS需要的。接收者实现预定后，基于接收者的模式能够实现一种分布式解决方案。

　 RSVP领域的发展是非常迅速的，但目前并没有在任何一种网络上得到证实，它的应用只是局限在测试的小Intranet网络上。因为RSVP的预定必须建立在完全流方式的基础上，其可扩展性问题倍受关注。

## AAC音频格式

### 5.1 音频初步介绍

DTS：全称为Digital TheaterSystems(数字影院系统)，是一种有损多声道家庭影院音频格式，但它用了很高的码率进行编码，通常为768-1536kbps，能够营造出比AC3更好的影院效果。两者都是基于多声道的编码技术，通用5.1声道（三组全频立体声）--左/右前置，中置/低音，左/右后置环绕。其中低音单元专门重放120HZ以下的低频，以增强效果。DTS的数据流量通常为768/1536Kbps,压缩比只有3：1，其从提高数字空间的利用率着手，使信息数据得以充分利用。理论上它的声场无论在真实性、细腻性、连续性、宽广性和层次性方面优于AC3。

AC3, 全称为Audio Coding version 3，是Dolby实验室所发展的有损音频编码格式。AC3最被广泛应用于5.1声道，是Dolby Pro Logic的继承者，不同的地方在于AC3提供6个独立的声道而Pro Logic混合其环绕声道。AC3普及度很高，以384-448 kbps的码率应用于LaserDisc和DVD，也经常以640 kbps的码率广泛用在电影院。AC3通常为384/448Kbps,采用了高压缩编码（12：1），通过大副删除在理论上认为多余的细节信号，从而达到减少数据量的目的。实际上AC3的编码技术相当优秀，在如此低的数据流量下带给观众的依然是较为真实的声音体验。

M4A：.M4A 是MPEG4音频标准的文件的扩展名。在MPEG4标准中提到，普通的MPEG4文件扩展名是.mp4。自从Apple开始在它的iTunes以及 iPod中使用m4a以区别MPEG4的视频和音频文件以来，.m4a这个扩展名变得流行了。目前，几乎所有支持MPEG4音频的软件都支持.m4a。最常用的.m4a文件是使用AAC编码格式的（文件），不过其他的格式，比如Apple Lossless甚至mp3也可以被放在.m4a容器里。可以安全的把只包含音频的.mp4 文件的扩展名改成.m4a，以便让它能在你喜欢的播放器里播放，反之亦然。

M4B：M4B格式是一种可以在iPod/iPhone及iPad上播放的称为“有声读物”的音频文件格式。主要优点：M4B有声读物是一种为“音频书籍”专门设计的格式。支持书签是它的主要优点。它允许书在任何时候暂停和恢复播放，恢复播放时，音频将从暂停的点开始。这一特点显然是非常重要的，因为有声读物的时间可能长很多。因此，它更经常用于有声读物和podcasts（播客）。

M4R：.M4R iPhone 手机的铃声格式

M4V：.M4V 是一个标准视频文件格式，由苹果公司创造。此种格式为 iPod 、 iPhone 和 PlayStation Portable 所使用，同时此格式基于 MPEG-4 编码第二版。M4V是一种应用于网络视频点播网站和移动手持设备的视频格式，是MP4格式的一种特殊类型，其后缀常为.MP4或.M4V，其视频编码采用H264或H264/AVC，音频编码采用AAC。采用H264高清编码，相比于传统On2 VP6、H263、Divx等，能够以更小的体积实现更高的清晰度。

### 5.2 AAC 音频格式

ACC 的特点：提升的压缩率：可以以更小的档大小获得更高的音质；支持多声道：可提供最多48 个全音域声道；更高的分辨率：最高支持96KHz 的取样频率；提升的解碼效率：译码播放所占的资源更少。AAC 是个大家族，目前已经制定了如下的9 种规格，以适应不同场合的需要：

MPEG-2 AAC LC 低复杂度规格（Low Complexity）--比较简单，没有增益控制，但提高了

编码效率，在中等码率的编码效率以及音质方面，都能找到平衡点

MPEG-2 AAC Main 主规格

MPEG-2 AAC SSR 可变采样率规格（Scaleable Sample Rate）

MPEG-4 AAC LC 低复杂度规格（Low Complexity）------现在的手机比较常见的MP4文件中的音频部份就包括了该规格音频文件

MPEG-4 AAC Main 主规格 ------包含了除增益控制之外的全部功能，其音质最好

MPEG-4 AAC SSR 可变采样率规格（Scaleable Sample Rate）

MPEG-4 AAC LTP 长时期预测规格（Long Term Predicition）

MPEG-4 AAC LD 低延迟规格（Low Delay）

MPEG-4 AAC HE 高效率规格（High Efficiency）-----这种规格适合用于低码率编码

（1）Main（主规格）包含了除增益控制（Gain Control）之外的全部功能，其音质最好。

（2）低复杂度规格（Low Complexity）则使用比较简单的时域噪声修整（Temporal Noise Shaping，TNS）模块，也缺少了预测（Predict ion）和增益控制（Gain Control）模块，以此降低复杂度，提高了编码效率。SSR（Scalable Sampling Rate，可变取样率规格）和LC 规格大体相似，但多了增益控制功能。

（3）MPEG-4 AAC LTP（Long TermPrediction，长时期预测规格）、MPEG-4 AAC LD（Low Delay，低延迟规格）和MPEG－4 AAC HE（High Efficiency，高效率规格）都是用在低码率下编码，尤其是MPEG－4 AAC HE 规格由于高效率高音质（低码率下而言），再加上有Nero AAC 编码器的支持，所以被越来越多地应用到低码率编码种。事实上，Main 规格和LC 规格的音质相差并不大，但是编码效率却相差明显，所以目前使用地最多的当数LC 规格。无怪乎很多编码器甚至只支持LC 规格了！

和MP3 一样，炙手可热的AAC 格式也催生了大量的编码器。包括商业性质的和完全免费的，技术保密的和源码开放的，对消费者提供的和对开发者提供的，各种各样的编码器都有。其中主要的编码器有：（1）FhG：Fraunhofer IIS 研发的权威编码器，拥有很好的音质，可惜现在不对外提供了！

（2）Nero AAC：可能是目前最完美的AAC 编码器了，同时支持LCAAC 和HE AAC 规格。这个编码器是商业性质的，随着包含了Nero Digital 的Nero 6 一起发布，可能很多有刻录机的计算机里已经安装了该软件。Nero AAC 编码器提供了质量最好的VBR LC AAC 格式，同时HE AAC 规格保证了它在低码率下也有不俗的表现。此外，这个编码器还支持多声道编码！

（3）QuickTime／iTune：苹果公司的两款多媒体软件都提供了AAC 编码功能，其编码技术来自杜比实验室（Dolby Laboratories）。起初，它们的编码器功能比较简单，只支持双声道的CBR 模式LC AAC 格式，不过最近苹果刚刚在QuickTime Pro 版中提供了VBR 编码。本文后面将介绍的iTune 尽管还不具备VBR 模式编码，但是它却是目前音质最好的中码率编码器！

（4）FAAC：最好的命令行编码器，只支持LC 规格。其表现中规中矩，最近几个月新版本的发布将其音质提升了一个台阶。这个免费的编码器已经完全能和商业性质的编码器相媲美了，是搭配EAC 的不二选择。

（5）Psytel：这是最早期的命令行编码器，曾经很是风光，其音质在以前是最好的，就是编码速度很慢。可惜的是其作者现在为Nero 工作了，这个编码器也从此停止了更新。因为有相似的FAAC，所以不建议使用它。

（6）CodingTechnologies：该公司的AAC 编码技术已经被Real 公司加入到他们的产品中了。最新的RealProducer 10 和RealPlayer 10 都包含了这个AAC 编码器。

（7）HHI／zPlane （Compaact！）：Compaact！是一款新出的AAC 编码器，和Nero 一样，也是商业性质的。它的特性相当诱人，支持LC 和Main 规格，支持CBR 和VBR 模式，支持多声道高取样率（24bi t／96KHz），还支持命令行操作！

（8）Dolby AAC：杜比公司的AAC 编码器，因为着重在网络广播，所以低码率下优势明显。据称它编码的48kbps 立体声音质比其它AAC 编码器好20％，甚至64kbps 的Dolby AAC 可媲美128kbps 的MP3！

前使用最多的是LC和HE(适合低码率)。流行的Nero AAC编码程序只支持LC，HE，HEv2这三种规格，编码后的AAC音频，规格显示都是LC。HE其实就是AAC（LC）+SBR技术，HEv2就是AAC（LC）+SBR+PS技术；

AAC原始码流（又称为“裸流”）是由一个一个的ADTS frame组成的。其中每个ADTS frame之间通过syncword（同步字）进行分隔。同步字为0xFFF（二进制“111111111111”）。AAC码流解析的步骤就是首先从码流中搜索0x0FFF，分离出ADTS frame；然后再分析ADTS frame的首部各个字段。

AAC能够在一条音轨中包括48条全带宽（直到96khz）音频声道，加上15条低频增强（LFE，限制到120Hz）声道，直到15条数据流并且更多。杜比也参与了AAC的开发，MPEG规格的一部分，及考虑继承MP3。AAC能输出AC-3的任何码率，胜过AC-3，压缩率更高，但技术上更加复杂。AAC在5.1声道中以400 Kbps和双声道中以180 Kbps输出，就已经能提供良好的聆听效果。

#### 5.2.1参考链接

1.ADTS头解析

https://blog.csdn.net/jay100500/article/details/52955232

### 5.3 AAC音频文件格式

#### 5.3.1 ADIF ＆ ADTS

AAC的音频文件格式有ADIF ＆ ADTS：

ADIF：Audio Data Interchange Format 音频数据交换格式。这种格式的特征是可以确定的找到这个音频数据的开始，不需进行在音频数据流中间开始的解码，即它的解码必须在明确定义的开始处进行。故这种格式常用在磁盘文件中。

ADTS：Audio Data Transport Stream 音频数据传输流。这种格式的特征是它是一个有同步字的比特流，解码可以在这个流中任何位置开始。它的特征类似于mp3数据流格式。

简单说，ADTS可以在任意帧解码，也就是说它每一帧都有头信息。ADIF只有一个统一的头，所以必须得到所有的数据后解码。且这两种的header的格式也是不同的，目前一般编码后的和抽取出的都是ADTS格式的音频流。

#### 5.3.2 ADIF

AC的ADIF格式见下图：

IMG_256

ADIF 的头信息：



ADIF头信息位于AAC文件的起始处，接下来就是连续的 raw data blocks。

 组成ADIF头信息的各个域如下所示：

#### 5.3.3 ADTS

ADTS格式如下：

AAC的ADTS的一般格式见下图：

IMG_257

 图中表示出了ADTS一帧的简明结构，其两边的空白矩形表示一帧前后的数据。

ADTS 的固定头信息：



syncword ：总是0xFFF, 代表一个ADTS帧的开始, 用于同步.解码器可通过0xFFF确定每个ADTS的开始位置.因为它的存在，解码可以在这个流中任何位置开始, 即可以在任意帧解码。

ID：MPEG Version: 0 for MPEG-4，1 for MPEG-2

Layer：always: '00'

protection\_absent：Warning, set to 1 if there is no CRC and 0 if there is CRC

profile：表示使用哪个级别的AAC，如01 Low Complexity(LC) -- AAC LCprofile的值等于 Audio Object Type的值减1.profile = MPEG-4 Audio Object Type - 1



sampling\_frequency\_index：采样率的下标



channel\_configuration：声道数，比如2表示立体声双声道



ADTS的可变头信息：



aac\_frame\_length：一个ADTS帧的长度包括ADTS头和AAC原始流。frame length, this value must include 7 or 9 bytes of header length:

aac\_frame\_length = (protection\_absent == 1 ? 7 : 9) + size(AACFrame)

protection\_absent=0时, header length=9bytes

protection\_absent=1时, header length=7bytes

adts\_buffer\_fullness：0x7FF 说明是码率可变的码流。

number\_of\_raw\_data\_blocks\_in\_frame：表示ADTS帧中有number\_of\_raw\_data\_blocks\_in\_frame + 1个AAC原始帧。所以说number\_of\_raw\_data\_blocks\_in\_frame == 0 表示说ADTS帧中有一个AAC数据块。

(一个AAC原始帧包含一段时间内1024个采样及相关数据)

（1）帧同步目的在于找出帧头在比特流中的位置，13818-7规定，aac ADTS格式的帧头同步字为12比特的“1111 1111 1111”.

（2）ADTS的头信息为两部分组成，其一为固定头信息，紧接着是可变头信息。固定头信息中的数据每一帧都相同，而可变头信息则在帧与帧之间可变。

AAC文件处理流程:

(1)判断文件格式，确定为ADIF或ADTS

(2)若为ADIF，解ADIF头信息，跳至第6步。

(3)若为ADTS，寻找同步头。

(4)解ADTS帧头信息。

(5)若有错误检测，进行错误检测。

(6)解块信息。

(7)解元素信息。

AAC解码流程图:



在主控模块开始运行后，主控模块将AAC比特流的一部分放入输入缓冲区，通过查找同步字得到一帧的起始，找到后，根据ISO/IEC 13818-7所述的语法开始进行Noisless Decoding(无噪解码)，无噪解码实际上就是哈夫曼解码，通过反量化(Dequantize)、联合立体声（JointStereo），知觉噪声替换（PNS）,瞬时噪声整形（TNS），反离散余弦变换（IMDCT），频段复制（SBR）这几个模块之后，得出左右声道的PCM码流，再由主控模块将其放入输出缓冲区输出到声音播放设备。

### 5.4 AAC 与 MP3比较

作为一种高压缩比的音频压缩算法，远胜MP3；在音质方面，由于采用多声道，和使用低复杂性的描述方式，使其比几乎所有的传统编码方式在同规格的情况下更胜一筹。一般来说，AAC可以在对比MP3文件缩小30%的前题下提供更好的音质。AAC是目前唯一一个，能够在所有的EBU试听测试项目的获得“优秀”的网络广播格式。

AAC与MP3规格对比

比特率：AAC – 最高超过400kbps /MP3 – 32~320kbps

采样率：AAC – 最高96kHz / MP3– 最高48kHz

声道数：AAC – （5.1）六声道 / MP3 – 两声道

采样精度：AAC – 最高32bit / MP3– 最高16bit

AAC格式扩展名

.AAC – 基于MPEG-2的音频编码技术，属于传统的AAC编码。

.MP4 – 基于MPEG-4的音频编码技术。

.M4A – 苹果(Apple)公司对纯音频MP4文件采用的扩展名，本质和音频MP4相同。（包含视频的MP4文件为”.M4V”）

## AVI封装格式

### 6.1 AVI概述

AVI是音频视频交错（Audio Video Interleaved）的英文缩写，它是Microsoft公司开发的一种符合RIFF文件规范的数字音频与视频文件格式。AVI格式允许视频和音频交错在一起同步播放，支持256色和RLE压缩，但AVI文件并未限定压缩标准，因此，AVI文件格式只是作为控制界面上的标准，不具有兼容性，用不同压缩算法生成的AVI文件，必须使用相应的解压缩算法才能播放出来。

在介绍AVI文件前，我们要先来看看RIFF文件结构。AVI文件采用的是RIFF文件结构方式，RIFF(Resource Interchange File Format,资源互换文件格式)是微软公司定义的一种用户管理Windows环境中多媒体数据的文件格式，波形音频wave,MIDI和数字视频AVI都采用这种格式存储。RIFF文件的实际数据中，使用了列表（List）和块（Chunk）的形式来组织。列表可以嵌套列表和块。整个RIFF文件可以看成一个数据库，其数据块ID为RIFF ，称为RIFF块。一个RIFF文件中只允许存在一个RIFF块。RIFF块中包含一系列的子块，其中有一种子块的ID为“List”，称为LIST块，LIST块中可以再包含一系列的子块，但除了LIST块外的其他所有的子块都不能再包含子块。 RIFF和LIST块分别比普通的数据块多了一个被称为形式类型（Form Type）或者列表类型（List Type）的数据域。整体就是RIFF-len-”avi”-LIST-len-”hdrl”-avih-strlaud-strh-strfh264-strl-strh-strf

RIFF块（RIFF结构）

“RIFF”表示字符串（Chunk ID）

RIFF文件大小（数据块的大小）

形式类型或者列表类型“AVI”,”WAVE” …….



RIFF文件大小=实际数据长度+4（形式类型或者列表类型的大小）。 也就是说RIFF文件大小不包括“RIFF”域和“文件大小”域本身的大小（共8个字节），如文件大小域数据为28 69 6B 00,则实际为0x006b6928，转换为十进制为7039272，实际在Windows系统下看到的文件大小为7039280，即换算成实际大小时要加上8。可能文件大小值是低字节在前。

LIST块（LIST结构）

“LIST”表示字符串（Chunk ID）

LIST块大小

形式类型或者列表类型

LIST实际数据



LIST块大小域=实际的列表数据+4（形式类型或者列表类型域的大小）也就是说listSize值不包括“LIST”域和listSize域本身的大小。如该位置数据为5E 00 00 00 ,即0x0000005E，转换成十进制为94，实际总长为102bytes,即换算成列表总长度是要加上8。但是有一点要特别注意的是，当前List具体包括到哪里，可能会有List嵌套。如截图中，选中的部分为一个hdrl的List；其中包含两个strl的List，一个音频，一个视频；每个strl List又包含一个strh和一个strf。

Chunk块（Chunk结构，普通块）

Chunk ID——表示块类型的四字符码

Chunk 块大小——记录了整个块的大小

Chunk数据

Chunk块大小=实际的块数据长度，而不包括ckID域和ckSize域本身的大小



从00000000-000007F3为一个WindowsAVI文件的信息区部分：

它是文件的第一个LIST块。在它的内部记录着整个文件的系统构成，如告诉播放软件“我是一个AVI文件”，“在我体内有几个数据流”，“每个数据流包含着什么数据——图像，声音或其他”，“如果是图像数据流，那么它的大小，颜色，压缩方式，播放速度等等是怎样规定的”，“如果是声音数据流，那么它的压缩方式，播放效果等等又将有何规定”……在这些信息区中还有多个附属的LIST块，也就是前面提到的“子块”，它们用来记录每个数据流的全部信息。而这些附属LIST块与数据流之间保持着一一对应的关系，即第一个附属LIST块对应与00号数据流；第二个附属LIST块对应01号数据流……要想解释数据流，我们必须先了解AVI文件中数据块是什么。在AVI文件中，数据块是被放置在数据区中的一个有起始标识（由“数据流识别码”和“数据块存储方式识别码”组成，请参见对数据区部分的说明），并指明大小和数据内容的数据段，那么，数据流就是那些相互之间具有练习的同种数据类型的数据块集合。

00000000-00000003、多媒体文件识别码：RIFF

00000004-00000007、文件大小（10EDICH字节）——8字节

00000008-0000000B、AVI文件识别码

0000000C-0000000F、第一个LIST块识别码

00000010-00000013、第一个LIST块的大小（168H字节）

00000014-00000017、hdrl部分识别码，代表后面的数据记录着此文件的格式

00000018-0000001B、hdrl部分所包含的avih块识别码，此模块记录着本文件的初始化信息

0000001C-0000001F、avi块大小（38H字节）

00000020-00000023、每帧画面显示所维持多少个百万分之一秒，本例为1046h，即6667百万分之一秒约0.07秒。所以在播放此文件时，你看到的画面约每秒15帧

### 6.2 AVI Layout(AVI文件层次划分)

AVI Layout(AVI文件层次划分)



**1. 信息块——**包括文件的通用信息，定义数据格式，所用的压缩算法等参数

**2. 数据块——**包含实际数据流，即图像和声音序列数据。这是文件的主体，也是决定文件容量的主要部分。视频文件的大小等于该文件的数据率乘以该视频播放的时间长度

**3. 索引块——**索引块包含数据块列表好它们在文件中的位置，以提供文件内数据随机存取能力。

#### 6.2.1 信息块

AVI文件时目前使用的最复杂的RIFF文件，他能同时存储同步表现的音频视频数据。AVI的RIFF块的形式类型是AVI，它包含3个子块，如下所述。

1. 信息块：ID为“hdrl”的LIST块，定义AVI文件的数据格式 “hdrl”LIST块包含两个子块，一个是ID为“avih”的子块和一个是ID为”strl”的LIST块。

**“avih”块结构：**用于记录AVI文件的全局信息，比如流的数量，视频图像的宽和高等：

typedef struct

{

DWORD ChunID; // 必须为'avih'

DWORD ChunkSize; //本数据结构的大小，不包括最初的8個位元組（ID和Size兩個域）

DWORD dwMicroSecPerFrame ; //显示每帧所需的时间ns，定义avi的显示速率

DWORD dwMaxBytesPerSec; //最大的数据传输率

DWORD dwPaddingGranularity; //记录块的长度需为此值的倍数，通常是2048

DWORD dwFlages; //AVI文件的特殊属性，如是否包含索引块，音视频数据是否交叉存储

DWORD dwTotalFrame; //文件中的总帧数

DWORD dwInitialFrames; //说明在开始播放前需要多少桢

DWORD dwStreams; //文件中包含的数据流个数

DWORD dwSuggestedBufferSize; //建议使用的缓冲区的大小，

//通常为存储一桢图像以及同步声音所需要的数据之和

DWORD dwWidth; //图像宽

DWORD dwHeight; //图像高

DWORD dwReserved[4]; //保留值

}MainAVIHeader;

接着就是”avih”块的数据部分：一个或多个“strl”子列表

1. 文件中有多少个流，这里就对应有多少个“strl”子列表）。

2. 每个“strl”字列表至少包含一个“strh”块和一个“strf”块，

3. “strd”块（保存编解码器需要的一些配置信息）和“strn”块（保存流的名字）是可选的。

4. 注意：“strl”子列表出现的顺序与媒体流的编号是对应的，比如第一个“strl”字列表说明的是第一个流（Stream 0）,第二个“strl”字列表说明的是第二个流（Stream 1），以此类推。

**“strh”块结构**：用于说明这个“strl”LIST块对应的数据流的头信息：

typedef struct

{

　FOURCC fccType; //4字节，表示数据流的种类，vids 表示视频数据流，auds 音频数据流

　FOURCC fccHandler; //4字节 ，表示数据流解压缩的驱动程序代号

　DWORD dwFlags; //数据流属性

　WORD wPriority; //此数据流的播放优先级

　WORD wLanguage; //音频的语言代号

　DWORD dwInitalFrames; //说明在开始播放前需要多少桢

　DWORD dwScale; //数据量，视频每桢的大小或者音频的采样大小

　DWORD dwRate; //dwScale /dwRate = 每秒的采样数

　DWORD dwStart; //数据流开始播放的位置，以dwScale为单位

　DWORD dwLength; //数据流的数据量，以dwScale为单位

　DWORD dwSuggestedBufferSize; //建议缓冲区的大小

　DWORD dwQuality; //解压缩质量参数，值越大，质量越好

　DWORD dwSampleSize; //音频的采样大小

　RECT rcFrame; //视频图像所占的矩形

}AVIStreamHeader;

**“strf” 块结构：**“strf”子块紧跟在“strh”子块之后，其结构是“strh”子块的类型而定，如下所述：

如果strh子块是视频数据流，则strf子块的内容是一个BITMAPINFO结构，如下。

typedef struct tagBITMAPINFO

{

　BITMAPINFOHEADER bmiHeader;

　RGBQUAD bmiColors[1]; //颜色表

}BITMAPINFO;

typedef struct tagBITMAPINFOHEADER

{

　DWORD biSize;

　LONG biWidth;

　LONG biHeight;

　WORD biPlanes;

　WORD biBitCount;

　DWORD biCompression;

　DWORD biSizeImage;

　LONG biXPelsPerMeter;

　LONG biYPelsPerMeter;

　DWORD biClrUsed;

　DWORD biClrImportant;

}BITMAPINFOHEADER;

如果strh子块是音频数据流，则strf子块的内容是一个WAVEFORMAT结构，如下：

typedef struct

{

　WORD wFormatTag;

　WORD nChannels; //声道数

　DWORD nSamplesPerSec; //采样率

　DWORD nAvgBytesPerSec; //WAVE声音中每秒的数据量

　WORD nBlockAlign; //数据块的对齐标志

　WORD biSize; //此结构的大小

}WAVEFORMAT

**strd”块结构：**“strd”子块紧跟在strf子块后，存储供压缩驱动程序使用的参数，不一定存在，也没有固定的结构。

“strl”List块定义的AVI数据流依次将“hdrl”LIST块中的数据流头结构与“movi”LIST块中的数据联系在一起，第一个数据流头结构用于数据流0，第二个用于数据流1，依次类推。

#### 6.2.2 数据块

二. 数据块：ID为“movi”的LIST块，包含AVI的音视频序列数据

用于保存真正的媒体流数据（视频图像帧数据或音频采样数据等）。保存方式为：

1. 将数据块直接嵌套在“movi”列表里面

2. 将几个数据块分组成一个“rec”列表后再编排进“movi”列表

（注意：在读取AVI文件内容时，建议将一个“rec”列表中的所有数据块一次性读出）

但是，当文件中包含有多个流的时候，数据块与数据块之间如何来区别呢？于是数据块使用了一个四字符吗来表征它的类型，这个四字符码由2个字节的类型吗和2个字节的流编号组成。

“db”——非压缩视频

“dc”——压缩视频

“pc”——改用新的调色板

“wb”——音缩视频

比如:

第一个流（Stream 0）是音频，则表征音频数据块的四字符码为“00wb”;

第二个流（Steam 1）是视频，则表征视频数据块的四字符码为“01db”或“01dc”。

对于视频数据来说，在AVI数据序列中间还可以定义一个新的调色板，每个改变的调色板数据块永“xxpc”来表征，新的调色板使用一个数据结构AVIPALCHANGE来定义。（注意：如果一个流的调色板中途改变，则应在这个流格式的描述中，也及时AVISTREMAHEADER结构的dwFlags中包含一个AVISF\_VIDEO\_PALCHANGES标记）另外，文字数据块可以使用随意的类型码表征。

#### 6.2.3 索引块

三. 索引块：ID为“idxl”的子块，定义“movi”LIST块的索引数据，是可选块。

最后紧跟在“hdr”列表块和“movi”列表块之后的，就是AVI文件可选的索引块。这个索引块为AVI文件中每一个媒体数据块进行索引，并且记录它们在文件中的偏移（可能相对于“movi”列表，也可能相对于AVI文件开头）。索引块使用一个四字符码“idxl”来表征，索引信息使用一个数据结构AVIOLDINDEXl来定义。

typedef struct \_avioldindex {

FOURCC fcc; // 必须为‘idx1’

DWORD cb; // 本数据结构的大小，不包括最初的8个字节（fcc和cb两个域）

struct \_avioldindex\_entry {

DWORD dwChunkId; // 表征本数据块的四字符码

DWORD dwFlags; // 说明本数据块是不是关键帧、是不是‘rec ’列表等信息

DWORD dwOffset; // 本数据块在文件中的偏移量

DWORD dwSize; // 本数据块的大小

} aIndex[]; // 这是一个数组！为每个媒体数据块都定义一个索引信息

} AVIOLDINDEX;

注意：如果一个AVI文件包含有索引块，则应在AVI信息头的描述中，也及时AVIMAINHEADER结构的dwFlags中包含一个AVI\_HASINDEX标记

还有一种特殊的数据块，用一个四字符码“JUNK”来表征，它用于内部数据的对齐（填充），应用程序应该忽略这些数据块的实际意义。

#### 6.2.4 索引解释

假如要将avi文件分成头、中、尾的话，RRFT开始到movi段之间应该算是头，movi算是中间，index就算是尾部了。index其实就是movi中的内容的索引，作用是在拖动视频进度条时，能让解码器迅速定位到要找的视频帧。索引以“idx1”开头，后面跟着的是整个索引的大小，其基本结构是：

struct AVI\_index\_entry{  
     unsigned char dwChunkId[4];//表征本数据块的四字符码比如01db压缩视频 00wb音频  
     DWORD dwFlags;//说明本数据块是不是关键帧、是不是‘rec’列表等信息  
     DWORD dwOffset;//本数据块在文件中的偏移量  
     DWORD dwSize;//本数据块的大小  
};

idx1”就是索引的头

30 5b 00 00是索引的大小

大家看看实例的第二行，它们对应avi文件中的每一帧或者没一段音频：

30 30 64 63（“00dc”）是数据结构中的dwChunkId，当该段数据时音频时取（“01wb")

10 00 00 00 是dwFlags，关键帧取值为10 00 00 00，否则取值为00 00 00 00

04 00 00 00 是dwOffset;//本数据块在文件中的偏移量

BF 0E 00 00 是dwSize，对应帧的大小

　　在最初合成avi视频的时候，我是没有添加index段的，视频一样可以播放。可是当你要拖动视频时，视频不会立刻跳到你所要的位置，而是加快播放速度，直到到达目标位置。

目前遇到一个问题：当我将索引加进去avi以后，播放器需要有一个缓冲的过程，有点像用迅雷下载电影，可是没等电影下完，改过后缀就去播放的状况。问题解决以后再来补充吧。

经过对avi视频的详细分析，发现音频视频的交叉有如下规律：

15帧 vedio

1段 audio

15帧 vedio

1段 audio

16帧 vedio

1段 audio

15帧 vedio

1段 audio

15帧 vedio

1段 audio

16帧 vedio

1段 audio

也就是基本上是每15段vedio就有一段的audio，可是每经过两个15段的vedio循环，就要有16段的vedio才能有一段audio。另外还有一个规律就是，每一段audio大小都是一定的4096字节。这是某个特定文件的交叉规律，个人认为音视频交叉的频率跟帧率和音频频率有关，只要是按一定的比率来交叉就行，例如上面按每30帧视频放一段音频，然后音频的长度变为8192应该可以，有空再试试。

#### 6.2.5 参考链接

1.AVI文件格式详解

<https://blog.csdn.net/chenyonken/article/details/79174500>

2.avi文件格式解析

<https://blog.csdn.net/occupy8/article/details/40678471>

## MP4封装格式

### 7.1 MP4概述

最开始MP4指的是音频（MP3的升级版），即MPEG-2 AAC标准。随后MP4概念被转移到视频上，对应的是MPEG-4标准。而现在我们流行的叫法，多半是指能播放MPEG-4标准编码格式视频的播放器。但是这篇文章介绍的内容跟上面这些都无关，我们要讨论的是MP4文件封装格式，对应的标准为ISO/IEC 14496-12，即信息技术 视听对象编码的第12部分：ISO 基本媒体文件格式（Information technology Coding of audio-visual objects Part 12: ISO base media file format）。ISO/IEC组织指定的标准一般用数字表示，ISO/IEC 14496即MPEG-4标准。

MP4视频文件封装格式是基于QuickTime容器格式定义的，因此参考QuickTime的格式定义对理解MP4文件格式很有帮助。MP4文件格式是一个十分开放的容器，几乎可以用来描述所有的媒体结构，MP4文件中的媒体描述与媒体数据是分开的，并且媒体数据的组织也很自由，不一定要按照时间顺序排列，甚至媒体数据可以直接引用其他文件。同时，MP4也支持流媒体。MP4目前被广泛用于封装h.264视频和AAC音频，是高清视频的代表。

#### 参考链接

1.MP4文件格式

<https://www.cnblogs.com/ranson7zop/p/7889272.html>

2.MP4文件格式解析

<https://www.cnblogs.com/boonya/p/8572052.html>

3.mp4文件格式解析

<https://www.jianshu.com/p/529c3729f357>

### 7.2 MP4结构

MP4文件中的所有数据都装在box（QuickTime中为atom）中，也就是说MP4文件由若干个box组成，每个box有类型和长度，可以将box理解为一个数据对象块。box中可以包含另一个box，这种box称为container box。一个MP4文件首先会有且只有一个“ftyp”类型的box，作为MP4格式的标志并包含关于文件的一些信息；之后会有且只有一个“moov”类型的box（Movie Box），它是一种container box，子box包含了媒体的metadata信息；MP4文件的媒体数据包含在“mdat”类型的box（Midia Data Box）中，该类型的box也是container box，可以有多个，也可以没有（当媒体数据全部引用其他文件时），媒体数据的结构由metadata进行描述。

下面是一些概念：

**track**表示一些sample的集合，对于媒体数据来说，track表示一个视频或音频序列。

**hint track** 这个特殊的track并不包含媒体数据，而是包含了一些将其他数据track打包成流媒体的指示信息。

**sample**对于非hint track来说，video sample即为一帧视频，或一组连续视频帧，audio sample即为一段连续的压缩音频，它们统称sample。对于hint track，sample定义一个或多个流媒体包的格式。

**sample table**指明sampe时序和物理布局的表。

**chunk** 一个track的几个sample组成的单元。

在本文中，我们不讨论涉及hint的内容，只关注包含媒体数据的本地MP4文件。下图为一个典型的MP4文件的结构树。



首先需要说明的是，box中的字节序为网络字节序，也就是大端字节序（Big-Endian），简单的说，就是一个32位的4字节整数存储方式为高位字节在内存的低端。Box由header和body组成，其中header统一指明box的大小和类型，body根据类型有不同的意义和格式。

标准的box开头的4个字节（32位）为box size，该大小包括box header和box body整个box的大小，这样我们就可以在文件中定位各个box。如果size为1，则表示这个box的大小为large size，真正的size值要在largesize域上得到。（实际上只有“mdat”类型的box才有可能用到large size。）如果size为0，表示该box为文件的最后一个box，文件结尾即为该box结尾。（同样只存在于“mdat”类型的box中。）

size后面紧跟的32位为box type，一般是4个字符，如“ftyp”、“moov”等，这些box type都是已经预定义好的，分别表示固定的意义。如果是“uuid”，表示该box为用户扩展类型。如果box type是未定义的，应该将其忽略。

**File Type Box（ftyp）**：该box有且只有1个，并且只能被包含在文件层，而不能被其他box包含。该box应该被放在文件的最开始，指示该MP4文件应用的相关信息。“ftyp” body依次包括1个32位的major brand（4个字符），1个32位的minor version（整数）和1个以32位（4个字符）为单位元素的数组compatible brands。这些都是用来指示文件应用级别的信息。该box的字节实例如下：

IMG_256

**Movie Box（moov）**：该box包含了文件媒体的metadata信息，“moov”是一个container box，具体内容信息由子box诠释。同File Type Box一样，该box有且只有一个，且只被包含在文件层。一般情况下，“moov”会紧随“ftyp”出现。一般情况下（限于篇幅，本文只讲解常见的MP4文件结构），“moov”中会包含1个“mvhd”和若干个“trak”。其中“mvhd”为header box，一般作为“moov”的第一个子box出现（对于其他container box来说，header box都应作为首个子box出现）。“trak”包含了一个track的相关信息，是一个container box。下图为部分“moov”的字节实例，其中红色部分为box header，绿色为“mvhd”，黄色为一部分“trak”。



**Movie Header Box（mvhd）：**“mvhd”结构如下表。

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 mvhd大小 |
| box type | 4 | box类型 “mvhd” |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 |  |
| creation time | 4 | 创建时间（相对于UTC时间1904-01-01零点的秒数） |
| modification time | 4 | 修改时间 |
| time scale | 4 | 文件媒体在1秒时间内的刻度值，可以理解为1秒长度的时间单元数 |
| duration | 4 | 该track的时间长度，用duration和time scale值可以计算track时长，比如audio track的time scale = 8000, duration = 560128，时长为70.016，video track的time scale = 600, duration = 42000，时长为70 |
| rate | 4 | 推荐播放速率，高16位和低16位分别为小数点整数部分和小数部分，即[16.16] 格式，该值为1.0（0x00010000）表示正常前向播放 |
| volume | 2 | 与rate类似，[8.8] 格式，1.0（0x0100）表示最大音量 |
| reserved | 10 | 保留位 |
| matrix | 36 | 视频变换矩阵 |
| pre-defined | 24 |  |
| next track id | 4 | 下一个track使用的id号 |

“mvhd”的字节实例如下图，各字段已经用颜色区分开：



**Track Box（trak）：**“trak”也是一个container box，其子box包含了该track的媒体数据引用和描述（hint track除外）。一个MP4文件中的媒体可以包含多个track，且至少有一个track，这些track之间彼此独立，有自己的时间和空间信息。“trak”必须包含一个“tkhd”和一个“mdia”，此外还有很多可选的box（略）。其中“tkhd”为track header box，“mdia”为media box，该box是一个包含一些track媒体数据信息box的container box。 “trak”的部分字节实例如下图，其中黄色为“trak”box的头，绿色为“tkhd”，蓝色为“edts”（一个可选box），红色为一部分“mdia”。



**Track Header Box（tkhd）：**“tkhd”结构如下表。

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 **tkhd大小** |
| box type | 4 | box类型 “**tkhd”** |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 | 按位或操作结果值，预定义如下：  0x000001 track\_enabled，否则该track不被播放；  0x000002 track\_in\_movie，表示该track在播放中被引用；  0x000004 track\_in\_preview，表示该track在预览时被引用。  一般该值为7，如果一个媒体所有track均未设置track\_in\_movie和track\_in\_preview，将被理解为所有track均设置了这两项；对于hint track，该值为0 |
| creation time | 4 | 创建时间（相对于UTC时间1904-01-01零点的秒数） |
| modification time | 4 | 修改时间 |
| track id | 4 | id号，不能重复且不能为0 |
| reserved | 4 | 保留位 |
| duration | 4 | track的时间长度 |
| reserved | 8 | 保留位 |
| layer | 2 | 视频层，默认为0，值小的在上层 |
| alternate group | 2 | track分组信息，默认为0表示该track未与其他track有群组关系 |
| volume | 2 | [8.8] 格式，如果为音频track，1.0（0x0100）表示最大音量；否则为0 0要不音量为0要不就是指定track视频 |
| reserved | 2 | 保留位 |
| matrix | 36 | 视频变换矩阵 |
| width | 4 | 宽 |
| height | 4 | 高，均为 [16.16] 格式值，与sample描述中的实际画面大小比值，用于播放时的展示宽高 |

“tkhd”的字节实例如下图，各字段已经用颜色区分开：



**Media Box（mdia）：** “mdia”也是个container box，其子box的结构和种类还是比较复杂的。先来看一个“mdia”的实例结构树图。

总体来说，“mdia”定义了track媒体类型以及sample数据，描述sample信息。一般“mdia”包含一个“mdhd”，一个“hdlr”和一个“minf”，其中“mdhd”为media header box，“hdlr”为handler reference box，“minf”为media information box。下面依次看一下这几个box的结构。

**Media Header Box（mdhd）：**“mdhd”结构如下表。

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 |
| box type | 4 | box类型 |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 |  |
| creation time | 4 | 创建时间（相对于UTC时间1904-01-01零点的秒数） |
| modification time | 4 | 修改时间 |
| time scale | 4 | 同前表 |
| duration | 4 | track的时间长度 |
| language | 2 | 媒体语言码。最高位为0，后面15位为3个字符（见ISO 639-2/T标准中定义） |
| pre-defined | 2 |  |

“mdhd”的字节实例如下图，各字段已经用颜色区分开：

IMG_256

**Handler Reference Box（hdlr）**：“hdlr”解释了媒体的播放过程信息，该box也可以被包含在meta box（meta）中。“hdlr”结构如下表。

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 |
| box type | 4 | box类型 |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 |  |
| pre-defined | 4 |  |
| handler type | 4 | 在media box中，该值为4个字符：  “vide”— video track  “soun”— audio track  “hint”— hint track |
| reserved | 12 |  |
| name | 不定 | track type name，以‘\0’结尾的字符串 |

“hdlr”的字节实例如下图，各字段已经用颜色区分开：



**Media Information Box（minf）**：“minf”存储了解释track媒体数据的handler-specific信息，media handler用这些信息将媒体时间映射到媒体数据并进行处理。“minf”中的信息格式和内容与媒体类型以及解释媒体数据的media handler密切相关，其他media handler不知道如何解释这些信息。“minf”是一个container box，其实际内容由子box说明。一般情况下，“minf”包含一个header box，一个“dinf”和一个“stbl”，其中，header box根据track type（即media handler type）分为“vmhd”、“smhd”、“hmhd”和“nmhd”，“dinf”为data information box，“stbl”为sample table box。下面分别介绍。下图为“minf”部分字节实例，其中红色为box header，蓝色为“smhd”，绿色为“dinf”，黄色为一部分“stbl”。



**Media Information Header Box（vmhd、smhd、hmhd、nmhd）**

Video Media Header Box（vmhd）

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 |
| box type | 4 | box类型 |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 |  |
| graphics mode | 4 | 视频合成模式，为0时拷贝原始图像，否则与opcolor进行合成 |
| opcolor | 2×3 | ｛red，green，blue｝ |

Sound Media Header Box（smhd）

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 |
| box type | 4 | box类型 |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 |  |
| balance | 2 | 立体声平衡，[8.8] 格式值，一般为0，-1.0表示全部左声道，1.0表示全部右声道 |
| reserved | 2 |  |

Hint Media Header Box（hmhd）

略

Null Media Header Box（nmhd）

非视音频媒体使用该box，略。

**Data Information Box（dinf）**：“dinf”解释如何定位媒体信息，是一个container box。“dinf”一般包含一个“dref”，即data reference box；“dref”下会包含若干个“url”或“urn”，这些box组成一个表，用来定位track数据。简单的说，track可以被分成若干段，每一段都可以根据“url”或“urn”指向的地址来获取数据，sample描述中会用这些片段的序号将这些片段组成一个完整的track。一般情况下，当数据被完全包含在文件中时，“url”或“urn”中的定位字符串是空的。

“dref”的字节结构如下表。

|  |  |  |
| --- | --- | --- |
| 字段 | 字节数 | 意义 |
| box size | 4 | box大小 |
| box type | 4 | box类型 |
| version | 1 | box版本，0或1，一般为0。（以下字节数均按version=0） |
| flags | 3 |  |
| entry count | 4 | “url”或“urn”表的元素个数 |
| “url”或“urn”列表 | 不定 |  |

“url”或“urn”都是box，“url”的内容为字符串（location string），“urn”的内容为一对字符串（name string and location string）。当“url”或“urn”的box flag为1时，字符串均为空。

下面是一个“dinf”的字节实例图。其中黄色为“dinf”的box header，由红色部分我们知道包含的“url”或“urn”个数为1，红色后面为“url”box的内容。紫色为“url”的box header（根据box type我们知道是个“url”），绿色为box flag，值为1，说明“url”中的字符串为空，表示track数据已包含在文件中。

IMG_256

**Sample Table Box（stbl）**： “stbl”几乎是普通的MP4文件中最复杂的一个box了，首先需要回忆一下sample的概念。sample是媒体数据存储的单位，存储在media的chunk中，chunk和sample的长度均可互不相同，如下图所示。



“stbl”包含了关于track中sample所有时间和位置的信息，以及sample的编解码等信息。利用这个表，可以解释sample的时序、类型、大小以及在各自存储容器中的位置。“stbl”是一个container box，其子box包括：sample description box（stsd）、time to sample box（stts）、sample size box（stsz或stz2）、sample to chunk box（stsc）、chunk offset box（stco或co64）、composition time to sample box（ctts）、sync sample box（stss）等。

“stsd”必不可少，且至少包含一个条目，该box包含了data reference box进行sample数据检索的信息。没有“stsd”就无法计算media sample的存储位置。“stsd”包含了编码的信息，其存储的信息随媒体类型不同而不同。

**Sample Description Box（stsd）**

box header和version字段后会有一个entry count字段，根据entry的个数，每个entry会有type信息，如“vide”、“sund”等，根据type不同sample description会提供不同的信息，例如对于video track，会有“VisualSampleEntry”类型信息，对于audio track会有“AudioSampleEntry”类型信息。视频的编码类型、宽高、长度，音频的声道、采样等信息都会出现在这个box中。



**Time To Sample Box（stts）**

“stts”存储了sample的duration，描述了sample时序的映射方法，我们通过它可以找到任何时间的sample。“stts”可以包含一个压缩的表来映射时间和sample序号，用其他的表来提供每个sample的长度和指针。表中每个条目提供了在同一个时间偏移量里面连续的sample序号，以及samples的偏移量。递增这些偏移量，就可以建立一个完整的time to sample表。



**Sample Size Box（stsz）**

“stsz” 定义了每个sample的大小，包含了媒体中全部sample的数目和一张给出每个sample大小的表。这个box相对来说体积是比较大的。四个字节其实就是每一帧得大小。



**Sample To Chunk Box（stsc）**

用chunk组织sample可以方便优化数据获取，一个thunk包含一个或多个sample。“stsc”中用一个表描述了sample与chunk的映射关系，查看这张表就可以找到包含指定sample的thunk，从而找到这个sample。



**Sync Sample Box（stss）**

“stss”确定media中的关键帧。对于压缩媒体数据，关键帧是一系列压缩序列的开始帧，其解压缩时不依赖以前的帧，而后续帧的解压缩将依赖于这个关键帧。“stss”可以非常紧凑的标记媒体内的随机存取点，它包含一个sample序号表，表内的每一项严格按照sample的序号排列，说明了媒体中的哪一个sample是关键帧。如果此表不存在，说明每一个sample都是一个关键帧，是一个随机存取点。四个字节其实和fps值也有关系。



**Chunk Offset Box（stco）**

“stco”定义了每个thunk在媒体流中的位置。位置有两种可能，32位的和64位的，后者对非常大的电影很有用。在一个表中只会有一种可能，这个位置是在整个文件中的，而不是在任何box中的，这样做就可以直接在文件中找到媒体数据，而不用解释box。需要注意的是一旦前面的box有了任何改变，这张表都要重新建立，因为位置信息已经改变了。



主要归纳为：co64指定每个chunk在媒体流中的位置。Mdata数据即为数据，占绝大空间。Stsc指定chunk中包含多少sample。Stsz指定每个sample得大小。Stss指定关键帧sample得位置即fps值。Stts即dts-sample映射。



ftyp：表明文件类型。

mdat：具体的媒体数据。

moov box：包含一系列的子box， 通过解析这些子box，获得编码信息、数据信息等。

Free Space Box（free或skip）**：**“free”中的内容是无关紧要的，可以被忽略。该box被删除后，不会对播放产生任何影响。

Meida Data Box（mdat）：该box包含于文件层，可以有多个，也可以没有（当媒体数据全部为外部文件引用时），用来存储媒体数据。数据直接跟在box type字段后面，具体数据结构的意义需要参考metadata（主要在sample table中描述）。普通MP4文件的结构就讲完了，可能会比较乱，下面这张图是常见的box的树结构图，可以用来大致了解MP4文件的构造。

### 7.3 MP4术语解释及知识补充

box：由唯一类型标识符和长度定义的面向对象的构件

container box：用来容纳一组相关box的box，container box通常都不是fullbox

chunk：同一轨道的一组连续的采样

hint track：不包含媒体数据，但包含了将一个或多个轨打包到流频道的指示

media data box：用来容纳实体数据的box

movie box：子box定义了元数据（metadata）的容器box

sample：与单个时间戳相关联的所有数据，video sample即为一帧视频，或一组连续视频帧，audio sample即为一段连续的压缩音频

sample description：定义和描述轨中的采样的格式的结构

sample table：指明sampe时序和物理布局的表

track：按时间排序的相关的采样，对于媒体数据来说，track表示一个视频或音频序列

整个box以box header开头，box header中包含了box的大小（size）和类型（type）等信息。其中，size指明了整个box所占用的大小，包括header部分，如果box很大(例如存放具体视频数据的mdat box)，超过了uint32的最大数值，size就被设置为1，并用接下来的8位uint64的largesize来存放大小。box中的字节序为网络字节序，也就是大端字节序（Big-Endian）。

“stbl”包含了关于track中sample所有时间和位置的信息，以及sample的编解码等信息。利用这个表，可以解释sample的时序、类型、大小以及在各自存储容器中的位置。“stbl”是一个container box，其子box包括：sample description box（stsd）、time to sample box（stts）、sample size box（stsz或stz2）、sample to chunk box（stsc）、chunk offset box（stco或co64）、composition time to sample box（ctts）、sync sample box（stss）等。“stsd”必不可少，且至少包含一个条目，该box包含了data reference box进行sample数据检索的信息。没有“stsd”就无法计算media sample的存储位置。“stsd”包含了编码的信息，其存储的信息随媒体类型不同而不同。

### 7.4 Mp4info工具各个参数介绍：

mp4文件基本信息

audio信息：

smplrate：sample rate(采样率)。

channel：通道个数。

bitrate：比特率。

audiosamplenum：音频sample的个数。

video信息：

width、height：视频的宽/高。

bitrate：比特率(码率)，秒为单位。等于视频总的大小/时长。

frames：视频帧数。

fps：帧率(frame per second)。

total\_time：时间长度，ms为单位。等于duration/timescale。

timescale：时间的粒度，1000表示1000个单位为1s。

duration：时间粒度的个数。

videosamplenum：视频sample的个数。

### 7.5MP4中如何seek

我们需要seek到30s。  
需要做如下工作：

1. 使用timescale将目标时间标准化。timescale为90000，30\*90000=2700000。
2. 通过time-to-sample box(stts)找到指定track的给定时间之前的第一个sample number。2700000/3000 = 900。900应该是一帧大概占用多少timescale=90000/100。
3. 通过sync sample table(stss)查询sample number之前的第一个sync sample。对应为795的sample。
4. 通过sample-to-chunk table(stsc)查找到对应的chunk number。对应的chunk号是假设是400。
5. 通过chunk offset box(stco或co64)查找到对应chunk在文件中的起始偏移量。第14个chunk的offset是3481072。
6. 最后使用sample-to-chunk box(stsc)和sample size box(stsz)的信息计算出该chunk中需要读取的sample在文件中的起始偏移量，即完成 seek。