Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42^{nd} Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

ISO/IEC MPEG-4 High-Definition Scalable Advanced Audio Coding

Ralf Geiger[1], Rongshan Yu[2], Jürgen Herre[1], Susanto Rahardja[2], Sang-Wook Kim[3], Xiao Lin[2], Markus Schmidt[1]

[1]*Fraunhofer IIS, Erlangen, Germany*

[2]*Institute for Infocomm Research, Singapore*

[3]*Samsung Electronics, Suwon, Korea*

Correspondence should be addressed to Ralf Geiger (ralf.geiger@iis.fraunhofer.de)

ABSTRACT

Recently, the MPEG Audio standardization group has successfully concluded the standardization process on technology for lossless coding of audio signals. This paper provides a summary of the Scalable Lossless Coding (SLS) technology as one of the results of this standardization work. MPEG-4 Scalable Lossless Coding provides a fine-grain scalable lossless extension of the well-known MPEG-4 AAC perceptual audio coder up to fully lossless reconstruction at word lengths and sampling rates typically used for high-resolution audio. The underlying innovative technology is described in detail and its performance is characterized for lossless and near-lossless representation, both in conjunction with an AAC coder and as a stand-alone compression engine. A number of application scenarios for the new technology are discussed finally.

1. INTRODUCTION

Perceptual coding of high-quality audio signals has experienced a tremendous evolution over the past two decades, both in terms of research progress and in worldwide deployment in products. Examples for successful applications include portable music players, Internet audio, audio for digital media (e.g. VCD, DVD) and digital broadcasting. Several phases of international standardization have been conducted successfully [1, 2, 16]. Many recent research and standardization efforts focus at achieving good sound quality at still lower bit-rates [4, 5, 6, 7, 8, 9] to accommodate storage/transmission channels with very limited quality (e.g. terrestrial and satellite broadcasting; 3rd generation mobile telecommunication). Nevertheless, for other scenarios with higher transmission bandwidth available, there is a general trend towards providing to the consumer a media experience with extremely high fidelity [10], as it is frequently associated with the terms "High Definition" and "High Resolution" and dedicated media types, such as DVD-Audio, SACD, HD-DVD, or Blue-Ray Disc. In the realm of audio, this is achieved by employing lossless formats with high resolution (word length) and/or high sampling rate.

In this context, the ISO/MPEG Audio standardization group decided to start a new work item exploring technology for lossless and near lossless coding of audio signals by issuing a Call for Proposals for relevant technology in 2002 [3]. Three specifications emerged from this call as amendments to the MPEG-4 Audio standard: Firstly, the standard on lossless coding of one-bit oversampled signals [12] specifies the lossless compression of highly oversampled 1-bit sigma-delta modulated audio signals as they are stored on the SACD media under the name Direct Stream Digital (DSD) [11]. Secondly, the Audio Lossless Coding (ALS) specification [13] describes technology for the lossless coding of PCM coded signals at sampling rates up to 192kHz as well as floating point audio.

This article provides an overview of the third descendant, i.e. the "Scalable Lossless Coding" (SLS) specification [14] which extends traditional methods for perceptual audio coding towards lossless coding of high-resolution audio signals in a scalable way. Specifically, it allows to scale up from a perceptually coded representation (MPEG-4 Advanced Audio Coding, AAC) to a fully lossless representation with high definition, including a wide range of intermediate near-lossless representations. The article starts with an explanation of the general concept, discusses the underlying novel technology components, and characterizes the codec in terms of compression performance and complexity. Finally, a number of application scenarios are briefly outlined.

2. CONCEPT AND TECHNOLOGY OVERVIEW

This section provides an overview of the principle and structure of the SLS technology and its combination with the AAC perceptual audio coder. This combination will be referred to as "High Definition Advanced Audio Coding" (HD-AAC) in the following.

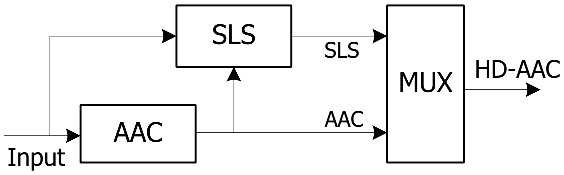

Fig. 1: HD-AAC Encoder

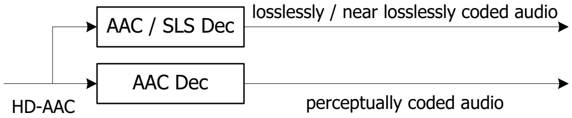

Fig. 2: HD-AAC Decoder

2.1. System structure

Fig. 1 shows a very high-level view of the structure of the HD-AAC encoder. The input signal is first coded by an AAC encoder. Then, the SLS uses the output to enhance the system's performance towards lossless coding. The two layers are then multiplexed into one high definition bitstream. The HD-AAC decoder is illustrated in Fig. 2. From the HD-AAC bitstream the decoder is able to decode either the perceptually coded AAC part only, or use the additional SLS information to produce high-definition audio. The signal representation accuracy can scale up to lossless.

2.2. AAC Background

Since the MPEG-4 SLS coder was designed to operate as an enhancement to the MPEG-4 AAC, its structure is closely related to that of the underlying AAC core coder [37]. This section briefly describes the architectural features of MPEG-4 AAC.

The underlying AAC codec provides efficient perceptual audio coding with high quality and achieves broadcast quality at a bitrate of ca. 64 kbit/s per

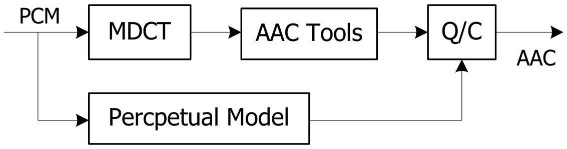

Fig. 3: Structure of an AAC encoder (simplified)

channel [18]. Fig. 3 shows a very condensed view of the AAC encoder structure. The audio signal is handled in a block-wise spectral representation using the Modified Discrete Cosine Transform (MDCT) [19]. The resulting 1024 spectral values are quantized and coded considering the required accuracy, as demanded by a perceptual model. This is done to minimize the perceptibility of the introduced quantization distortion by exploiting masking effects. Several neighboring spectral values are grouped into so-called Scalefactor Bands (sfbs) sharing the same scalefactor for quantization. Prior to the quantization/coding tool, a number of processing tools operate on the spectral coefficients in order to improve coding performance for certain situations, most importantly:

- The Temporal Noise Shaping (TNS) tool [20] carries out predictive filtering across frequency in order to achieve a temporal shaping of the quantization noise according to the signal envelope and in this way optimize temporal masking.
- The M/S Stereo Coding tool [21] provides sum/difference coding of channel pairs, exploits inter-channel redundancy for near-monophonic signals, and protects from binaural unmasking.

2.3. Scalable to Lossless Enhancement

The scalable to lossless enhancement is achieved using SLS codec. The more specific structures of an HD-AAC encoder and decoder are shown in Figs. 4 and 5.

In SLS encoder, the audio signal is represented in frequency domain using the Integer Modified Discrete Cosine Transform (IntMDCT) [22, 23]. This transform provides an invertible integer approximation of the MDCT and is well-suited for lossless coding in frequency domain. Other AAC coding tools, such as TNS or M/S Coding, are also considered and performed on the IntMDCT spectral coefficients in an invertible integer fashion, thus maintaining the similarity between the spectral values used in the AAC coder and in the lossless enhancement.

The link between the perceptual core and the scalable lossless enhancement layer of the coder [26] is provided by an error mapping process. In principle, the error mapping process removes the information that has already been coded in perceptual part from the IntMDCT spectral coefficients so that only the resulted IntMDCT residuals are coded in the SLS encoder. In addition, this error mapping process also preserves the probability distribution skew of the original IntMDCT coefficients so that they can be very efficiently coded by the scalable coding process used in the enhancement layer. Specifically, the scalable coding process is bit-plane coding with several entropy coding schemes include Bit-plane Golomb Code (BPGC)[28], Context-based Arithmetic Coding (CBAC) and low energy mode coding [29].

In the following sections, the principles underlying the design of SLS that include IntMDCT filterbank, error mapping strategy and scalable coding tools will be described in details.

3. FILTERBANK

3.1. IntMDCT

The Integer Modified Discrete Cosine Transform (IntMDCT), introduced in [22], is an invertible integer approximation of the Modified Discrete Cosine Transform (MDCT), which is obtained by utilizing the "Lifting Scheme" [42], or "Ladder Network" [43].

The IntMDCT allows efficient lossless coding [22]. Furthermore, as the IntMDCT closely approximates the MDCT, it allows to combine the techniques of perceptual and lossless audio coding in frequency domain [26].

Figure 6 illustrates this close approximation for a typical audio signal portion. The difference between the MDCT and the IntMDCT values only shows up as a noise floor, which is typically much lower than the error introduced in perceptual coding. Thus, the IntMDCT allows to efficiently code the quantization error of an MDCT-based perceptual codec in frequency domain.

In the following, some details on how to derive the IntMDCT from the MDCT are explained.

3.2. Decomposition of MDCT

The MDCT, defined by

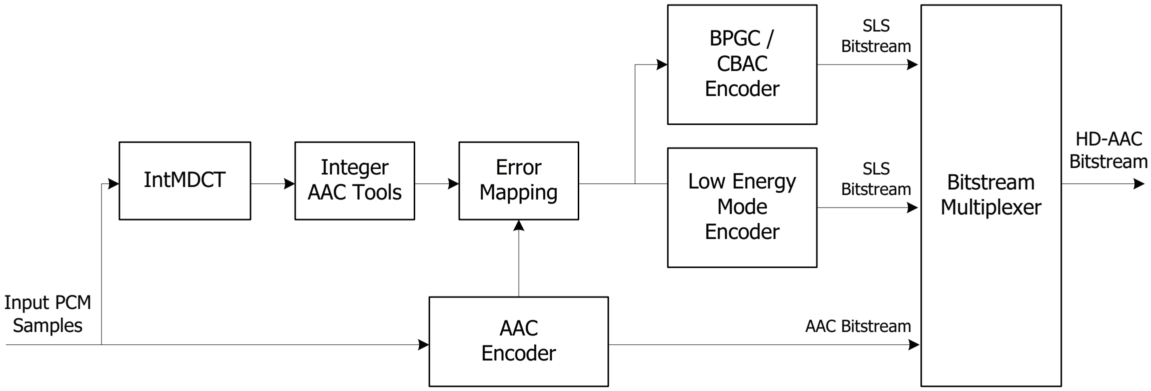

Fig. 4: Structure of HD-AAC encoder

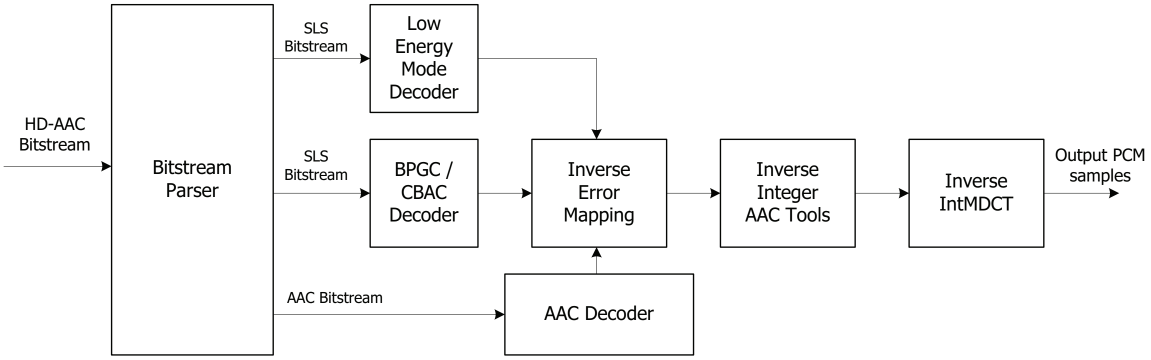

Fig. 5: Structure of HD-AAC decoder

$$X(m) = \sqrt{\frac{2}{N}} \sum_{k=0}^{2N-1} w(k)x(k)\cos\frac{(2k+1+N)(2m+1)\pi}{4N} \qquad m = 0, ..., N-1 \tag{1}$$

with the time domain input $x(k)$ and the windowing function $w(k)$, is decomposed into two blocks:

- Windowing and Time Domain Aliasing (TDA)
- Discrete Cosine Transform of Type IV (DCT-IV)

This is illustrated in Figure 7 for the MDCT and the inverse MDCT.

In the forward IntMDCT the Windowing/TDA block is calculated by $3N/2$ so-called *lifting steps*:

$$\begin{pmatrix} x(k) \\ x(N-1-k) \end{pmatrix} \mapsto \begin{pmatrix} 1 & -\frac{w(N-1-k)-1}{w(k)} \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -w(k) & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -\frac{w(N-1-k)-1}{w(k)} \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x(k) \\ x(N-1-k) \end{pmatrix} \qquad k = 0, ..., \frac{N}{2}-1 \tag{2}$$

After each lifting step, a rounding operation is applied to stay in the integer domain. Every lifting step can be inverted by simply adding the subtracted value, and vice versa.

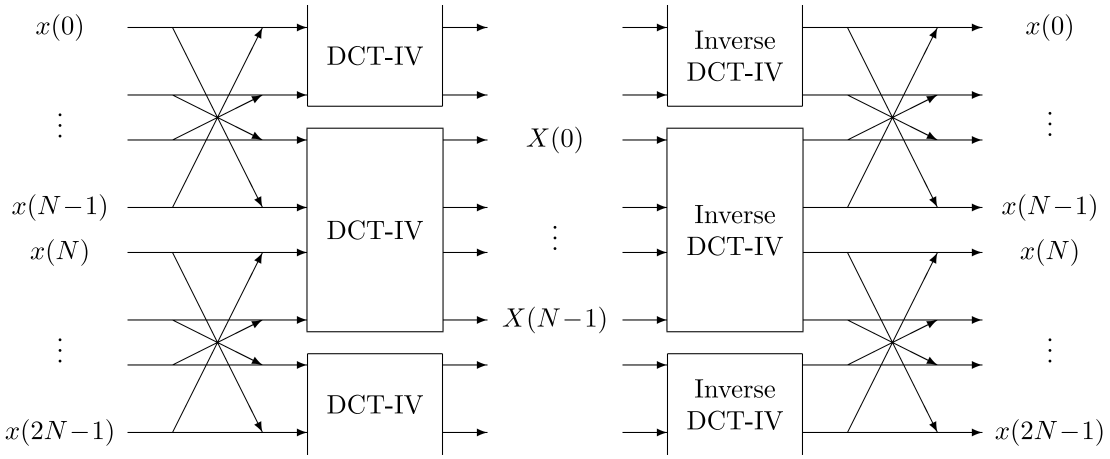

Fig. 7: MDCT and inverse MDCT by Windowing/TDA and DCT-IV

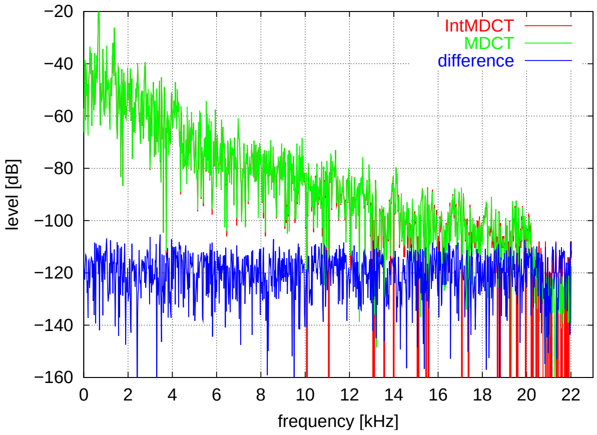

Fig. 6: IntMDCT and MDCT magnitude spectra

3.3. Integer DCT-IV

For the IntMDCT, the DCT-IV is calculated in an invertible integer fashion, called the Integer DCT-IV. The Multi-Dimensional Lifting (MDL) Scheme [23, 24] is applied in order to reduce the required rounding operations in the invertible integer approximation as much as possible.

The following block matrix decomposition for an invertible matrix T and the identity matrix I shows the basic principle behind the MDL scheme:

$$\begin{pmatrix} T & 0 \\ 0 & T^{-1} \end{pmatrix} = \begin{pmatrix} -I & 0 \\ T^{-1} & I \end{pmatrix} \begin{pmatrix} I & -T \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & I \\ I & T^{-1} \end{pmatrix} \tag{3}$$

The three blocks in this decomposition are so-called *Multi-Dimensional Lifting Steps*. Similar to the conventional lifting steps, they can be transferred to invertible integer mappings by rounding the floating point values after being processed by T resp. T^{-1}, and they can be inverted by subtracting the values that have been added.
In case of stereo signals this decomposition is used to obtain an integrated calculation of the M/S matrix and the Integer DCT-IV for the left and the right channel. The number of required rounding operations is $3N$ per channel pair, i.e. $3N/2$ per channel, which is the same number as for the Windowing/TDA stage. Overall, the Stereo IntMDCT including M/S requires only 3 rounding operations per sample. In case of mono signals the same structure can be utilized. It only has to be extended by some additional lifting steps to obtain the Integer DCT-IV of one block, see [24]. For this Mono IntMDCT 4 rounding operations per sample are required.

3.4. Noise Shaping

The lossless coding efficiency of the IntMDCT is further improved by utilizing a noise shaping technique,

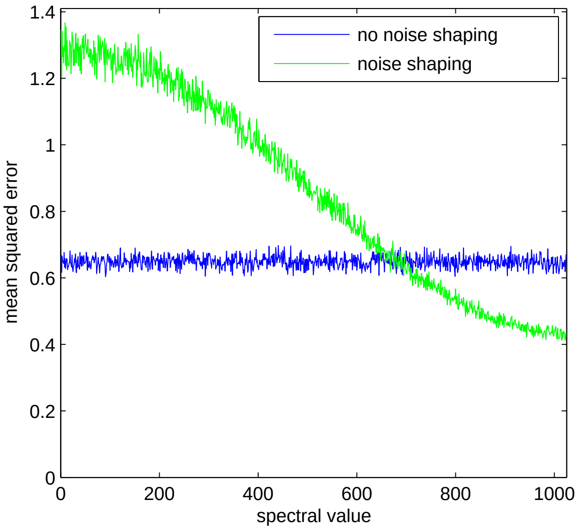

Fig. 8: Mean-squared approximation error of Stereo IntMDCT (including M/S) with and without noise shaping

introduced in [25]. In the lifting steps where time domain signals are processed, the rounding operations are connected to an error feedback to provide a spectral shaping of the approximation noise. This approximation noise affects the lossless coding efficiency mainly in the high frequency region where audio signals usually contain a very small amount of energy, especially at sampling rates of 96 kHz and higher. Hence, a low-pass characteristic of the approximation noise improves the lossless coding efficiency. A first-order noise shaping filter is used. For the IntMDCT this filter is applied in the three stages of lifting steps in the Windowing/TDA processing and in the first rounding stage of the Integer DCT-IV processing. Figure 8 compares the resulting approximation error between the IntMDCT values and the MDCT values rounded to integer, where the IntMDCT operates both with and without noise shaping.

4. ERROR MAPPING / RESIDUAL CALCULATION

The purpose of error mapping process is to remove the information that has already been coded in the AAC core layer. The generated residual signals will be further coded in the enhancement coding layer.

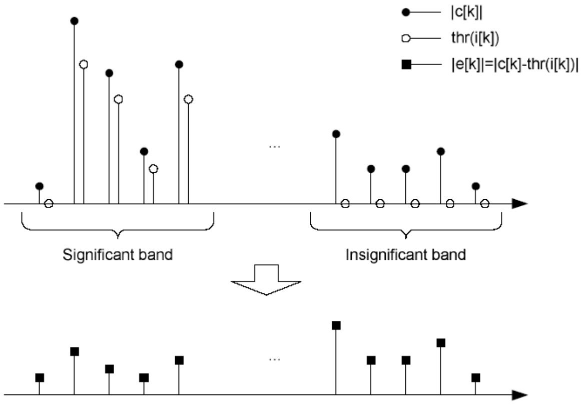

Fig. 9: Illustration of the error mapping process.

The residual signals $e[k]$, $k = 0, ..., N-1$ where N is the length of IntMDCT spectrum, are calculated as

$$e[k] = \begin{cases} c[k] - thr(i[k]), & i[k] \neq 0 \\ c[k], & i[k] = 0 \end{cases} \quad (4)$$

where $c[k]$ is the IntMDCT coefficient, $i[k]$ is the AAC quantized value and $thr(i[k])$ is the quantization thresholds closer to zero of $i[k]$. This error mapping process is illustrated in Fig. 9, where two different cases are given. In the first case, the IntMDCT coefficients $c[k]$ are from an sfb that has been quantized and coded at the AAC core encoder (Significant band). For these coefficients, the residual coefficients are obtained by subtracting the quantization thresholds from their original IntMDCT coefficients, resulting in a residual spectrum with reduced amplitude. In the other case $c[k]$ are from an sfb that is not coded in the AAC core encoder (Insignificant band). In this scenario, the residual spectrum is in fact the IntMDCT spectrum itself.

5. SCALABLE CODING

5.1. Scalable Coding Tools

5.1.1. Bit-Plane Coding

In SLS codec, the bit-plane coding technology is used in coding the residual signals to generate the Lossless Enhancement (LLE) bitstream.

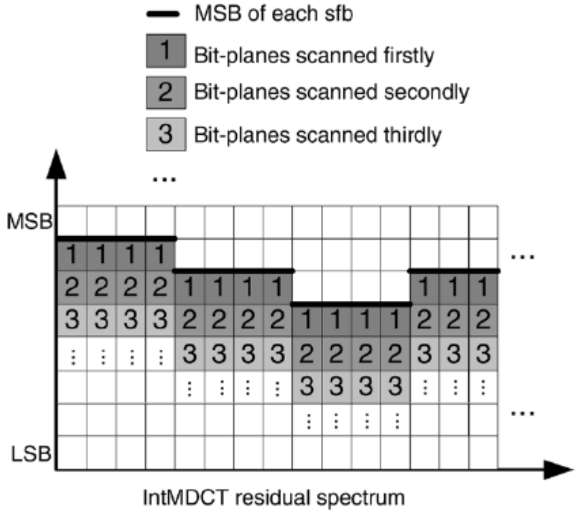
MSB of each sfb
1 Bit-planes scanned firstly
2 Bit-planes scanned secondly
3 Bit-planes scanned thirdly
...
MSB
LSB
IntMDCT residual spectrum

Fig. 10: Bit-plane scan process in SLS

The residual signals $e[k]$, $k = 1, ..., N-1$ where N is the dimension of the input signals, is first represented in a binary format as

$$e[k] = (2s[k] - 1) \sum_{j=0}^{M-1} b[k,j] \cdot 2^j \tag{5}$$

which comprises of a sign symbol

$$s[k] \triangleq \begin{cases} 1, & e[k] \geqslant 0 \\ 0, & e[k] < 0 \end{cases} \tag{6}$$

and bit-plane symbols $b[k,j] \in \{0,1\}$ with M is the MSB for $e[k]$.

By using the bit-plane coding for the above format residuals, the lossless enhancement is performed in a fine-grain scalable way. A lossless reconstruction is obtained if all the bit-planes of the residual are coded and transmitted completely; while it is still possible to obtain a high-quality lossy reconstruction if only part of the bit-planes are decoded. To achieve optimal perceptual quality at intermediate bit-rates, the bit-plane coding is started from the Most Significant Bit (MSB) for all sfbs, and progressed to subsequent bit-planes until it reaches the Least Significant Bit (LSB) for all sfbs (Fig. 10). In such a way the spectral shape of the core layer quantization noise, which has been shaped by the noise shaping process of the core layer perceptual coder, is preserved during the bit-plane coding process.

5.1.2. Bit-Plane Coding with BPGC/CBAC

The bit-plane symbols are then coded with arithmetic code with fixed frequency tables. There are two different types of frequency tables used in SLS codec. The first one is followed by BPGC frequency assignment which is derived from the statistical properties of a geometrically distributed source. In BPGC, a symbol at bit-plane j is coded with probability assignment Q_j^L given by

$$Q^L[j] = \begin{cases} \frac{1}{2}, & j < L \\ \frac{1}{1+2^{2^{j-L}}}, & j > L \end{cases} \tag{7}$$

where the Lazy Plane parameter L can be selected using the adaptation rule [28].

To further improve the coding efficiency, SLS also introduces a more sophisticated probability assignment named CBAC to complement the BPGC coding method. The idea of CBAC is inspired by the fact that the probability distribution of bit-plane symbols is usually correlated with their frequency location and the significance state of the adjacent spectral lines. In order to capture these correlations, three types of the contexts are used in SLS. The detailed context assignments are summarized as follows:

- Context 1: Frequency Band (FB)
  It is found in [29] that the probability distribution of bit-plane symbols of IntMDCT varies for different frequency bands. Therefore, in CBAC the IntMDCT spectral data are classified into three different FB contexts, namely, Low Band (0 ∼ 4 kHz, FB = 0), Mid Band (4 kHz ∼ 11 kHz, FB = 1) and High Band (above 11 kHz, FB = 2).

- Context 2: Distance to Lazy (D2L)
  The D2L context is defined as the distance of the current bit-plane j to the BPGC Lazy Plane parameter L, as defined in the following equation

  $$\text{D2L} = \begin{cases} 3 - j + L, & j - L \geq -2 \\ 6, & \text{else} \end{cases} \tag{8}$$

  The rationale behind follows the BPGC frequency assignment rule (7), which is based on the fact that the skew of the probability distribution of the bit-plane symbols from a source with Laplacian or near-Laplacian distribution tends to decrease as the number of D2L decrease. To reduce the total number of the D2L context, all the bit-planes with D2L < -2 are grouped into one context where all the bit-plane symbols are coded with probability 0.5.

- Context 3: Significant state (SS)
  The SS context tries to group the factors that may correlate with the distribution of the amplitude of the IntMDCT residual in one place. These include the amplitude of the adjacent IntMDCT spectral lines and the quantization interval of the AAC core quantizer if it has previously quantized in the core encoder. The detailed configuration of the SS context can be found at [38].

5.1.3. Low Energy Mode Coding

The BPGC/CBAC coding process described above works well for source with Laplacian or near-Laplacian distribution, which is usually the case for most audio signals [40] . However, it is also found that for some music items, there always exist some Time/Frequency (T/F) regions with very low energy level where the IntMDCT spectral data are in fact dominated by the rounding errors of IntMDCT algorithm whose distribution is far from Laplacian. In order to efficiently encode those low energy regions, the BPGC/CBAC coding process is replaced with the low energy mode coding described as follows.

The low energy mode coding is evoked for sfb for which L is smaller or equal to 0. At low energy mode coding, the amplitude of the residual spectral data $e[k]$ is first converted into unitary binary string $\mathbf{b} = \{b[0], b[1], ..., b[pos], ...\}$ as illustrated in Table 1. It can be seen that the probability distributions of these symbols is a function of position pos, and the distribution of $e[k]$:

$$Pr\{b[pos] = 1\} = Pr\{e[k] > pos|e[k] \geq pos\} \qquad (9)$$

where $0 \leq pos < 2^M$. $b[pos]$ is then arithmetic coded conditioned on its position pos and L with a trained frequency table.

5.2. Smart Decoding

With using characteristics of arithmetic decoding, this tool provides an efficient way to decode an intermediate layer corresponding to a given target bitrate in scalable decoding. This tool decodes additional symbols in the absence of incoming bits when a decoding buffer still contains meaningful information for arithmetic decoding in the CBAC/BPGC mode and/or low energy mode. This decoding continues up to the point where there exists no ambiguity in determining a symbol [44]. This tool can be effective when transmitting truncated bits at lower bitrate.

Table 1: Binarization of IntMDCT error spectrum at low energy mode.

Amplitude of $e[k]$	Binary string $\{b[pos]\}$
0	0
1	1 0
2	1 1 0
...	...
$2^M - 2$	1 1 ... ... 1 0
$2^M - 1$	1 1 ... ... 1 1
pos	0 1 2 3 ...

6. BIT-STREAM MULTIPLEXING

The HD-AAC data, including the core layer AAC bit-stream and the LLE bit-stream, can be carried in multiple elementary streams (ES) in an MPEG-4 system [41]. As shown in Fig. 11, the AAC bit-stream is carried in the base-layer ES, and the enhancement bit-stream is carried in one or more enhancement layers ES(s). Each ES is thus constructed by a sequence of access units (AU) where one AU contains one audio frame from AAC bit-stream or the enhancement bit-stream. Each ES also associated with an ES_Descriptor that contains necessary information for decoding the corresponding ES in the decoder, and an ES_ID point to the stream of the actual coded data. In an HD-AAC decoder, the AAC and SLS raw data are reconstructed from the received ES(s) before the actual decoding process is performed.

Such a bit-stream structure provides great flexibility in constructing either a large-step scalable system or a FGS system with SLS.

7. OTHER TOOLS

As a counterpart of the underlying AAC perceptual audio coder, SLS provides a number of integer versions of AAC coding tools.

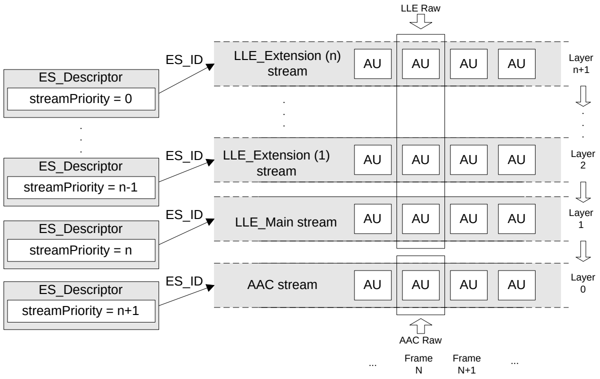

Fig. 11: Structure of MPEG-4 SLS bit-stream.

7.1. Integer M/S

In the AAC codec, the M/S tool allows to choose between mid/side and left/right coding for each scale factor band individually. On the other hand, the IntMDCT in the lossless enhancement only allows to globally choose between a left/right and a mid/side spectral representation. In order to make the integer spectral values fit to the spectral values from the AAC core, an invertible integer version of the M/S mapping is used. It is based on a lifting decomposition of the normalized M/S matrix, i.e. of a rotation by $\pi/4$.

$$\begin{aligned}&\frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix} = \\ &\begin{pmatrix}1 & 1-\sqrt{2}\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ \frac{1}{\sqrt{2}} & 1\end{pmatrix}\begin{pmatrix}1 & 1-\sqrt{2}\\ 0 & 1\end{pmatrix}\end{aligned} \tag{10}$$

7.2. Integer TNS

When the Temporal Noise Shaping (TNS) tool is used in the AAC core, the resulting MDCT spectral values deviate from the IntMDCT spectral values. In order to compensate for that, the same TNS filter as in the AAC core is applied to the integer spectral values in the lossless enhancement. To assure lossless operation, the TNS filter is converted to a deterministic invertible integer filter.

8. SOME FEATURES

This section introduces some of the features offered by SLS, such as oversampled mode (running the enhancement layer at a higher sampling rate than the AAC core coder), and the combination with AAC Scalable or AAC/BSAC.

8.1. Oversampling

MPEG-4 SLS uses an additional feature called "oversampling". It referres to the possibility to let the lossless enhancement operate at a higher sampling rate than that of the AAC core codec. The ratio between the SLS sampling rate and the AAC sampling rate is called "oversampling factor". It can be either 1, 2 or 4.

For example, the lossless enhancement can operate at a rate of 192 kHz, while the AAC core operates at 48 kHz, see table 2.

The mapping between the two coding layers is achieved by using a time-aligned framing and a correspondingly longer IntMDCT in the lossless enhancement. For example, a IntMDCT size of 4096 spectral values is used in case of oversampling by 4. The 1024 MDCT values from the AAC core are mapped to the lower 1024 IntMDCT values.

This approach simultaneously provides two advantages:

- The lossless performance is improved when using longer transforms. It has turned out that a transform length of 2048 or 4096 provides a better lossless performance for stationary signals than a transform length of 1024.
- The AAC core can operate at a sampling rate which is more appropriate for perceptual coding (e.g. 48 kHz), while the lossless enhancement can support high resolution input (e.g. 192 kHz).

8.2. Combination with MPEG-4 Scalable AAC

The MPEG-4 Scalable AAC (with one or more AAC mono or stereo layers) can be integrated with the SLS. This is equivalent as replacing the AAC core codec for SLS with Scalable AAC codec. In this integrated structure, the scalability can be achieved in both the core layer (scalable AAC) bitstream and lossless enhancement layer (SLS) bistream.

8.3. Combination with MPEG-4 AAC/BSAC

The SLS object is supported by the scalable to lossless tool which provides fine-grain scalable to lossless

	AAC @ 48kHz	AAC @ 96kHz	AAC @ 192kHz
SLS @ 48kHz	x		
SLS @ 96kHz	x	x	
SLS @ 192kHz	x	x	x

Table 2: Example combinations of sampling rates for AAC core and lossless enhancement

enhancement of MPEG perceptual audio codecs, allowing multiple enhancement steps in audio quality. In this mode of combination with core layer codec MPEG-4 AAC/BSAC, the SLS object allows multiple enhancement steps in core layer audio quality and let fine grain scalable decoding with 1kbps enhancements steps per channel with core layer codec. This combination enables a scalable decoding service to extend the lower range of scalable decoding bitrate of MPEG-4 SLS from perceptually lossless quality bitrate to very low bitrates.

8.4. SLS stand-alone mode

The SLS lossless enhancement can also operate as a stand-alone codec, without any underlying core codec.

9. PERFORMANCE

In this section, the performance of SLS is evaluated. For lossless operation the compression ratio is evaluted, while for near-lossless operation audio quality measurements are performed.

9.1. Lossless Performance

This section reports the compression ratio performance of SLS. We evaluate the compression ratio performance for both systems by using the MPEG-4 lossless audio coding testing sets donated by Matsushita Corporation. Meanwhile, we also evaluate compression ratio performance of SLS on 20 commercial music CDs (which comprises of many different type of music styles and the size of the total testing set is 11.128 GBytes). The results are listed in Table 3 and Table 5. Here the compression ratio is defined as

$$\text{Compression ratio} = \frac{\text{original file size}}{\text{compressed file size}}$$

For MPEG-4 Losselss Audio testing set a compression of, on average, factor 2:1 can be achieved easily at a sampling rate of 48 kHz and 16 bits word length. It can also be observed that for the AAC-based mode, where AAC is operating at 128kbps, an additional bitrate of only 30 to 40 kbps is required compared to the non-core mode for lossless representation. This reduces the bitrate consumption by 90 to 100 kbps compared to simulcast solutions that transmits an AAC bitstream and a non-core lossless bitstream simultaneously.

9.2. Near-Lossless Performance

The bit-plane coding of residual spectral values, i.e. of the AAC quantization error, allows to refine the initial AAC quantization successively. With each additional bit-plane the quantization error is reduced by 6 dB. Consequently, an increasing safety margin with respect to audibility is added as the bitrate is increased.

9.2.1. Evaluation of Near-Lossless Audio Quality

While it may seem sufficient for most purposes to provide perceptually transparent reproduction of audio signals by using conventional perceptual audio coders (e.g. AAC at sufficient bitrate), there are applications which demand still higher audio quality. This is especially the case for professional audio production facilities, such as archiving and broadcasting in which audio signals may undergo many cycles of encoding / decoding ("tandem coding") before being delivered to the consumer. This leads to an accumulation of introduced coding distortion and may lead to unacceptable final audio quality, unless substantial headroom towards audibility is provided by each coding step, e.g. by using coding algorithms with very high quality resp. bitrate.

The ITU-R recommendation BS.1548-1 [31] defines requirements for audio coding systems for digital broadcasting, assuming a codec chain consisting of so-called *contribution*, *distribtion*, and *emission* codecs. According to this recommendation, and based on ITU-R BS.1116 [32], audio codecs for contribution and distribution should fulfill the following requirements:

"The quality of sound reproduced after a reference

	SLS + AAC @ 128kbps/stereo (AAC @ 48kHz sampling rate)		SLS stand-alone	
	Compression ratio	Average bitrate	Compression ratio	Average bitrate
48kHz/16bit	2.09	735	2.20	698
48kHz/24bit	1.55	1490	1.58	1454
96kHz/24bit	2.09	2201	2.13	2160
192kHz/24bit	2.60	3543	2.63	3509
Overall	2.08	1992	2.12	1955

Table 3: Lossless compression results for MPEG-4 Lossless Audio Testing Set

contribution/distribution cascade [...] should be subjectively indistinguishable from the source for most types of audio programme material. Using the triple stimuli double blind with hidden reference test, described in Recommendation ITU-R BS.1116 [...], this requires mean scores generally higher than 4.5 in the impairment 5-grade scale, for listeners at the reference listening position. The worst rated item should not be graded lower than 4."

In accordance with these recommendations, tests were run on signals en/decoded with AAC/SLS. The PEAQ measurement [33] provides methods for objective measurements of perceived audio quality. It focuses on applications which are normally assessed in the subjective domain by applying an ITU-R BS.1116 test. The most essential results can be seen in Figures 12, 13 and 14.

The graphs show the *Objective Difference Grade* (ODG) values which have been computed by a PEAQ system. The evaluation procedure consisted of multiple cycles of tandem coding/decoding with up to 16 cycles. The standard set of critical MPEG-4 audio items for perceptual audio coding evaluations was used. An ODG value of 0, -1, -2, -3, -4 means 'indistinguishable', 'perceptible but not annoying', 'slightly annoying', 'annoying', 'very annoying', respectively.

Figure 12 shows the achieved ODG values as a function of tandem cycles for a traditional AAC coder running at a bitrate of 128 kbps/stereo. As expected, it can be observed that the audio quality significantly degrades with increasing number of tandem cycles, depending on the test item. Clearly, tandem coding is not a recommended practice for such coders.

Figure 13 shows the corresponding tandem coding results for the AAC+SLS combination running at 512 kbps/stereo (AAC @ 128 kbps + SLS enhancement @ 384 kbps). It can be observed that the audio quality remains at a very high level, even after a total of 16 tandem cycles. This illustrates the high robustness of such a representation against tandem coding. According to this measurement, the aforementioned audio quality requirement on BS.1548-1 is fulfilled.

Furthermore, when placed in tandem with AAC, the resulting audio quality is not significantly degraded by the SLS tandem cascade. More details can be found in [34].

9.2.2. Stand-alone SLS Operation

The SLS codec can also operate as a stand-alone lossless codec when the AAC core codec is not used, referred to as "non-core mode". Despite the simple structure in this mode (only IntMDCT and BPGC/CBAC modules are used), this mode allows efficient lossless coding, see [34]. Furthermore, fine-grain scalability by truncated bit-plane coding is also possible in this mode. Given that the stand-alone SLS codec does not include any perceptual model to estimate masking thresholds, it is interesting to investigate the audio quality resulting from a truncation of the SLS bitstream.

A closer look into the behavior of the bit-plane coding in this mode reveals that a constant SNR per scalefactor band is achieved. With each additional bit-plane the SNR is improved by 6 dB. While this

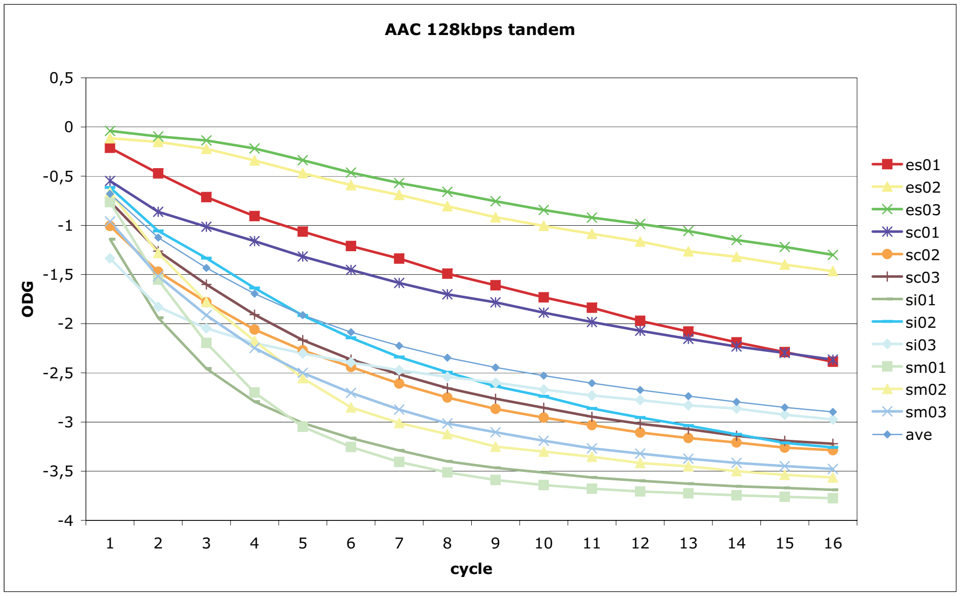

Fig. 12: Test results: AAC in tandem coding

behaviour does not allow to compete with efficient perceptual codecs at low bitrates (e.g. AAC at 128 kbps stereo), this simple approach works quite well at higher bitrates in the near-lossless range.
Figure 14 shows tandem coding results for the stand-alone SLS codec operating at 512 kbps/stereo. It reaches about the same near-lossless audio quality as in the AAC-based mode presented in the previous section.

Further increasing the bitrate towards 768 kbps, most test items still require some truncation in order to guarantee this constant bitrate. Nevertheless, the corresponding PEAQ measurements indicate that both for the AAC-based mode and for the stand-alone mode no degradation of audio quality occurs in this tandem coding scenario, see [34].
This provides an interesting operating point for SLS, corresponding to a guaranteed 2:1 compression. While other stand-alone lossless codecs can also provide an average compression of 2:1 for suitable test material, their peak compression performance can be much worse, depending on the audio material to be encoded. In contrast, SLS is able to guarantee a certain compression ratio while providing lossless or near-lossless signal representation, depending on the input signal.

10. DECODER COMPLEXITY

The computational complexity for SLS is evaluated by counting the total numbers of standard instructions (multiplications, additions, bit-shifts, comparisons, memory transfers, etc) required for performing the decoding process on a generic 32-bit fixed-point CPU.

The main components contributing to SLS computational complexity are:

1. IntMDCT filterbank
2. Bit-plane arithmetic decoder
3. AAC huffman decoding
4. AAC+SLS inverse error mapping
5. Integer M/S stereo coding
6. Unpacking of tables

Items 3 to 5 are only required in the AAC based mode, item 6 only if the necessary tables are not precomputed.

10.1. Number of instructions

Table 4 lists both the amount of instructions required for decoding in the AAC-based mode with

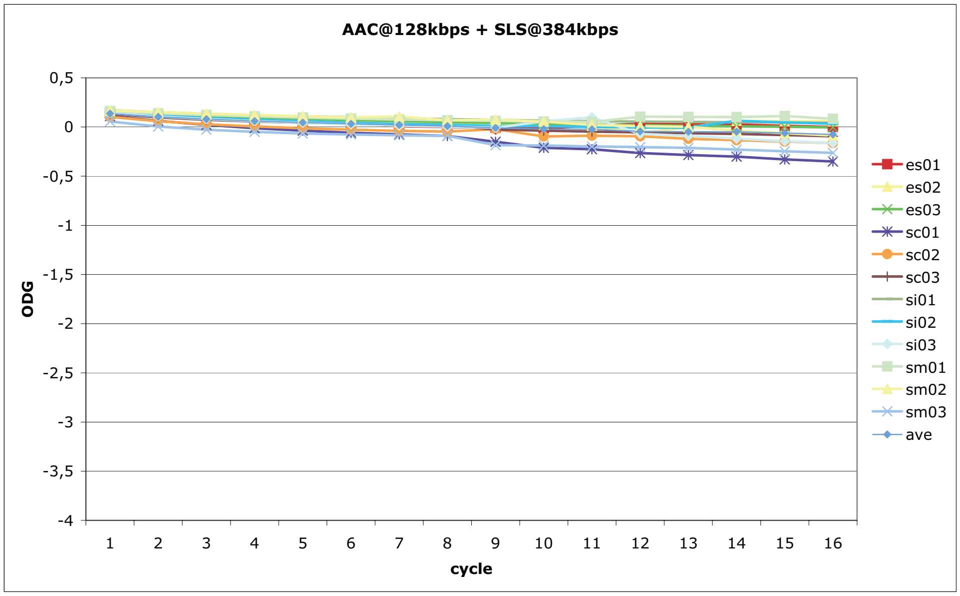

Fig. 13: Test results: AAC + SLS in tandem coding

the AAC core operating at 64kbps/channel and for the non-core mode (without AAC core layer).

10.2. ROM requirements

The memory requirement of the SLS implementation in terms of ROM usage is summarized as follows:

- For SLS non-core mode, the ROM requirement is 4K bytes.
- For AAC-based mode, the ROM requirement is 45K bytes.

As can be seen from Tables 4, a trade-off between ROM requirement and number of instructions can be made by pre-computing the necessary table values.

More details on the computational complexity can also be found in [34].

11. APPLICATIONS

As the primary functionality of SLS audio coding is lossless audio coding, it can be used in applications that require bit-exact reconstruction, such as studio operation, music disc delivery, audio archiving, etc. Due to its scalability feature, the SLS audio coding technology in fact fits into virtually every application that requires audio compression. Several potential application scenarios for SLS audio coding technology are listed below.

- Studio Operations
  The SLS audio coding technology is useful for storage of audio at various points in the studio operations such as recording, editing, mixing and premastering as studio procedures are designed to preserve the highest levels of quality. The scalability of SLS also provides a nice solution for situations where the bandwidth is not sufficient to support lossless quality.
- Archival
  Archives of sound recordings are very common in studios, record labels, libraries, etc. These archives are tremendously large and certainly compression are essential. In addition, the scalability of SLS technology enables the possibility that low bit-rate versions of the archives lossless audio items can be extracted at any time to allow applications such as remote data browsing.
- Broadcast Contribution/Distribution Chain

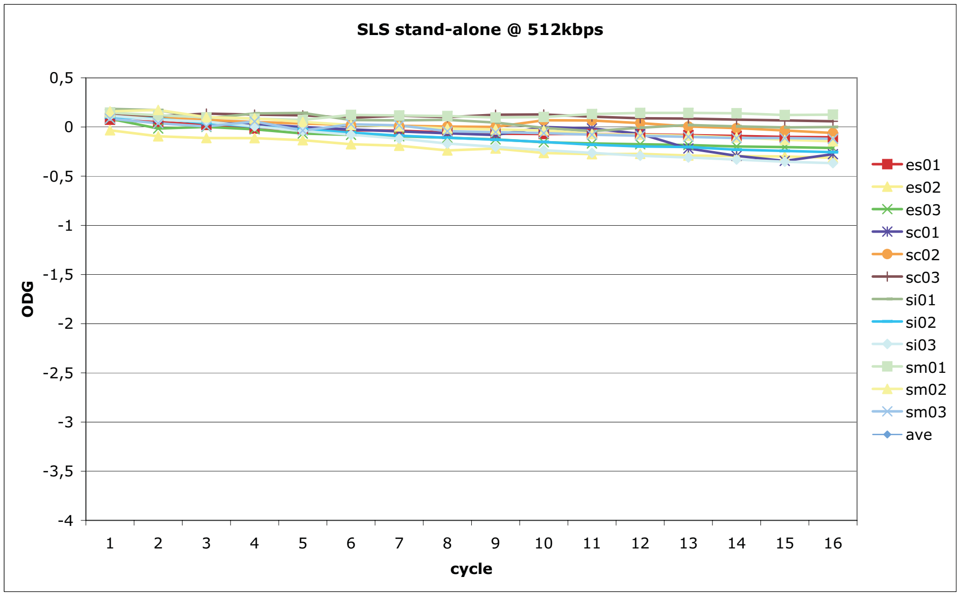

Fig. 14: Test results: SLS stand-alone in tandem coding

In a broadcast environment, SLS audio coding technology could be used in all stages comprising archiving, contribution/distribution and emission. In the broadcast chain, one main feature of the SLS technology can be used: In every stage where lower bit rates are required, the bit stream is just truncated, and no re-encoding is therefore required.

- Consumer Disc-Based Delivery
  The SLS technology can also be used in consumer disc-based delivery of music contents. It also enables the music disc to deliver both lossless and lossy audio on the same disc.

- Internet Delivery of Files
  In such an application model the bandwidth condition can vary dramatically over different access network technologies. As a result, same audio contents at a variety of bit-rates and qualities may need to be provisioned at the server side. SLS technology provides one-file solution for such a requirement.

- Streaming
  The SLS audio coding technology delivers the vital bit-rate scalability for streaming applications on channel with variable QoS conditions. Examples for this kind of streaming applications include the Internet audio streaming, multicast streaming applications that feeds several channels of differing capacity, etc.

- Digital Home
  The idea of digital home is to create an open and transparent home network platform where the consumers can easily create, use, manage and share digital content such as audio, video, image, and others. In a typical setup for audio, the user can download the SLS coded bit-streams in lossless quality from the service provider and archive them in the home music server; these bit-stream are then streamed, or downloaded to different audio terminal at various quality for playback.

12. CONCLUSIONS

The new ISO/MPEG specification for Scalable Lossless Coding extends the well-known Advanced Audio Coding (AAC) perceptual coding scheme towards lossless and near-lossless operation, and in this way enables its use in the context of high definition applications. The scheme offers competitive lossless com-

Tables pre-unpacked		
Frame length	With AAC core	Non-core mode
	1 cycle	1 cycle
4096 or 512	270.86	245.70
2048 or 256	265.05	239.89
1024 or 128	250.83	225.67
Tables unpacked in place		
Frame length	With AAC core	Non-core mode
	1 cycle	1 cycle
4096 or 512	316.10	290.05
2048 or 256	303.30	278.14
1024 or 128	273.58	248.42

Table 4: Maximum numbers of INT32 operations per sample in the HD-AAC decoder

pression rates at all typical operating points (word lengths / sampling rates). For distribution on bandwidth limited channels, a perceptually coded compatible AAC bitstream can simply be extracted from the composite AAC/SLS stream. Alternatively, SLS can also be run as a simple and versatile stand-alone compression engine. In both cases, the fidelity of the signal representation can be scaled with fine granularity within a wide range of near-lossless representations. This enables lossless / near-lossless transmission of high definition audio with a guaranteed maximum rate. We anticipate that this flexibility will make HD-AAC the technology of choice for many applications.

13. ACKNOWLEDGEMENTS

The authors would like to thank all the colleagues at the MPEG audio subgroup who supported the lossless standardization activity, especially Takehiro Moriya (NTT) for inspiring these standardization activities, Tilman Liebchen (Technical University of Berlin) for chairing the ad-hoc group on lossless audio coding and Yuriy Reznik (Real Networks / Qualcomm) for the thorough complexity evaluations.

14. REFERENCES

[1] ISO/IEC 11172-3, "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s, Part 3: Audio", 1992.

[2] ISO/IEC 13818-3, "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 3: Audio", 1994.

[3] ISO/IEC JTC1/SC29/WG11, "Final Call for Proposals on MPEG-4 Lossless Audio Coding", MPEG2002/N5208, Shanghai, China, October 2002.

[4] ISO/IEC 14496-3:2001/Amd.1:2003, “Coding of Audio-Visual Objects - Part 3: Audio, Amendment 1: Bandwidth extension”, 2003.

[5] M. Dietz, L. Liljeryd, K. Kjoerling, O. Kunz, “Spectral Band Replication, a Novel Approach in Audio Coding”, 112th Convention of the AES, Munich, Germany, April 2002.

[6] ISO/IEC 14496-3:2001/Amd.2:2004, “Coding of Audio-Visual Objects - Part 3: Audio, Amendment 2: Parametric coding for high quality audio”, 2004.

[7] A.C. den Brinker, E. Schuijers, A.W.J. Oomen, “Parametric Coding for High-Quality Audio”, 112th Convention of the AES, Munich, Germany, April 2002.

[8] ISO/IEC FCD 23003-1, “MPEG-D (MPEG audio technologies), Part 1: MPEG Surround”, 2006.

[9] J. Breebaart, S. Disch, C. Faller, J. Herre, G. Hotho, K. Kjoerling, F. Myburg, M. Neusinger, W. Oomen, H. Purnhagen, J. Roeden, “MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status”, 119th Convention of the AES, New York, NY, USA, October 2005.

[10] “Journal of the Audio Engineering Society - Special Issue: High-Resolution Audio”, Volume 52, Number 3, March 2004.

[11] E. Knapen, D. Reefman, E. Janssen, F. Bruekers, “Lossless Compression of 1-Bit Audio”, J. Audio Eng. Soc., Vol. 52, No. 3, pp. 190-199, March 2004.

[12] ISO/IEC 14496-3:2001/Amd.6:2005, "Coding of Audio-Visual Objects - Part 3: Audio, Amendment 6: Lossless coding of oversampled audio", 2005.

[13] ISO/IEC 14496-3:200X/Amd.2, "Coding of Audio-Visual Objects - Part 3: Audio, Amendment 2: Audio Lossless Coding (ALS), new audio profiles and BSAC extensions", to be published.

[14] ISO/IEC 14496-3:200X/Amd.3, "Coding of Audio-Visual Objects - Part 3: Audio, Amendment 3: Scalable Lossless Coding (SLS)", to be published.

[15] R. Geiger, M. Schmidt, J. Herre, R. Yu, "MPEG-4 SLS - Lossless and Near-Lossless Audio Coding Based on MPEG-4 AAC," International Symposium on Communications, Control and Signal Processing, Marrakech, Morocco, 13-15 March 2006.

[16] ISO/IEC 14496-3:2001, "Coding of Audio-Visual Objects, Part 3 Audio", 2001

[17] J. Herre, H. Purnhagen, "General Audio Coding", in F. Pereira, T. Ebrahimi (Eds.), *The MPEG-4 Book*, Prentice Hall IMSC Multimedia Series, ISBN 0-13-061621-4, 2002

[18] ISO/IEC JTC1/SC29/WG11, "Report on the MPEG-2 AAC Stereo Verification Tests", MPEG1998/N2006, San Jose, USA, Feb. 1998

[19] J. Princen, A. Johnson, A. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", *ICASSP*, 1987, pp. 2161 - 2164

[20] J. Herre, J.D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)", *101st AES Convention*, Preprint 4384, Los Angeles, USA, Nov. 1996

[21] J.D. Johnston, J. Herre, M. Davis, U. Gbur, "MPEG-2 NBC Audio - Stereo and Multichannel Coding Methods", *101st AES Convention*, Preprint 4383, Los Angeles, USA, Nov. 1996

[22] R. Geiger, T. Sporer, J. Koller, K. Brandenburg, "Audio Coding based on Integer Transforms", *111th AES Convention*, New York, USA, Sep. 2001

[23] R. Geiger, Y. Yokotani, G. Schuller, "Improved integer transforms for lossless audio coding", *Proc. of the Asilomar Conf. on Signals, Systems and Computers*, 2003

[24] R. Geiger, Y. Yokotani, G. Schuller, J. Herre, "Improved Integer Transforms using Multi-Dimensional Lifting", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 17-21, 2004, Montreal, Quebec, Canada.

[25] Y. Yokotani, R. Geiger, G. Schuller, S. Oraintara, K. R. Rao, "Improved Lossless Audio Coding using the Noise-Shaped IntMDCT", IEEE 11th DSP Workshop, August 1-4, 2004, Taos Ski Valley, New Mexico, USA.

[26] R. Geiger, J. Herre, J. Koller, K. Brandenburg, "IntMDCT - A link between perceptual and lossless audio coding", *ICASSP*, Orlando, USA, May 2002

[27] R. Yu, R. Geiger, S. Rahardja, J. Herre, L. Xiao, H. Haibin, "MPEG-4 Scalable to Lossless Audio Coding", *117th AES Convention*, San Francisco, USA, Oct. 2004

[28] R. Yu, C.C. Ko, S. Rahardja, X. Lin, "Bit-plane Golomb code for sources with Laplacian distributions", *Proc. ICASSP*, 2003, pp. 277-280

[29] R. Yu, X. Lin, S. Rahardja, C. C. Ko, H. Huang, "Improving Coding Efficiency for MPEG-4 Audio Scalable Lossless Coding", *IEEE 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, May 2005

[30] R. Yu, X. Lin, S. Rahardja, C.C. Ko, "Advanced Audio Zip - A fine granular scalable to lossless audio coder", to appear in *IEEE Trans. Speech Audio Processing*

[31] Recommendation ITU-R BS.1548-1, "User requirements for audio coding systems for digital broadcasting", International Telecommunications Union, Geneva, Switzerland, 2001-2002

[32] Recommendation ITU-R BS.1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems", International Telecommunications Union, Geneva, Switzerland, 1994

[33] Recommendation ITU-R BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality", International Telecommunications Union, Geneva, Switzerland, 1998

[34] ISO/IEC JTC1/SC29/WG11, "Verification Report on MPEG-4 SLS", MPEG2005/N7687, Nice, France, Oct. 2005

[35] M. Militzer, M. Suchomski, K. Meyer-Wegener, "LLV1: layered lossless video format supporting multimedia servers during realtime delivery", *Proc. SPIE*, Vol. 6015, Oct. 2005, pp. 436-445

[36] M. Suchomski, M. Militzer, K. Meyer-Wegener, "RETAVIC: using meta-data for real-time video encoding in multimedia servers", *Proc. ACM SIGMM NOSSDAV*, Stevenson, Washington, USA, Jun. 2005, pp. 81-86

[37] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding", J. Audio Eng. Soc., Vol. 45, No. 10, pp. 789-813, Oct. 1997.

[38] R. Yu, X. Lin, S. Rahardja, and H. Haibin, "Proposed Core Experiment for improving coding efficiency in MPEG-4 audio scalable coding (SLS)", ISO/IEC JTC1/SC29/WG11, M10683, March. 2004, Munich, Germany.

[39] IA-32 Intel Architecture Optimization Reference Manual, ON: 248966-009, Intel Corp.

[40] R. Yu, X. Lin, S. Rahardja and C.C. Ko, "A Statistics Study of the MDCT Coefficient Distribution for Audio," Proc. ICME 2004

[41] ISO/IEC JTC1/SC29/WG11, "Coding of Audiovisual Objects, Part 1 System," International Standard 14496-1

[42] I. Daubechies, W. Sweldens, "Factoring Wavelet Transforms into Lifting Steps", Tech. Rep., Bell Laboratories, Lucent Technologies, 1996.

[43] F. Bruekers, A. Enden, "New networks for perfect inversion and perfect reconstruction", IEEE JSAC, vol. 10, no. 1, pp. 130 137, Jan. 1992.

[44] K.-H. Choo, J.-H. Kim, E. Oh, C.-Y. Son, "Enhanced Performance in the Functionality of Fine Grain Scalability", 119th Convention of the AES, New York, NY, USA, October 2005.

1. ANNEX

CD items	SLS+AAC @ 128kbps/stereo	SLS non-core
	Compression ratio	Compression ratio
ACDC - Highway to Hell (Sony 80206)	1.31	1.36
Avril Lavigne - Let Go (Arista 14740)	1.36	1.41
Backstreet Boys - Greatest Hits Chapter One (Jive 41779)	1.39	1.45
Brian Setzer - The Dirty Boogie (Interscope 90183)	1.43	1.49
Cowboy Junkies - Trinity Session (RCA-8568)	1.93	2.04
Grieg - Peer Gynt - von Karajan (DG 439010)	2.63	2.83
Jannifer Warnes - Famous Blue Raincoat (BMG 258418)	2.07	2.20
Marlboro Music Festival, DISC A (Bridge 9108)	2.23	2.35
Marlboro Music Festival, DISC B (Bridge 9108)	2.20	2.33
Nirvana - Nirvana (Interscope 493523)	1.50	1.56
Philip Jones - 40 Famous Marches, CD1 (Decca 416241)	1.99	2.11
Philip Jones - 40 Famous Marches, CD2 (Decca 416241)	2.00	2.11
Pink Floyd - Dark Side of the Moon (Capitol 46001)	1.76	1.85
Rebecca Pidgeon - The Raven (Chesky 115)	1.88	1.97
Ricky Martin (Sony 69891)	1.33	1.38
Schubert Piano Trio in E-flat (Sony 48088)	2.74	2.90
Spaniels - The Very Best 0f (Collectables 7243)	2.41	2.62
Steeleye Span - Below the Salt (Shanachie 79039)	1.85	1.95
Suzanne Vega - Solitude Standing (A&M 5136)	1.74	1.83
Westminster Concert Bell Choir - Christmas Bells (Gothic Records 49055)	2.55	2.71
Overall	1.85	1.94

Table 5: Lossless compression results for commercial CD Testing Set