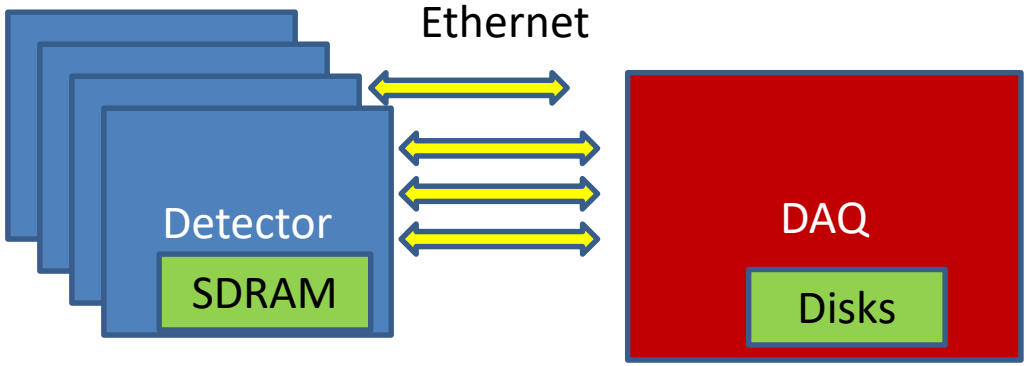
# Data Compression Techniques

Grzegorz Pastuszak
Warsaw University of Technology

Trieste 22.05.2019

# Need for compression

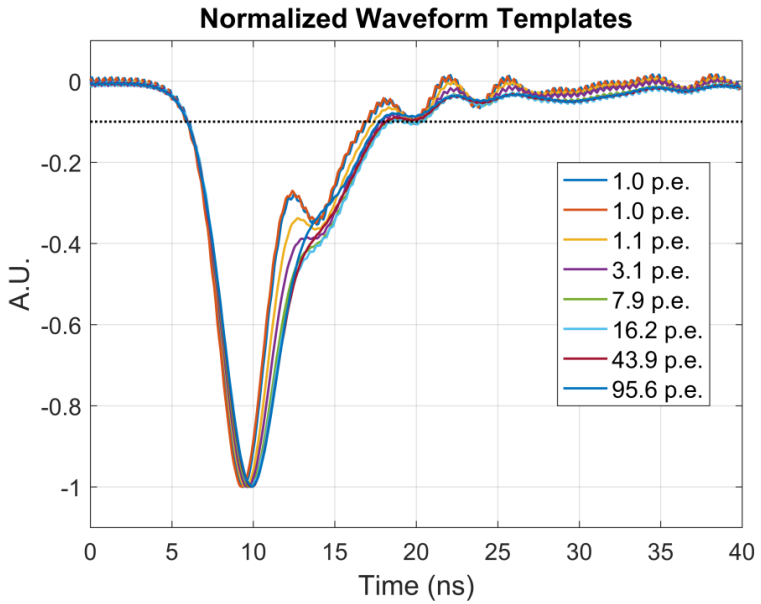
- Saving disk space for the archiving
- Limited bandwidth between detectors and the data acquisition system (DAQ)
- Saving RAM capacity in detector modules in case of pile-ups



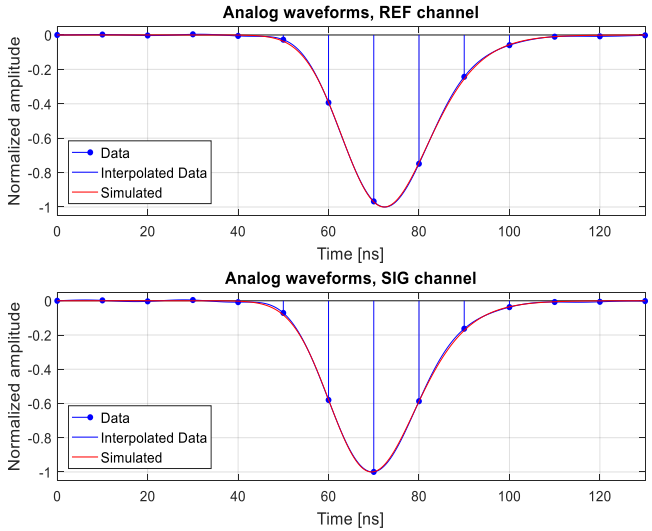

- Constraints on resources and power

# Input Waveforms

- Acquired PMT waveforms:
  - seems to be similar,
  - Stability is limited,
  - Shaping changes original signal from PMT.
- Allowable losses in processing should be small to preserve key waveform features
- How strong is the correlation of waveforms from neighboring PMTs?



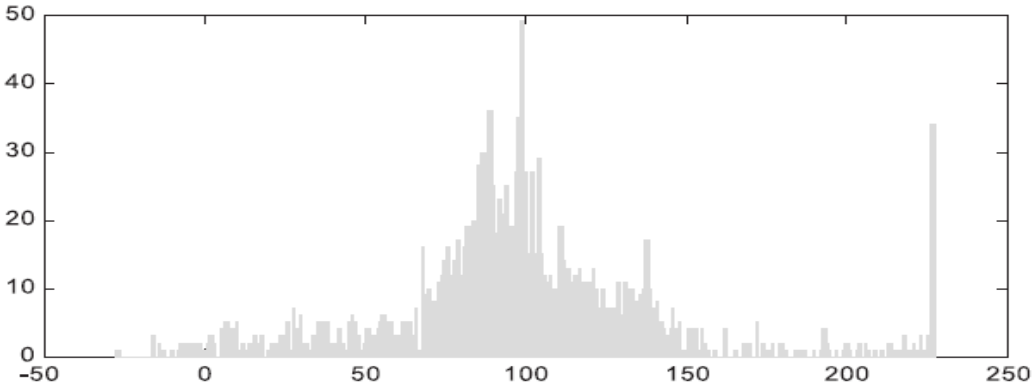


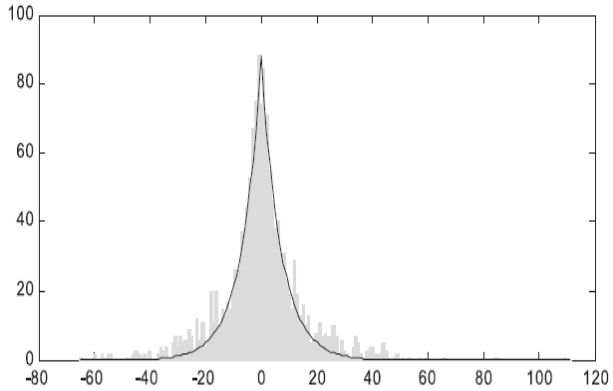


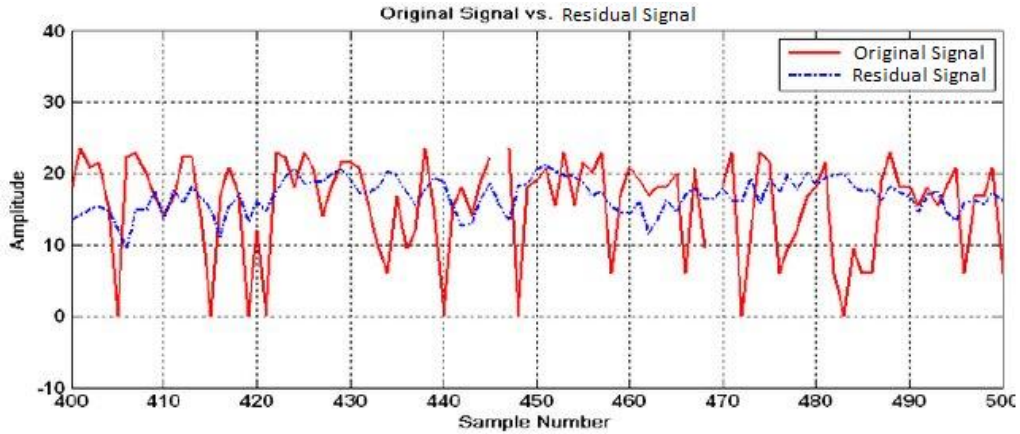
# Compression Methods

- Modeling
  - Linear Prediction
  - Signal Models
  - Transforms
- Quantization
  - Scalar quantization
  - Vector quantization – using signal models
- Entropy Coding
  - Variable length coding
  - Arithmetic coding – more complex and better compression

# Signal Modelling




- Predictions, Transformations decrease the dynamics
- Distributions of residual signal concentrated around zero
- Signal reconstruction using reverse operations

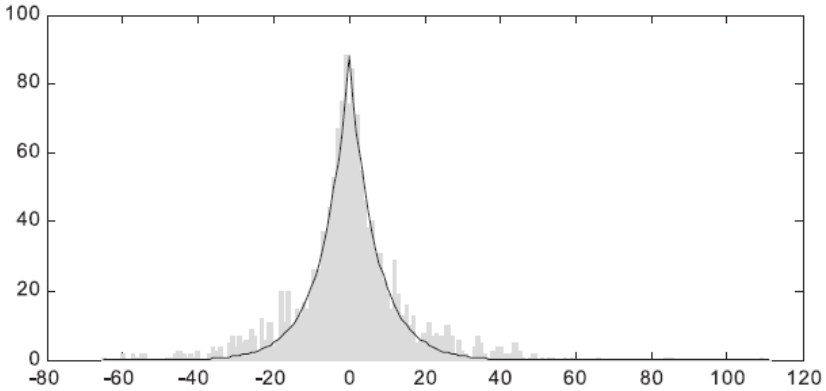
# Linear Prediction

- Prediction as a sum of previous samples multiplied by coefficients/weights

$$x_{predicted}[t] = \sum_{i=1}^{N} a_i x[t-i]$$

- Residuals (equal to difference between input samples and their predictions) have much lower values and energy

$$\Delta x[t] = x[t] - \sum_{i=1}^{N} a_i x[t-i]$$




- Coefficients must be known at the decoder -> precomputed or sent with residuals
- Error energy:

$$E = \sum_{t=0}^{T} (\varepsilon[t])^2 = \sum_{t=0}^{T} \left( x[t] - \sum_{i=1}^{N} a_i x[t-i] \right)^2$$

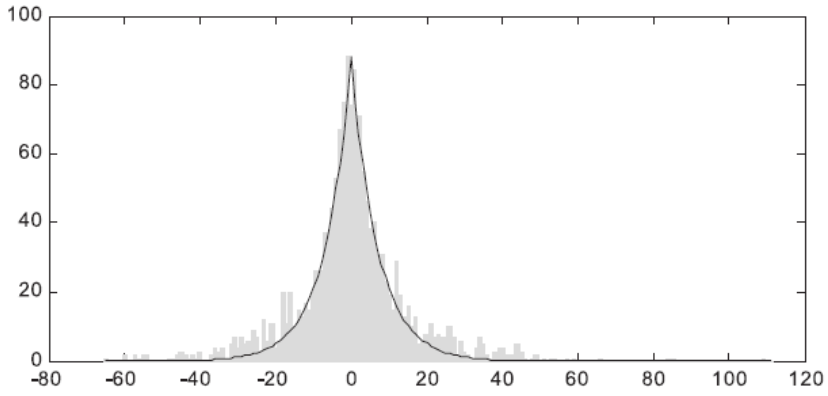
# Signal Models

- Set of representative waveforms are compared with acquired samples to find the best matching in terms of SAD or MSE

SAD: $$i = \arg\min\left( \sum_t | x[t,i] - x[t] | \right)$$

MSE: $$i = \arg\min\left( \sum_t \left( x[t,i] - x[t] \right)^2 \right)$$

$$x_{predicted}[t] = x[t,i]$$



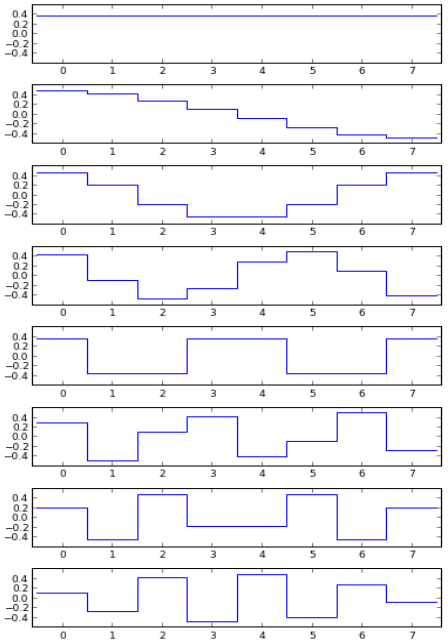

- Residuals (equal to difference between input samples and their predictions) have much lower values and energy
- In vector quantization residuals are neglected

# Transforms

- Karhunen-Loeve Transform (KLT)
  - Best efficiency expected
  - Computed based on a number of waveforms
  - Required similarity of signals to obtain better energy compaction
- DWT, FFT, and DCT seems to be less efficient

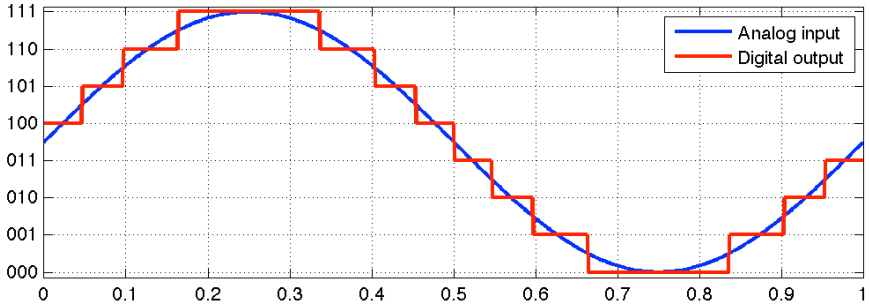

DWT base:


DCT base:



# Quantization

- Scalar Quantization – division by quantization step
- Scalar Dequantization – multiplication by quantization step
- Quantization step can be dependent on charge to keep sufficient SNR




- Possible to apply quantization from video coding
  - Quantization parameter (QP: 6 bits) determines quantization step
  - Increments decrease SNR by about 1dB
  - Division replaced by equivalent multiplication by multiple constants

Quantizer: $$X_q = sign\{X\} \cdot \left[\left(|X| \cdot A(QP\%6) + f \cdot 2^{17+QP/6}\right) \gg (17 + QP/6)\right]$$

Tables of constants

Dequantizer: $$X_r = sign\{X_q\} \cdot \left[X_q \cdot B(QP\%6) \ll QP/6\right]$$

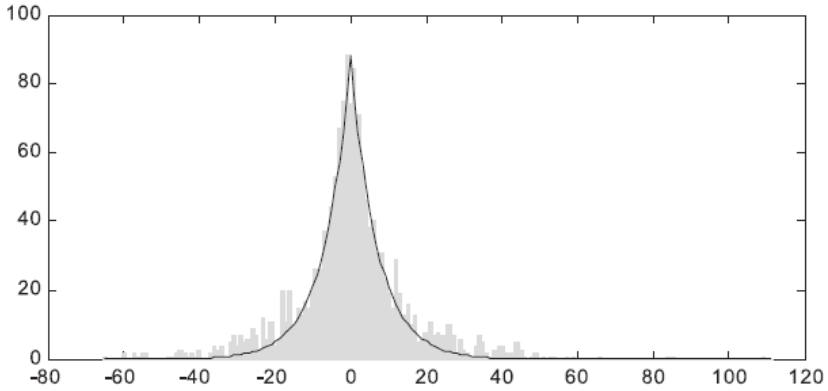
# Entropy coding (1)

- Assignment of input values to codewords
  - Codewords have variable lengths proportional to the logarithm of inversed probabilities of a symbol/value

    $L \approx \log(1/p)$
- Variable Length Coding:
  - Simple in implementation
  - Bit rate greater than the information entropy by a fraction of bit per sample
- Arithmetic Coding:
  - Higher implementation complexity
  - Achieve entropy $-\sum_{i=1}^{n} P(a_i) \log P(a_i) = H(S_{DMS})$

# Entropy coding (2)

- Golomb/Rice codes suitable for geometric distribution
- Exp-Golomb codes suitable for exponential distribution

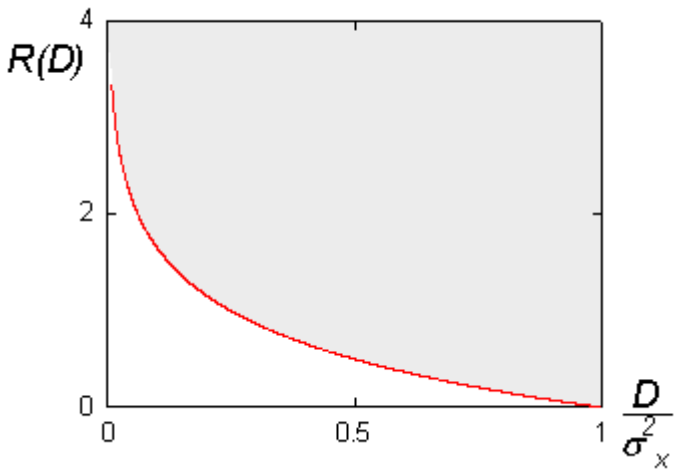



Golomb Codewords for different orders

| $i$ | $m=1$ | $m=2$ | $m=3$ | $m=4$ | $m=5$ | $m=6$ | $m=7$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0\|0 | 0\|0 | 0\|00 | 0\|00 | 0\|00 | 0\|00 |
| 1 | 10 | 0\|1 | 0\|10 | 0\|01 | 0\|01 | 0\|01 | 0\|010 |
| 2 | 110 | 10\|0 | 0\|11 | 0\|10 | 0\|10 | 0\|100 | 0\|011 |
| 3 | 1110 | 10\|1 | 10\|0 | 0\|11 | 0\|110 | 0\|101 | 0\|100 |
| 4 | 11110 | 110\|0 | 10\|10 | 10\|00 | 0\|111 | 0\|110 | 0\|101 |
| 5 | 111110 | 110\|1 | 10\|11 | 10\|01 | 10\|00 | 0\|111 | 0\|110 |
| 6 | 1111110 | 1110\|0 | 110\|0 | 10\|10 | 10\|01 | 10\|00 | 0\|111 |
| 7 | 11111110 | 1110\|1 | 110\|10 | 10\|11 | 10\|10 | 10\|01 | 10\|00 |
| 8 | 111111110 | 11110\|0 | 110\|11 | 110\|00 | 10\|110 | 10\|100 | 10\|010 |
| 9 | 1111111110 | 11110\|1 | 1110\|0 | 110\|01 | 10\|111 | 10\|101 | 10\|011 |
| 10 | 11111111110 | 111110\|0 | 1110\|10 | 110\|10 | 110\|00 | 10\|110 | 10\|100 |
| ⋮ | … | … | … | … | … | … | … |

# Compression Efficiency

- Lossless Coding of waveforms
  - Compression ratio: about 2-6
  - Depends on SNR, sampling frequency, signal dynamics
- Lossy Coding of waveforms
  - Compression ratio: more than 3, e.g. 10, 20 …
  - Distortion (D) and bit rate (R) depend on quantization step
  - RD Tradeoff
  - Allowable losses should be lower than signal noise

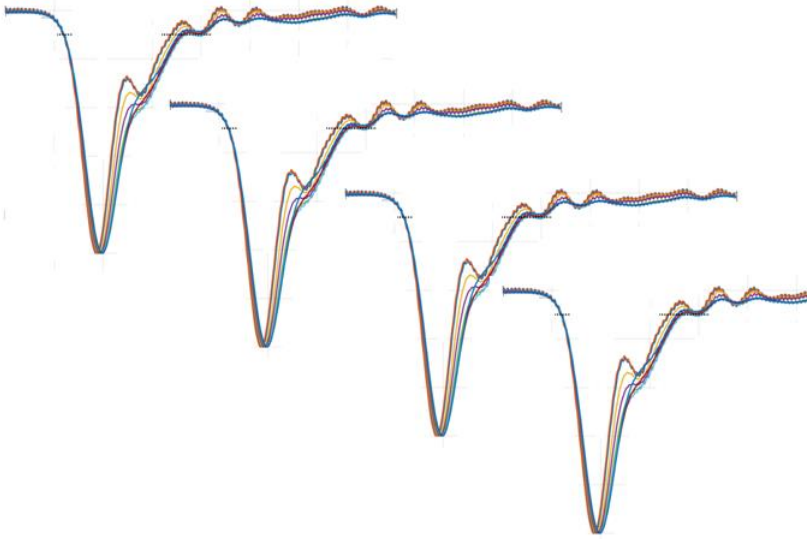



More accurate estimation of compression ratios after the statistical analysis

# Multi-channel Compression

- Neighboring PMTs may be excited in similar moments in the case of Cherenkov photons
  - common packets where one-bit flags can indicate the presence of the hit in each channel
  - each separate time descriptor consumes 27 bits
  - common time descriptor (offset) for 19 channels is useful
  - Time Delta values for each channel should be close to zero ->suitable variable length coding

- Waveforms from neighboring PMTs may be similar
  - Use of one waveform to predict others



# Data in Super-Kamiokande (SK)

Hit-DATA cell Format

| | Bit 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1st_Word | Ch# (HEADER) | | | | | QTC gate count | | | | | | | | | | |
| 2nd_word | Range | | TRG | | | Event # | | | | | | | | | | |
| 3rd_Word | TDC count | | | | | | | | | | | | | | | |

48 bits

| | Bit Location | Length | Function |
|---|---|---|---|
| Ch # | MSB: 1st_Word Bit15<br>LSB: 1st_Word Bit11 | 5bits | Channel Number:<br>0～23 (00000B～10111B) |
| QTC gate count | MSB: 1st_Word Bit10<br>LSB: 1st_Word Bit0 | 11bits | QTC gate count value:<br>0～2047 (000H～7FFH)→0～1064ns, 0.52ns/bit |
| Range | MSB: 2nd_Word Bit15<br>LSB: 2nd_Word Bit14 | 2bits | Range code:<br>00B = small, 01B = medium, 10B = large |
| TRG | MSB: 2nd_Word Bit13<br>LSB: 2nd_Word Bit12 | 2bits | Trigger ID:<br>00B = Narrow, 01B = Wide,<br>10B = Pedestal, 11B = Calibration |
| Event # | MSB: 1st_Word Bit11<br>LSB: 2nd_Word Bit0 | 12bits | TDC Event Number:<br>0～4095 (000H～FFFH) |
| TDC count | MSB: 3rd_Word Bit15<br>LSB: 3rd_Word Bit0 | 16bits | TDC count (T1 leading edge):<br>0～65535 (0000H～FFFFH) |

- Time: Event + TDC count = 28 bits
- Charge: QTC gate count = 11 bits

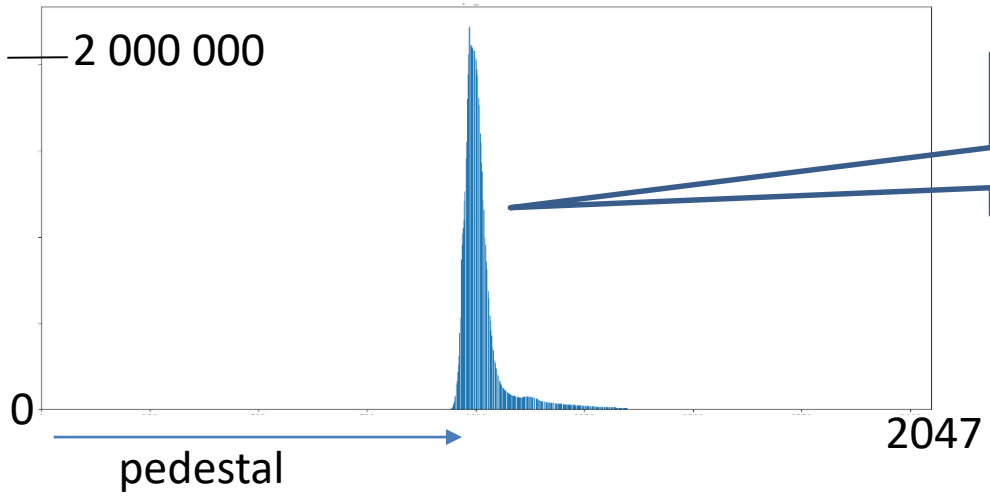
# Time-Stamp Compression

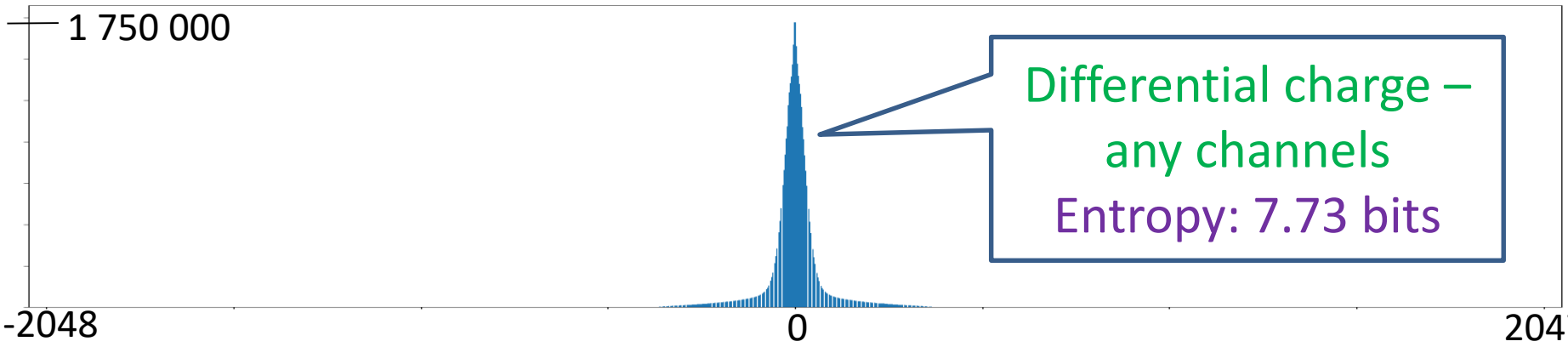
- Efficiency limited by the entropy
- Differential coding
  - Difference between successive time stamps of any channel
  - Data dominated by dark counts
- Division of bits into two parts: variable-length code (VLC) and fixed-length code
  - More bits to VLC -> better compression and complex code

| Division of bits | MSB Entropy + fixed length | Entropy | Entropy gain |
|---|---|---|---|
| 12/15 | 1.1728 +15 | 16.1728 | |
| 13/14 | 1.7035 +14 | 15.7035 | -0.4693 |
| 14/13 | 2.3766 +13 | 15.3766 | -0.7962 |
| 15/12 | 3.1632 +12 | 15.1632 | -1.0096 |
| 16/11 | 4.0454 +11 | 15.0454 | -1.1274 |
| 17/10 | 4.9914 +10 | 14.9914 | -1.1814 |
| 18/9 | 5.9600 + 9 | 14.9600 | -1.2128 |
| 19/8 | 6.9390 + 8 | 14.9390 | -1.2338 |
| 20/7 | 7.9216 + 7 | 14.9216 | -1.2512 |
| 21/6 | 8.9051 + 6 | 14.9051 | -1.2677 |
| 22/5 | 9.8891 + 5 | 14.8891 | -1.2837 |
| 23/4 | 10.8736 + 4 | 14.8736 | -1.2992 |
| 24/3 | 11.8582 + 3 | 14.8582 | -1.3146 |
| 25/2 | 12.8426 + 2 | 14.8426 | -1.3302 |
| 26/1 | 13.8266 + 1 | 14.8266 | -1.3462 |
| 27/0 | 14.8106 + 0 | 14.8106 | -1.3622 |

# Charge Compression (1)

- 11 bits in original representation

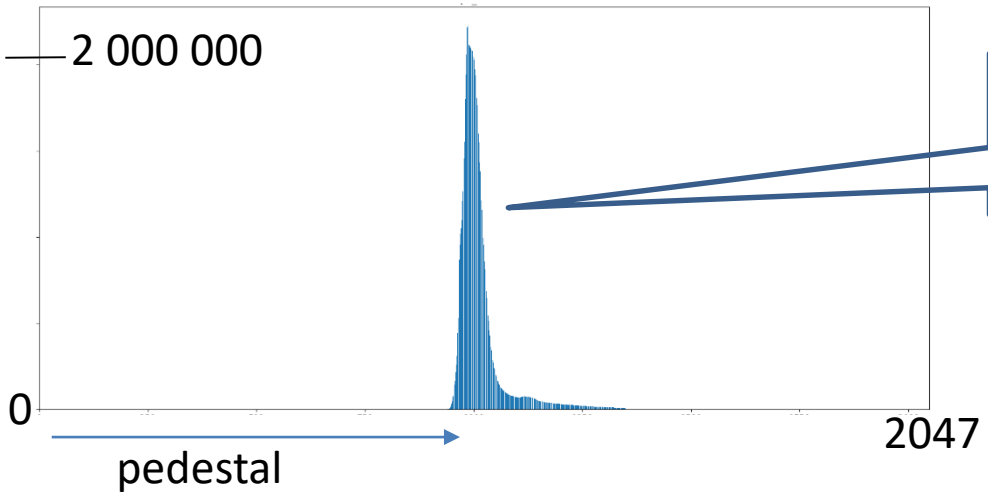



Other predictions will be searched to improve entropy

# Charge Compression (2)

- 11 bits in original representation




Huffman Coding

Bit-Rate: 6.996 bits

Simplified Code Table:

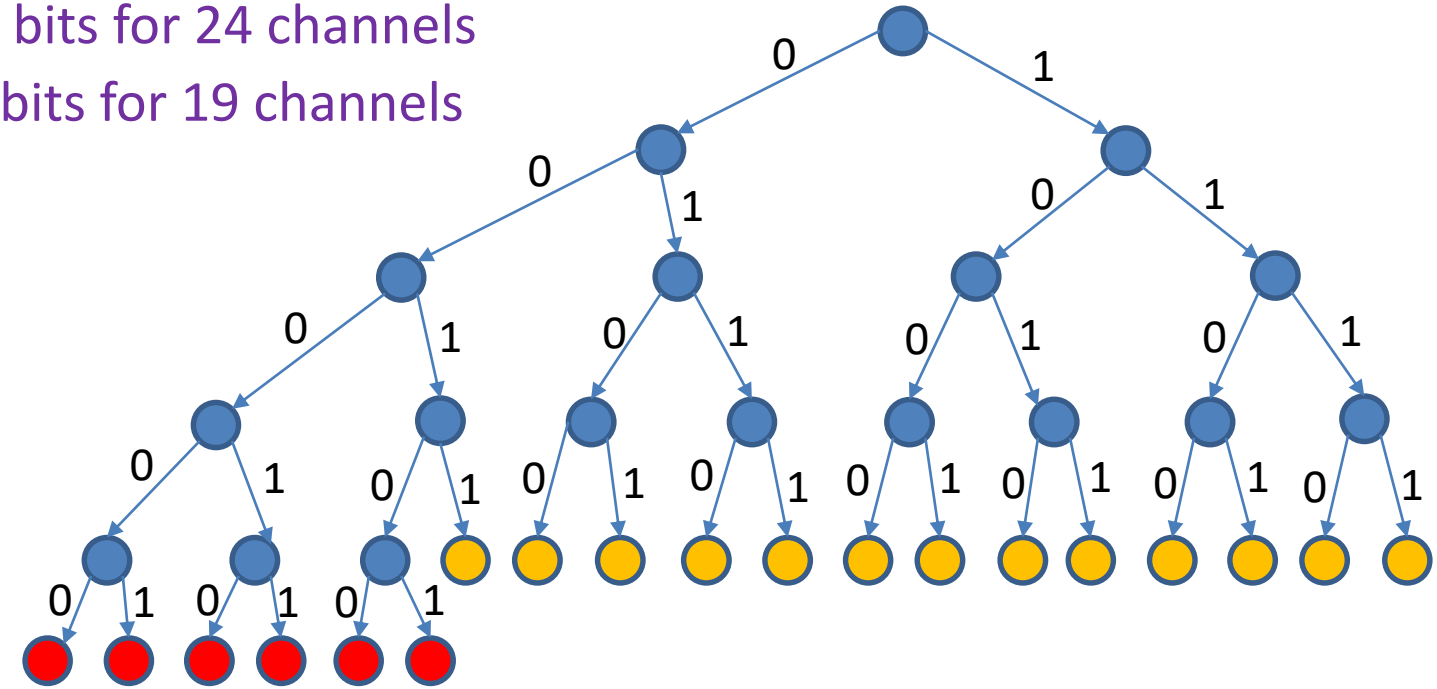
Bit-Rate: 7.466 bits

Loss: 0.49 bits

| Subrange | Prefix | Suffix | Length |
|---|---|---|---|
| 0-947 | 1110 | +10 bits | =14 bits |
| 948-979 | 110 | +5 bits | =8 bits |
| 980-1011 | 0 | +5 bits | =6 bits |
| 1012-1075 | 10 | +6 bits | =8 bits |
| 1076-2047 | 1111 | +10 bits | =14 bits |

# Channel number

- Identification one of 24/19 channels in mPMT
  - Equal probabilities prevent the compression gain
  - Fixed-length codes require 5 bits
  - Almost fixed-length codes uses 4-bit and 5-bit codewords. Average bit rate is:
    - 4.66 bits for 24 channels
    - 4.32 bits for 19 channels



# Triger Type and Range

- Originally coded with 4 bits
  - Three ranges of signal dynamic: small/medium/large
  - Four trigger types: narrow/wide/pedestal/calibration

| Statistics in SK | Small (S) | Medium (M) | Large (L) | All |
|---|---|---|---|---|
| Narrow (N) | 48 | 0 | 0 | 48 |
| Wide (W) | 122653704 | 850692 | 494 | 123504890 |
| Pedestal (P) | 438115 | 437982 | 437887 | 1313984 |
| Calibration (C) | 0 | 0 | 0 | 0 |

- Common code built with the Huffman method

| | | | |
|---|---|---|---|
| 0 | W_S | 111110 | N_S |
| 100 | W_M | 11111100 | N_M |
| 101 | P_S | 11111101 | N_L |
| 110 | P_M | 11111110 | C_S |
| 1110 | P_L | 111111110 | C_M |
| 11110 | W_L | 111111111 | C_L |

- Entropy: 0.16 bits
- Bit-Rate: 1.0581 bits

# Summary

- A number of compression methods must be examined for signal waveforms
  - The level of loss must be decided
- The compression of extracted parameters allows reduction 48 bits ->28 bits ≈ 0.58 ratio
  - Optimized methods can slightly improve the ratio
- Compression oriented to dark counts