balesio AG

balesio Native Format Optimization Technology (NFO)

White Paper

Abstract

balesio provides the industry’s most advanced technology for unstructured data optimization, providing a fully system-independent optimization workflow to unstructured data. It is neither deduplication nor compression, it is *Native Format Optimization* - our optimization technology composed of a comprehensive set of content-aware native optimization algorithms which are especially developed for unstructured file formats such as Microsoft Office files, PowerPoint presentations, images and PDF files.

Native Format Optimization by balesio is delivering end-to-end savings for customers with large-scale unstructured file collections typically found on primary storage. NFO delivers bandwidth savings, data traffic benefits and backup time reduction in addition to storage savings.

Chris Schmid, balesio AG

Table of Contents

Introduction

balesio delivers true Native Format Optimization (NFO) for unstructured files via the FILEminimizer engine, the world's only true content-aware data optimization solution for unstructured files on primary storage. balesio has gained worldwide recognition for its post-process architecture and ability to reduce pre-compressed file format types in large-scale, and to deliver substantial customer savings where traditional deduplication and compression technologies deliver no benefit.

Solution Background and Approach

balesio's Native Format Optimization technology can be divided into three major components which are critical for a scale-out unstructured file optimization solution.

1. A comprehensive series of intelligent, content-aware algorithms that provide intra-file data optimization without any need for rehydration or decompression
2. A software management framework that supports flexible policy-based and fast multi-threaded optimizations
3. Optionally, optimizations can be performed via a reliable hardware platform (appliance) that is sized to deliver a maximum of parallel optimizations for the customer's data optimization workflow.

Native Format Optimization (NFO) is not just a matter of using naive LZ or LZW compression or applying ignorant file-based or block-based deduplication. The NFO optimization workflow includes the following categories of algorithms:

- Intelligent Content Analysis, Separation and Segmentation
    - File Delayering: NFO first opens files to identify, analyze and segment underlying file objects for separate algorithm treatment. This process is done successfully even for proprietary or compressed file types such as Microsoft Office, PDF or JPEGs.
    - File Object Delayering: NFO then identifies, analyzes and segments single objects within file objects for separate algorithm treatment. This process is done successfully even for proprietary or compressed file types such as Microsoft Office
- Multi-Stage Native Format Optimization
    - Deduplication: Eliminating of duplicate data at the sub-object level (intra-file)
    - Compression: Applying content-aware specific compression algorithms to different object types
    - Object Encoding: Re-evaluate the native coding efficiency for optimization opportunities to maximize utilization of physical blocks.

These optimization steps have been designed individually for each supported unstructured file format. Together, the supported unstructured file formats represent between 70-80% of all data stored on primary storage[1], proven effective and tested even for file types that were considered by the industry to be no further “compressible”.

Independent Workflow

In contrast to standard compression or deduplication workflows, the NFO workflow does not require any rehydration of an optimized file when reading data back. In fact, the NFO workflow is entirely different and INDEPENDENT from any further applied workflow or optimization: files are optimized but stored in their native file format (for example the image file "example.jpg" is natively optimized and stored as "example.jpg") - a read-back of an optimized file requires no rehydration.

Benefits of Native Format Optimization (NFO)

By intelligently reducing single file sizes and leaving them in their optimized state through the whole information lifecycle, customers using our NFO technology are realizing benefits in bandwidth cost and improved end-user experience, in addition to reducing storage costs and backup data volume.

While with standard compression or deduplication workflow the read-back of files requires rehydration, the NFO workflow is entirely different: files are optimized but stored in their native file format (for example the image file "example.jpg" is natively optimized and stored as "example.jpg") - a read-back of an optimized file requires no rehydration.

Therefore, the data can move completely through the workflow in its optimized form. For example, when unstructured files are compressed by 40%, then storage consumption and distribution bandwidth are also reduced by 40%, power and A/C are potentially reduced by 40%, and finally, the movement of data between datacenters or applications is accelerated by 40%, potentially reducing backup times and improving enterprise reliability. Simply put, NFO optimization provides benefits to the entire customer storage, datacenter and network infrastructure.

The Native Format Optimization Technology

balesio‘s NFO technology applies a comprehensive set of content-aware optimization mechanisms which are both “lossless” and “visually lossless” meaning they are technically “lossy”, but in practice deliver optimized files which are visually identical to the original. Carefully developed enhanced image coding algorithms align precisely with the sensitivities of the human visual system to deliver visually lossless results to the user. In addition, a set of content-aware optimization algorithms which have been especially developed for main unstructured file formats such as PowerPoint, Word, Excel and PDF provide even more powerful optimization results. Techniques employed to reduce single

[1] According to Research by Hewlett-Packard
http://news.techworld.com/applications/106661/it-grossly-underplaying-unstructured-data-growth/

file sizes via native format optimization include intra-file object deduplication, object format optimization, object slimming, noise reduction, quantization adjustment, optimization of non-visual data, enhanced image encoding and other methods.

NFO for Microsoft Office Formats

The main Microsoft Office formats, namely PowerPoint, Word and Excel, are de facto standards today for every computer user to create spreadsheets, documents, letters and presentations. The capabilities of these programs are immense offering the use of graphics, objects, pictures and images during the creation of a file.

Our NFO technology for Microsoft Office file formats works on three major fields to realize the enormous space savings within a file:

1. Correct end-user storage inefficiencies when creating a file
2. Optimize format structures and correct storage inefficiencies which are deliberately allowed by the file format itself
3. Optimize all image content within the Microsoft Office format container (see chapter “NFO for Images” for details)

From a technical perspective, one or all of the following Native Format Optimization processes are applied on the object-level of a Microsoft Office file:

- **Deduplication:** Elimination of duplicated elements within the single file
- **Formatting:** Calculating of and conversion to the most appropriate object format
- **Interpolation:** Enhanced image optimization (see Chapter “NFO for Images” for details)
- **Slimming and encoding:** Optimization and compaction of non-visual data, e.g. Huffman coding tables
- **Flattening:** Optimization and conversion of embedded objects
- **Compression:** Content-aware specific compression for single objects

Applying this native optimization process can produce results in storage space savings of up to 98%, depending on the input file. Average optimization ratios vary between 50-90%.

NFO for PDF Formats

The PDF format is a standard format for exchanging information from different sources. PDF files can originate from scanned paper or from different content, e.g. a PDF file is “created” from a PowerPoint presentation (PDF Conversion). As such, visual information is frequently stored in a PDF file. The format itself is very powerful offering the possibility to display graphics, objects, pictures and images and offering great potential for NFO as well.

Our NFO technology for PDF files works on three major fields to realize the enormous space savings within a file:

1. Correct storage inefficiencies from end-users and inefficient PDF converters
2. Optimize format structures and correct storage inefficiencies which are deliberately allowed by the file format itself
3. Optimize all image content within the PDF format container (see chapter “NFO for Images” for details)

From a technical perspective, one or all of the following Native Format Optimization processes are applied on the object-level of a PDF file:

- **Deduplication:** Elimination of duplicated elements within the single PDF file
- **Structure:** Optimization of internal PDF file structure
- **Formatting:** Calculating of and conversion to the most appropriate object format
- **Interpolation:** Enhanced image optimization (see Chapter “NFO for Images” for details)
- **Image/Color quantization:** Calculation and optimization of used color scheme tables
- **Slimming and encoding:** Optimization and compaction of non-visual data, e.g. Huffman coding tables
- **Non-referenced objects:** Clearing non-referenced “dead” objects
- **Compression:** Content-aware specific compression for single objects

Applying this native optimization process can produce results in storage space savings of up to 98%, depending on the input file. Average optimization ratios vary between 50-90%.

NFO for Images

NFO for JPEG Images

The JPEG image format is a standard for digital photography and web-based imaging. It provides a variety of opportunities for compression, depending on the underlying subject matter (the “motive”), and the encoding parameters used to capture the initial JPEG image.

Our NFO technology for JPEG files works on three major fields to realize the enormous space savings within a file:

- **Elimination of non-visual information**

JPEG images can contain thumbnails, as well as a variety of metadata including EXIF headers and comments. Other information related to the image, such as Huffman coding tables, can be reduced through table reformatting. Our NFO technology is highly flexible and able to strip non-visual information such as EXIF headers, other comment information and thumbnails from images. Furthermore, space can be reduced by optimizing Huffman coding and color space tables.

- **Optimization of visual information**

In JPEG images, many standard JPEG encoders - especially those in cameras with low CPU resources - treat luminance and chrominance with the same quantization weight. Our NFO process is revisiting quantization levels to capitalize on this initial inefficiency. By applying lowering quantization levels to chrominance information than luminance information, the weight of JPEG images can be reduced significantly while maintaining a visually identical experience to the human eye. In addition to re-weighting the luma (black-and-white) and chroma (color) channel levels, balesio's NFO process is non-linearly attenuating the higher frequency components of the DCT function, which aligns the pixel block encoding with actual visual sensitivity.

Applying this native optimization method can produce results in file size reductions of up to 20-80%, depending on the input image and the chosen optimization policies. It is worth noting that balesio provides a "gradient" compression; high quality source images can be better compressed than images with a low base quality. This is due to the nature of the NFO process; low quality images simply do not contain as much "material" which leaves less room for image optimization.

NFO for BMP Images

balesio's NFO technology also supports optimization for bitmap images (BMP). As this image format does not use the variable quantization of DCT-based transforms nor allows for LZ compression, the achievement of further data optimization for BMP files requires different techniques encompassed by the balesio NFO technology.

From a technical perspective, one or all of the following Native Format Optimization processes are applied to BMP image files:

- **Color adjustment**

Often, 24-bit or 32-bit RGB pixel colors are specified by default. Bitmap images, which in reality do not use that high spectrum of colors, e.g. 2-color or 4-color illustrations or graphs, are adjusted by our NFO technology. balesio therefore leverages this color mismatch to reduce file size of BMP images.

- **Format normalization**

balesio's NFO technology offers the possibility of format normalization for bitmap images. Modern standard image file formats such as PNG are more efficient than BMP for high-color images. balesio's NFO technology can in such a case normalize the BMP image in an optimized PNG file resulting in significant file size reductions.

NFO for GIF and PNG Images

balesio's NFO technology also supports optimization of GIF and PNG image file formats, which are lossless, compressed images using LZW and LZ77 compression algorithms respectively. As these image formats are (contrary to JPEG) not lossy by definition, they do not use the variable quantization of DCT-based transforms. Therefore, the achievement of further data optimization requires different techniques encompassed by the balesio NFO technology.

From a technical perspective, one or all of the following Native Format Optimization processes are applied on PNG and GIF image files:

- **Color adjustment and color map optimizations**

In many GIF files, 24-bit RGB pixel colors are specified by default. However, in reality that many colors are unlikely in a GIF image and therefore not found in the embedded color table. balesio leverages this mismatch to further reduce file size.

- **Context-based per-line filter selection**

The PNG specification accommodates 5 different filters (prediction methods) that can be used for encoding the data. Although not always employed, the PNG ISO standard can support multiple filter methods per image. balesio takes advantage of this by choosing the most efficient filter on a line-by-line basis. Often, the most efficient filter is not applied in the PNG file as it requires more CPU to calculate it (CPU speed is traded over space efficiency). balesio leverages this mismatch in the compression process to find the optimum size resulting in significant size reductions.

- **LZW and LZ77 compression inefficiencies**

All LZW and LZ77 encoders exhibit different levels of efficiency. In general, the efficient LZ-based compressors used by balesio can provide an up to 10% better compression than those typically used in GIF and PNG image encoders.

NFO for TIF Images

balesio’s NFO technology also supports optimization of TIF image file formats. As this image format does not use the variable quantization of DCT-based transforms nor allows for LZ compression, the achievement of further data optimization for TIF files requires different techniques encompassed by the balesio NFO technology.

From a technical perspective, one or all of the following Native Format Optimization processes are applied on TIF image files:

- **Color adjustment**

Often, 24-bit or 32-bit RGB pixel colors are specified by default. However, many TIF files originate from scans or represent plans or graphic illustrations which in reality do not use that high spectrum of colors. balesio leverages this mismatch to reduce file size of TIF images.

- **Format normalization (single-page TIF files)** balesio's NFO technology offers the possibility of format normalization for single-page, complex (high-color) TIF images. The TIF file format origins from the days of MS-DOS, which makes it an inefficient format for displaying images containing many colors. balesio's NFO technology can in such a case normalize the TIF image in an optimized PNG file resulting in significant file size reductions.

NFO Quality Measurement and Optimum Configuration

balesio delivers far more than the most advanced Native Format Optimization technology using a comprehensive set of industry-leading, content-aware and file-specific compression algorithms. balesio provides a full management framework that assures reliability, data quality, and provides an enterprise class management framework using manifold policy-based options for adapting balesio's NFO technology to individual customer needs.

NFO processes may be governed by optimization policies based on file type, file subtype, file location, file name, last access date, last modified date as well as file characteristics such as file size and file patterns.

In addition, file optimization may be invoked by capturing hash keys of single files (file ID).

Files meeting the specified criteria will be read into either the software-based NFO solution or into the data reduction appliance for the native optimization of the same.

The compression itself is determined by other policies establishing the desired level of compression:

- **Lossless pixels** - A setting that maintains the identical pixel visual experience of unstructured files. Size reduction is achieved by reducing non-visual image data including comments and thumbnails.
- **Visually lossless** - The most commonly used model by balesio customers. Enhances the coding using techniques listed above, to reduce file size while preserving quality levels and visual experience, even for large-scale printing.
- **Visually lossless display** - Provides native format optimization preserving all quality when files are displayed on a screen. Minimal compression artifacts are visible when printed as large-scale. Ideal for document and file archiving.
- **Visually lossy (strong)** - Provides maximum space savings while retaining acceptable image quality, ideal for smaller devices such as smartphones.
- **Custom** - Provides close to 100 settings to customize the Native Format Optimization for the customer's specific requirements.

NFO vs. Compression

Native Format optimization technology is designed for unstructured files with a mixture of content elements (e.g. images, graphics, elements, objects, text, etc.) and is therefore far more effective on unstructured files than traditional compression methods, such as ZIP compression is. Natively optimizing a large PowerPoint presentation will often reduce its file size by 50-90% while zipping the same presentation will result in a low compression rate of below 10% as included elements are already compressed. And, as natively optimized files remain in their original format, there's no need to unzip an optimized file before using it.

NFO vs. Primary Deduplication

Native Format optimization technology is designed for unstructured files which usually represent between 70-80% of the total data volume on primary storage. Native format optimization offers content-aware optimization methods which work on an intra-file level, i.e. within a given file container. On unstructured files, this technology provides higher data reduction ratios than primary deduplication for the fact that redundancies across files are less frequent on primary storage. Therefore, the effect of primary deduplication on unstructured files is limited. Native format optimization on the other hand finds the optimization potential within each single file and reduces primary storage data volume of unstructured files by 50-85%.

Summary

balesio offers with its Native Format Optimization (NFO) technology the most advanced technology platform for data reduction on the market. The ability to manipulate and optimize image data at a pixel, DCT, object, sub-object and file format level presents many opportunities to deliver economically significant, permanent benefits without degrading perceived image quality (visually lossless optimization).

The whole NFO process is an intra-file optimization requiring no rehydration or decompression (independent NFO process). The optimization starts at the beginning of the file lifecycle on primary storage saving customers not just on storage capital expenditures and operational expenses for primary storage, but as well on backup and archiving space and on bandwidth and network traffic.

- **Reduction in primary storage consumption by up 50-90%**
- **Reduction in backup storage consumption by 50-90%**
- **Reduction in archive storage by 50-90%**
- **Reduction in distribution bandwidth by 50-90%**
- **Accelerated backup times by 50-90%**

The ability to normalize incoming images to a specified level of quality as appropriate for a specific application, workflow and process increases further the efficiency of the above.

balesio delivers the NFO technology both as software and as part of a fully integrated platform (appliance), tuned for performance and reliability, and supporting policy-based operation to align data optimization with the customer workflow.