5G Air Interface – Part I
from NRZ to OFDM

OFDM

The LTE air interface uses **O**rthogonal **F**requency **D**omain **M**odulation which is the modulation type used by almost all modern systems, including:

- ADSL and VDSL (where it is called **D**iscrete **M**ulti**t**one **T**ransmission)
- G.hn home networking over twisted pair
- WiFi 802.11g (54 Mbps) and 802.11n (up to 600 Mbps)
- **D**igital **V**ideo **B**roadcasting – **T**errestial

OFDM is now preferred over other modulation techniques due to its efficiency and flexibility

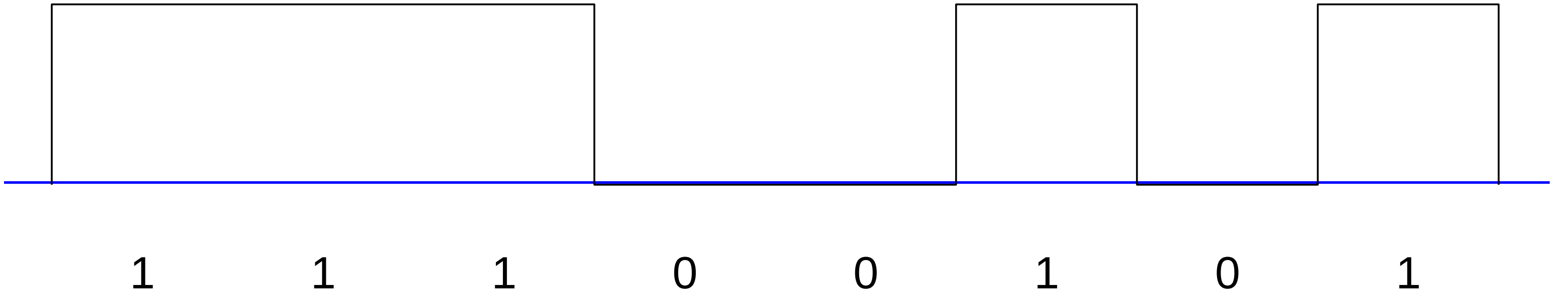
OFDM was invented and patented by Chang (Bell Labs) in 1966 (US3,488,445)
extended by Salzberg (Bell Labs) in 1967
made practical by Weinstein and Ebert (Bell Labs) in 1971
with further improvements by Cioffi (Stanford/Amati) in 1991

We have already mentioned FDM as the mux method used in the analog PSTN

FDM, along with any bandwidth-limited modulation method can be used to construct an efficient wideband modulation scheme that can be naturally extended to a multiple access scheme

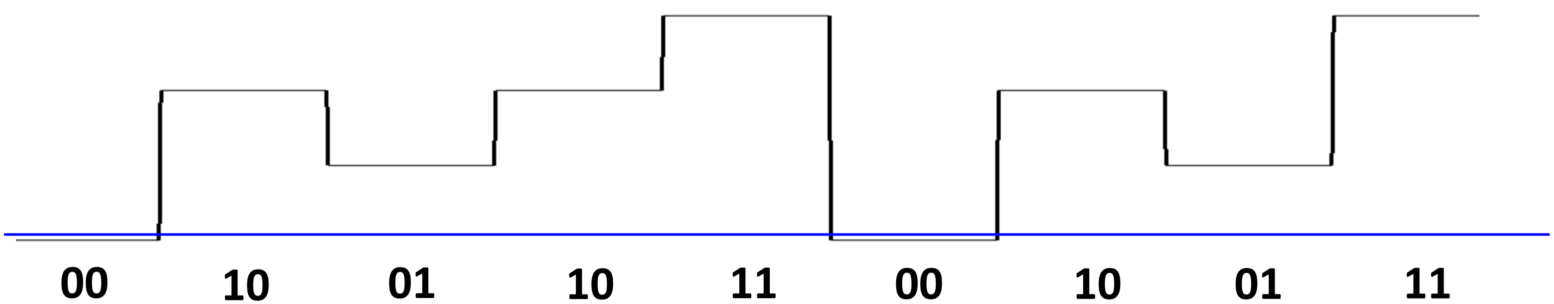
NRZ and PAM

The simplest modulation technique is **N**on**R**eturn to **Z**ero (NRZ)
which simply encodes 0 or 1 using 2 possible signal values (e.g., 0 or 1)



NRZ is only applicable to media that can pass DC (base-band)

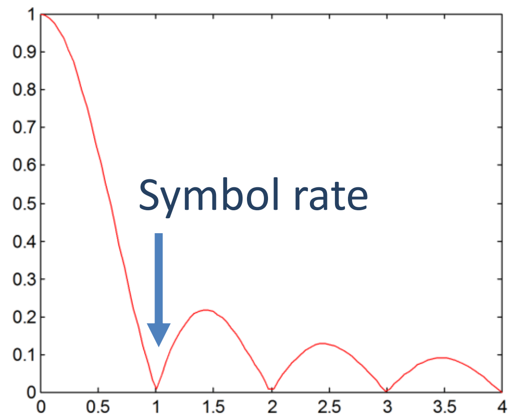
NRZ can be extended to multi-level **P**ulse **A**mplitude **M**odulation
where each level represents a *symbol* (AKA *baud* after Jean-Maurice-Émile Baudot)



The number of bits per symbol is $\log_2$(number of different symbols)

NRZ/PAM spectrum

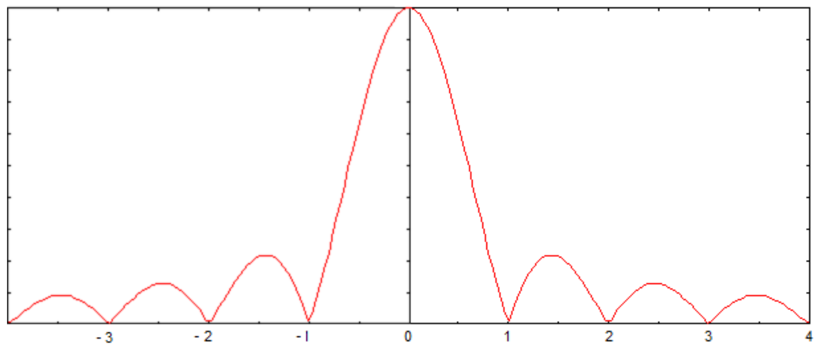
The spectrum of NRZ or PAM is (from Wiener-Khintchin)



The first zero is at the symbol rate

independent of the number of bits per symbol!

If we show the negative frequencies (not very interesting for real signals) ...



InterSymbol Interference

One problem impacting all modulation techniques is ISI

ISI is when one information-bearing symbol interferes with another

There are two main causes of ISI

1. limited bandwidth
   multiplying by a window in the *frequency domain*
      is equivalent to a *convolution* in the *time domain*
   which means that each symbol spreads into following symbol(s)

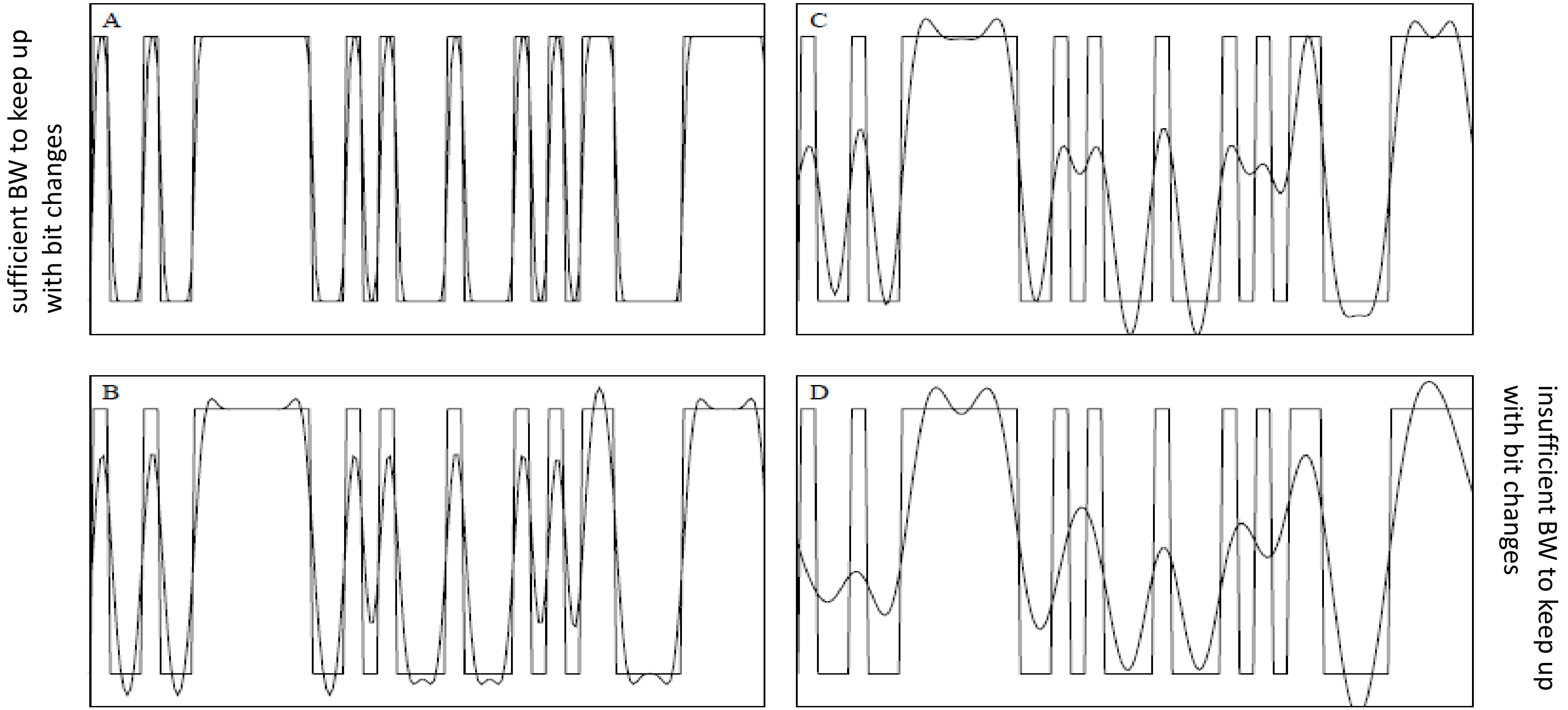
2. multipath
   when a signal propagates in space over multiple paths of different length
      a delayed symbol may be received during a subsequent symbol
   this results in a convolution
      with coefficients nonzero for physical delay differences

ISI increases with symbol rate
   since a given delay overlaps more symbols at high rates

Limited Bandwidth ISI

What happens if we try to squeeze an NRZ signal
into a channel with insufficient bandwidth?

The signal exiting the channel simply can't vary fast enough
resulting in ISI

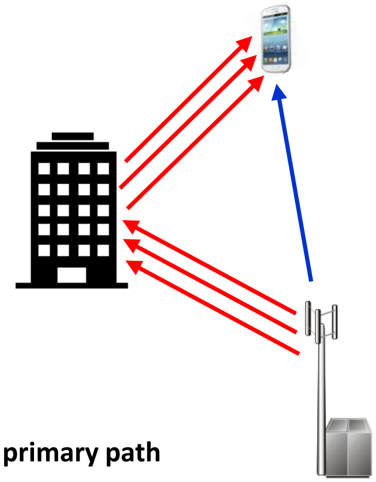


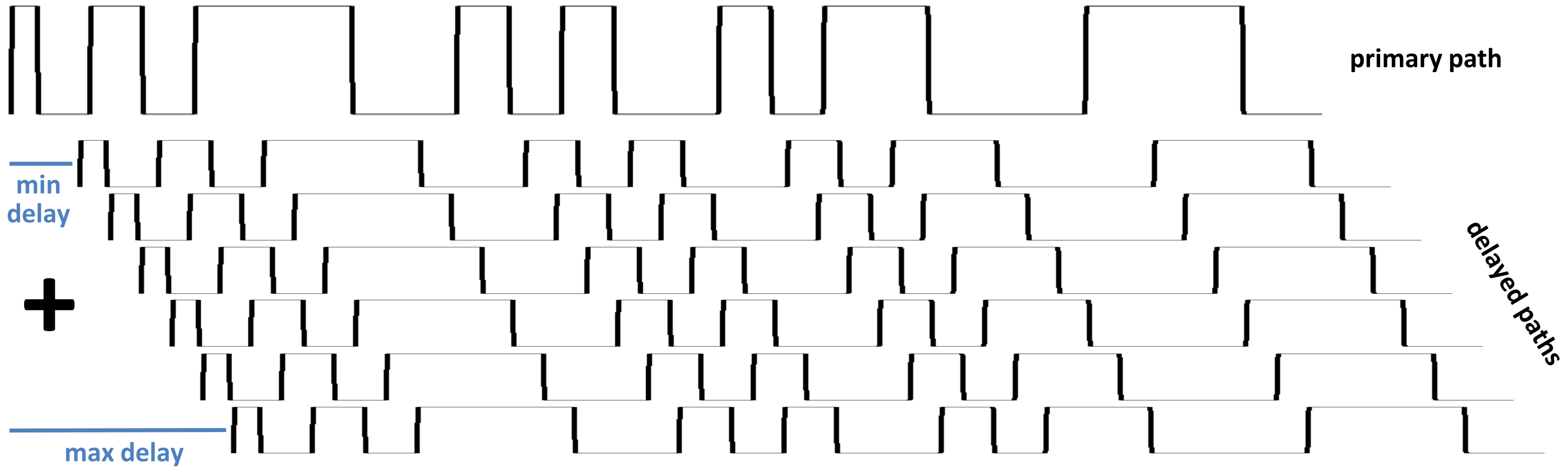
Multipath ISI

What happens when a signal is received over a primary (shortest) path and also over delayed paths (e.g. reflections off buildings) ?

The composite signal displays ISI

Note that this ISI may change rapidly for mobile users



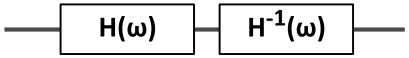


Combatting ISI

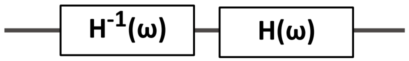
How can we combat ISI ?

There are basically 5 strategies, each with challenges/disadvantages

1. leave large *guard times* between symbols
   – inefficient since it wastes time



2. using an *equalizer* at the receiver (in time or frequency domain)
   – apply an inverse filter to compensate for the channel response
   – equalizers generally need training phases
   – linear equalizers suffer from noise amplification at near-zeros (...)
   – decision feedback equalizers need to close a large loop
3. Tomlinson-Harashima *precoding*

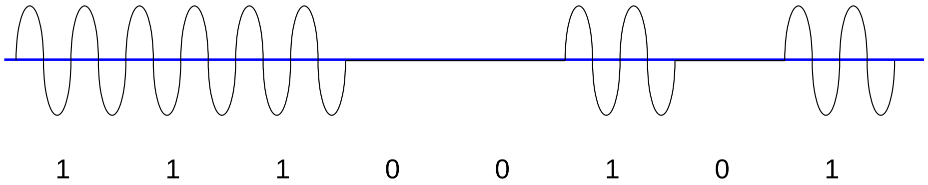


   – THP applies the inverse filter before transmission
   – but requires a feedback channel to learn the channel response
4. using very low symbol rates (*OFDM* exploits this!)
   – (similar to 1.) may eliminate ISI but seems to strongly constrain data-rate
5. trellis encoding and using the Viterbi algorithm to decode
   – requires encoding transmitted symbols and complex decoding

OOK and ASK

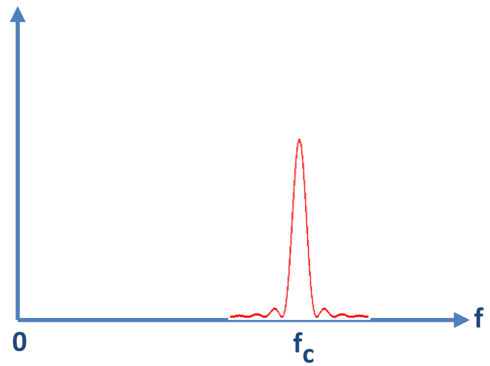
On-**O**ff **K**eying is formed by AM modulating a carrier frequency f_c with NRZ
(alternatively you can think of this as frequency-shifting NRZ)

OOK is applicable to arbitrary media (pass-band)



OOK can be extended to multi-amplitude **A**mplitude **S**hift **K**eying

The spectrum of OOK or ASK is (from frequency shifting NRZ's spectrum)



with spread between zeros of twice the symbol rate

The extra bandwidth can be utilized by more efficient modulation

Limitations on data rate

We saw that since the spectral width equals the symbol rate
(we can make do with the primary spectral lobe)
the symbol rate must not exceed the physical bandwidth

But the spectrum of PAM is independent of the number of bits per symbol
so why can't we use arbitrarily many bits per symbol ?

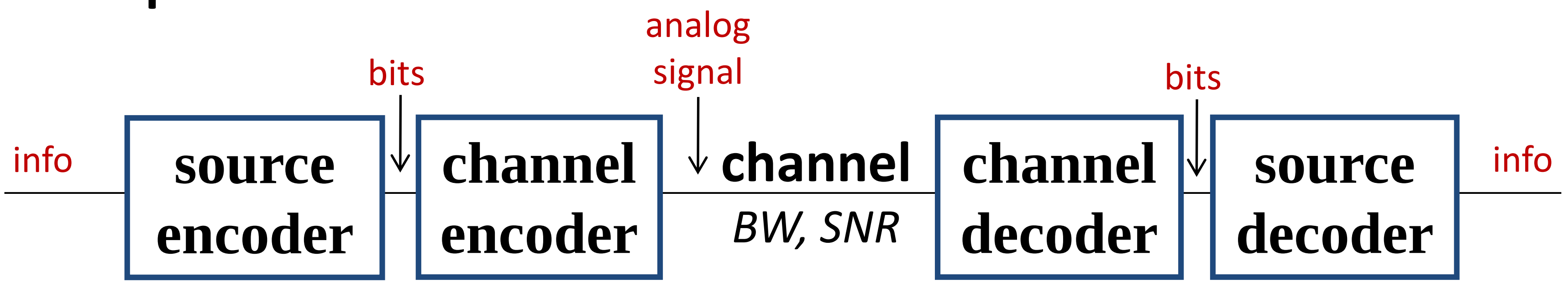
The problem here is noise –
if the symbol values are closer than the noise level
then we won't distinguish between them

To summarize:
if there is no limit on the signal's physical bandwidth or
if there is no noise added to the signal between transmitter and receiver
then there is no limitation on the number of bits/sec we can transfer

But, any physical channel with finite bandwidth and non-zero noise level
has a maximum capacity (Shannon's *channel capacity theorem*)

Shannon's Theorems

1. **Separation Theorem**



2. **Source Encoding Theorem**

Information can be quantified (in bits)

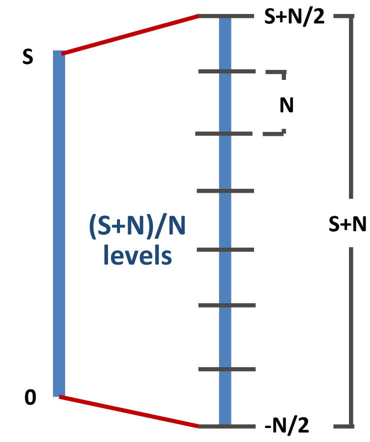
3. **Channel Capacity Theorem**

$C = BW \log_2 (SNR + 1)$

Proof of Shannon's capacity theorem

A simplistic justification of the capacity theorem is as follows:

- assume (*wolg*) that we transmit some signal with values between **0** and **S**
- assume that the channel adds DC-less noise with uniform distribution
  we'll call the peak-to-peak noise N
  so that the noise values are between –N/2 and +N/2
- the receiver always sees values between –N/2 and S+N/2
  so the receiver's dynamic range is S+N
- to maximize the information per symbol w/o overlap
  we space the signal levels by N
- so there can be (S+N)/N = S/N + 1 = SNR+1 different symbols
- hence each symbol contains $\log_2(\mathrm{SNR} + 1)$ bits
- but there can be BW symbols per second



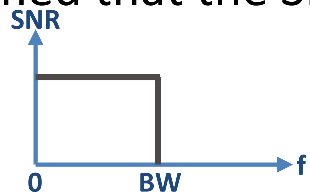
Hence, the maximum information rate $C = BW \log_2(\mathrm{SNR} + 1)$ bit/s

The maximum spectral efficiency is $C/BW = \log_2(\mathrm{SNR} + 1)$ bit/s/Hz

Capacity for frequency-dependent SNR

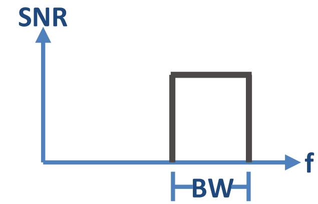
The capacity theorem assumed that the SNR was

- constant from DC to BW
- zero over BW



or more generally

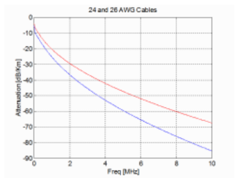
- constant in some passband
- zero outside the passband

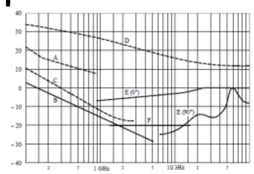


Which will not be the case if either

- the signal attenuation (including multipath cancellation)
- the noise (including interference)

or both, vary with frequency



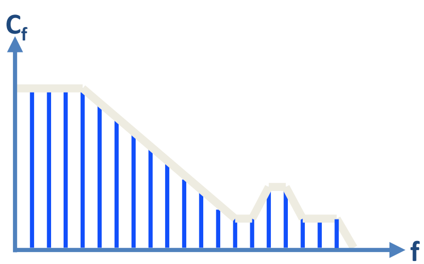


The extension of the theorem is simple

- divide the passband into channels of bandwidth Δf centered at f
- the capacity theorem states that for the channel $C_f \approx \log_2(SNR(f) + 1)\ \Delta f$
- the composite capacity is $\sum_f \log_2(SNR(f) + 1)\ \Delta f$

The theorem becomes exact when we send $\Delta f \rightarrow 0$

and then $C = \int \log_2(SNR(f) + 1)\ df$

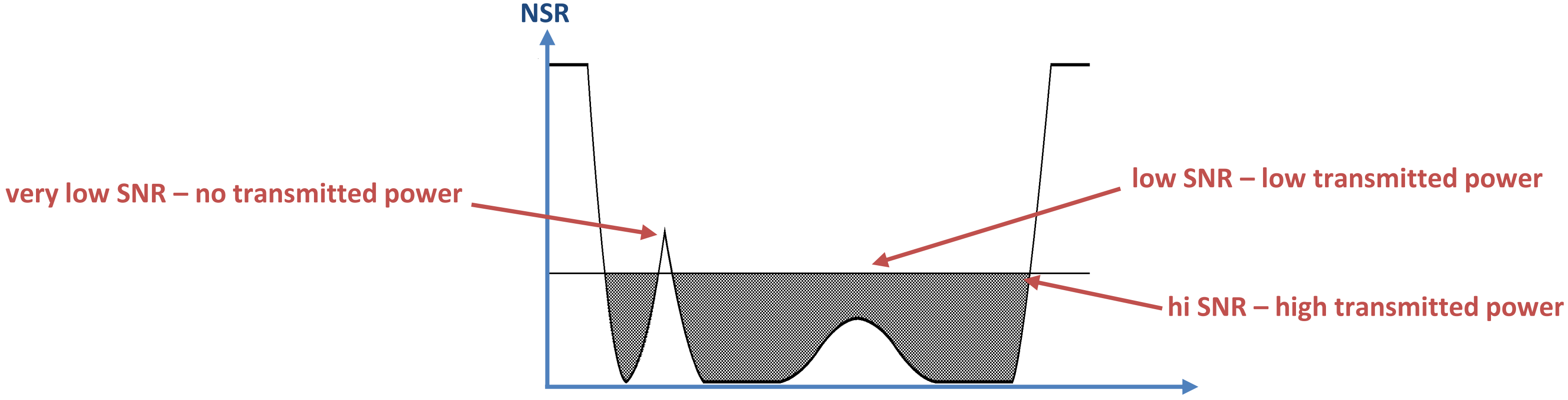


Water pouring theorem

The converse to Shannon's channel capacity theorem
is Gallager's water pouring theorem

The spectrum of the optimum (capacity-achieving) modulated signal
is given by the *water pouring* algorithm

- fill a pitcher with a volume of water
  representing the total power to be transmitted
- pour the water over the reciprocal SNR(f) graph



This power spectrum can be achieved by

- spectral shaping
- explicit power loading

To FEC or not to FEC

Shannon's separation theorem implies

that it is suboptimal to use *bit-layer* Forward Error Correction schemes

which is why Ungerboeck developed **T**rellis **C**oded **M**odulation

Yet, all mobile systems are full of error detection/correction codes:

- CRC (Cyclic Redundancy Code) at MAC layer before passed to physical layer
- repetation codes for very small data blocks
- TBCC (Tail-Biting Convolutional Codes) or polar codes for control channels
- turbo codes or LDPC (Low Density Parity Code) for data channels
- HARQ (Hybrid Automatic Repeat Request)

This is because the SNR is often not stationary

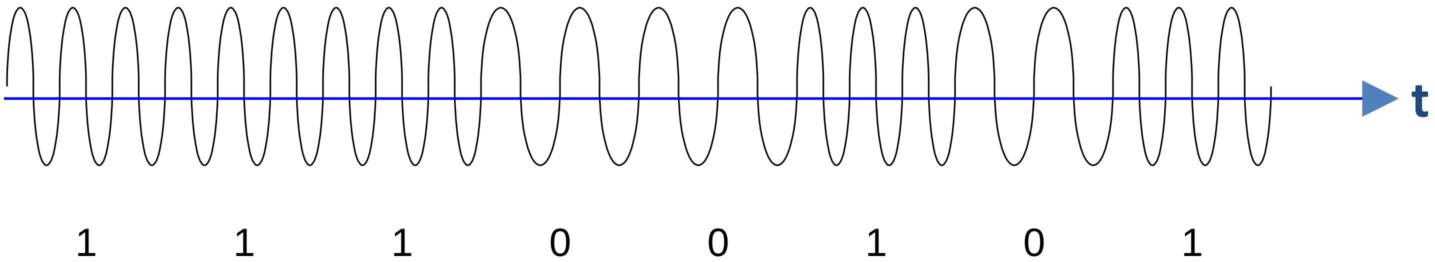
in contrast with the assumptions of the capacity theorem

FSK

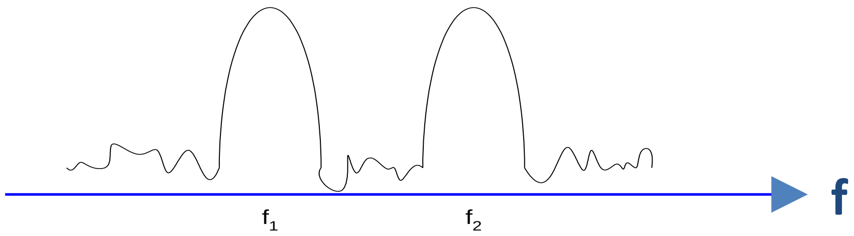
What can we do about noise?

We can gain resilience by using *frequency diversity*

i.e., using two independent OOKs with basically the same information



This is called **F**requency **S**hift **K**eying

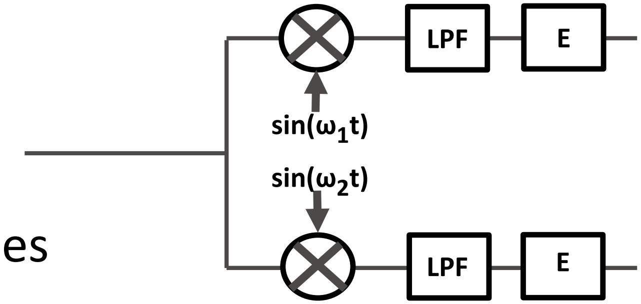


We can extend this to multiple tones (mFSK) with N bits/tone (symbol)

The problem with FSK

Note that we are exploiting the fact
  that sinusoids of different frequencies are *orthogonal*

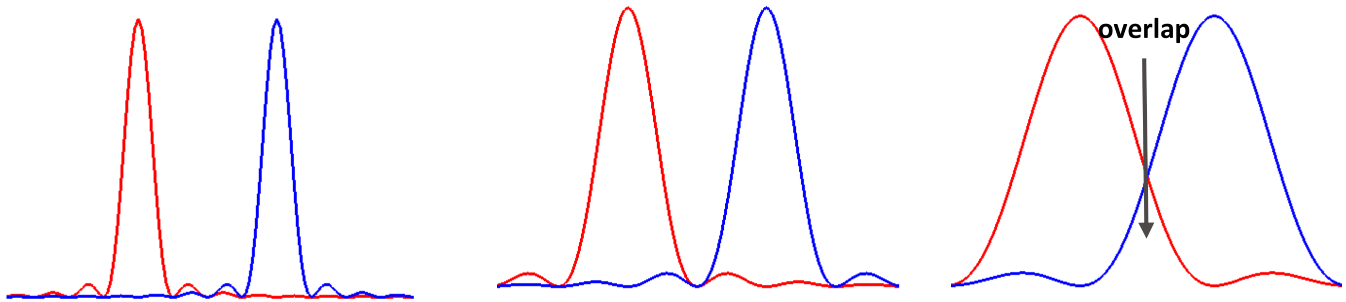
The optional receiver
  mixes down to zero
  filters
  and compares energies



So, why can't we use many frequencies to bypass Shannon?

Because sinusoids of different frequencies are only orthogonal
  when we integrate over all time (from -∞ to + ∞)

If each bit only lasts a short time
  the uncertainty theorem says that its frequency is uncertain
    so we can't be sure which tone was transmitted!



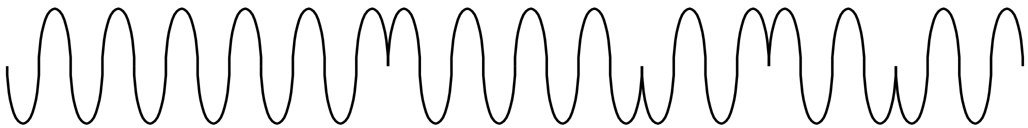
PSK

But we can differentiate between sinusoids of the same frequency - but different phases after a single cycle !

$$\int \sin(\omega t)\cos(\omega t)\, dt = 0$$

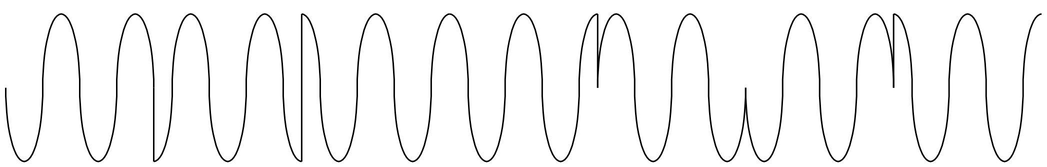
This is the basis of **P**hase **S**hift **K**eying

BPSK has 1 bit / symbol



1 1 1 0 0 1 0 1

QPSK has 2 bits / symbol

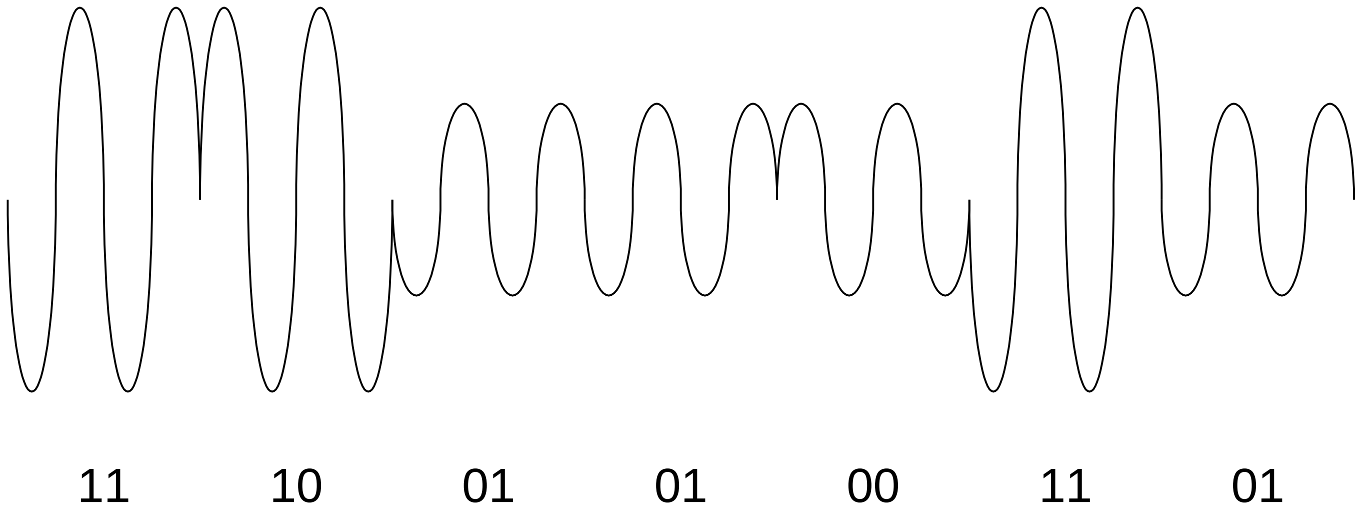


11 10 01 01 00 11 01

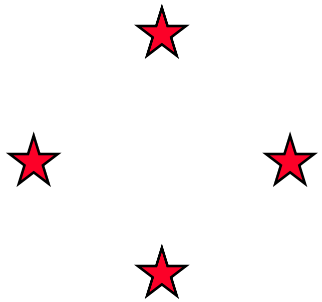
QAM

Finally, we can combine PSK and ASK (but not FSK)
to get **Q**uadrature **A**mplitude **M**odulation

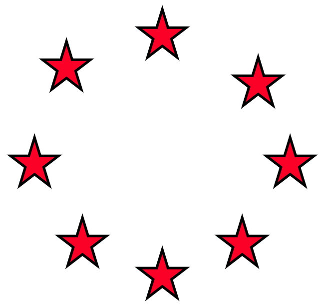
2-QAM has 4 different symbols (+sin, -sin, +3sin, -3sin)
and thus 2 bits/symbol



Some *constellations*



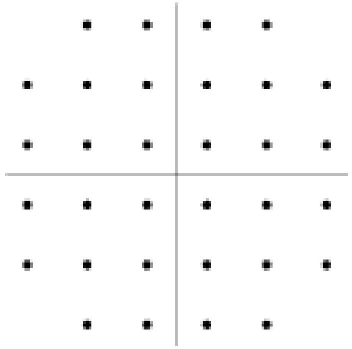
QPSK
2 bits/symbol



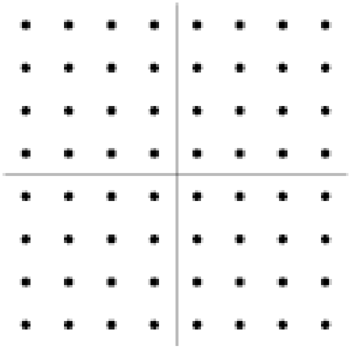
8PSK
3 bits/symbol

16QAM
4 bits/symbol

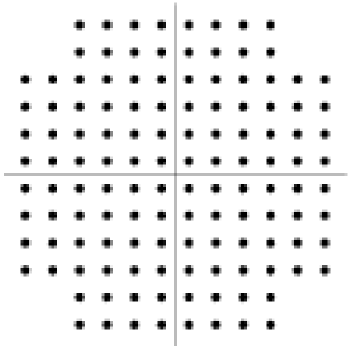
V.29
4 bits/symbol



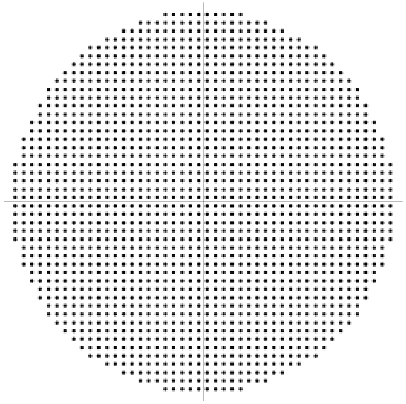
V.32 9600 bps
5 bits/symbol



V.32 12000 bps
6 bits/symbol



V.32 14400 bps
7 bits/symbol



V.34 33600 bps

Spectral efficiency of simple modulations

We defined spectral efficiency as the bit/s per Hz

We can now compare the modulation techniques we have seen so far

modulation	bit/symbol	BW/symbol rate	spectral efficiency
NRZ	1	1	1
BPSK	1	2	0.5
QPSK	2	2	1
8PSK	3	2	1.5
16QAM	4	2	2
64QAM	6	2	3
256QAM	8	2	4

Note that we are using *raw* (DSB) efficiencies

we will see that the values for PSK and QAM can be improved by a factor of 2 !

FDM

QAM (including PSK as a special case) is a very efficient modulation
approaching the Shannon capacity for simple bandpass channels

But the spectrum of a QAM signal is inflexible
like all discretely keyed modulations of a single carrier
it is a sinc centered at f_c with first zeros at $f_c \pm$ symbol-rate

The spectrum can be shaped using a Tomlinson precoder
but this requires a feedback channel and precomputing the filter

So, QAM does not lend itself to water-pouring

The trick is to use Frequency Domain Multiplexing

We saw FDM as a technique to mux different information sources

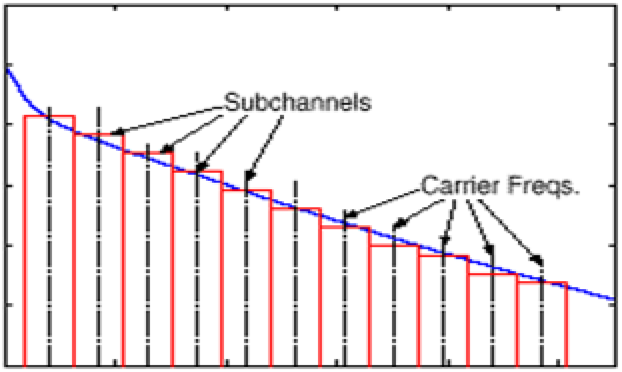
Here we divide a single information stream into blocks of bits
and mux them together using distinct carrier frequencies (sub-carriers)

Each sub-carrier signal can
- have its own power level
- use its own modulation technique (PSK, QPSK, 16QAM, 64QAM, ...)

thus directly implementing water pouring

FDM combats ISI but creates ICI

FDM uses many subchannels, each with low bandwidth but low data rate



Since the data rate is low, there is essentially no ISI

And since each subchannel is localized in frequency
we can perform equalization in the frequency domain (FEQ)
i.e., simply multiply each frequency and shift its phase

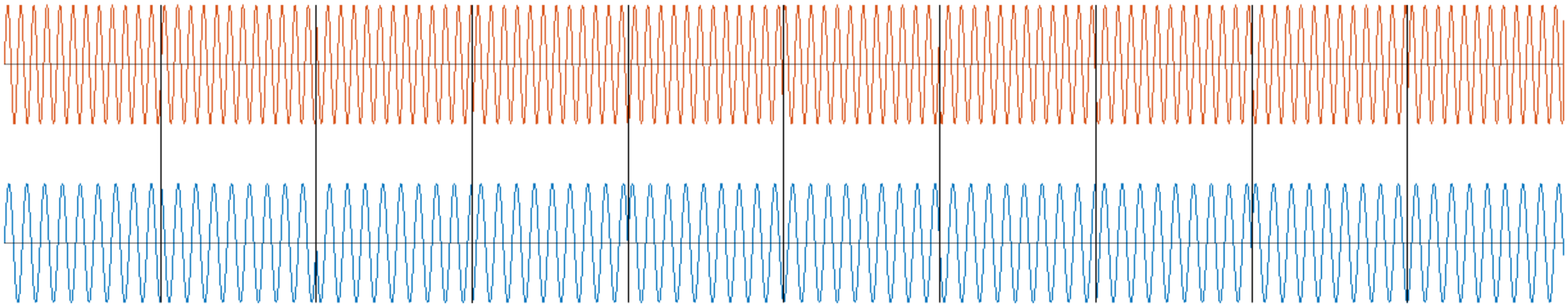
But, we need to space the sub-carriers far enough apart
to avoid **I**nter**C**hannel **I**nterference

This squanders bandwidth, distancing us from attaining the Shannon capacity
(similar to what we saw with FSK)

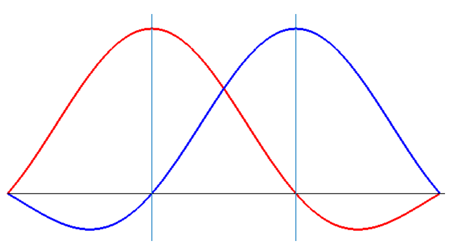
OFDM

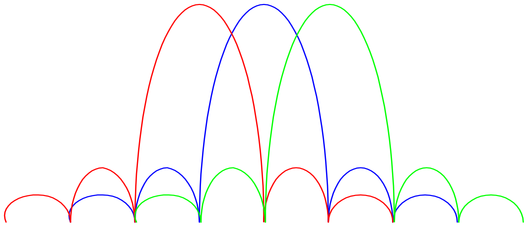
The solution is called Orthogonal FDM (OFDM)

- all sub-channels use the same symbol rate (even if different modulations)
- sub-carriers are spaced at precisely the symbol rate
- the sub-carriers are the precisely orthogonal
  and hence do not interfere with each other



- ICI is eliminated with no guard frequencies needed
- simple implementation based on the FFT



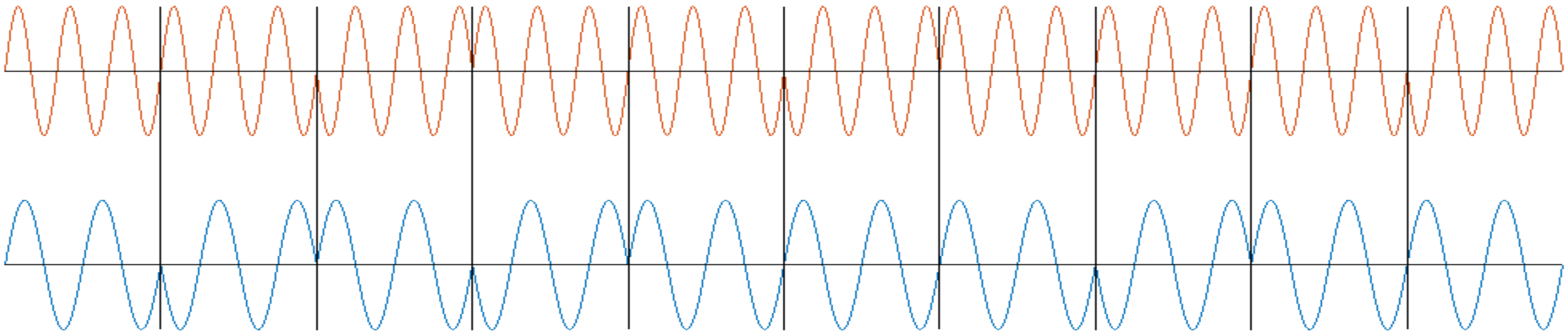


Why are the channels orthogonal?

Let's look at the baseband signal

where all sub-carriers are multiples of the symbol rate

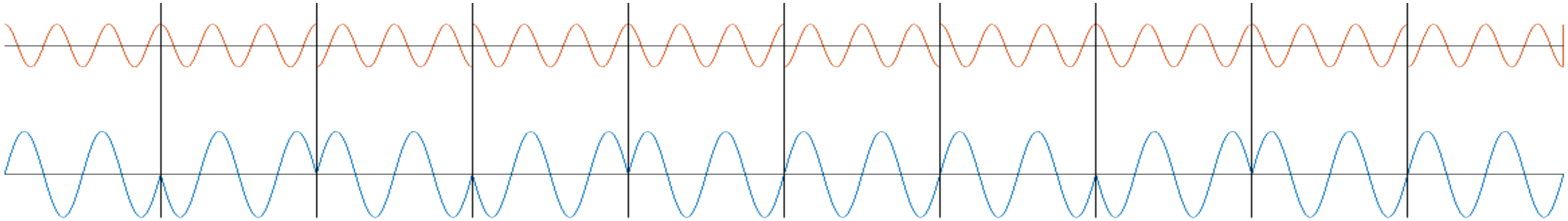
so that there are an integral number of sinusoid cycles in a symbol



Any two such signals are precisely orthogonal

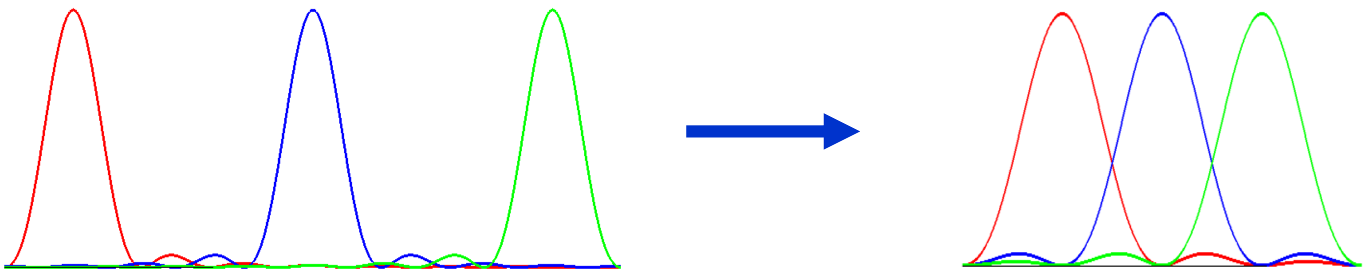
$\int_0^{\frac{2\pi}{\omega}N} e^{in_1\omega t}\, e^{-in_2\omega t}\, dt = 0$ (the only requirement is for a whole number of cycles of sin **Δω** t !)

and the same is true if we arbitrarily phase shift or amplify the signals



Spectral efficiency

OFDM provides the minimum possible sub-carrier spacing
and hence eliminates the need for *guard bands*



Now each sub-channel effectively occupies only symbol-rate of bandwidth
So the spectral efficiencies improve by a factor of 2 !

modulation	**bit/symbol**	**BW/symbol rate**	**spectral efficiency**
BPSK	1	1	1
QPSK	2	1	2
8PSK	3	1	3
16QAM	4	1	4
64QAM	6	1	6
256QAM	8	1	8

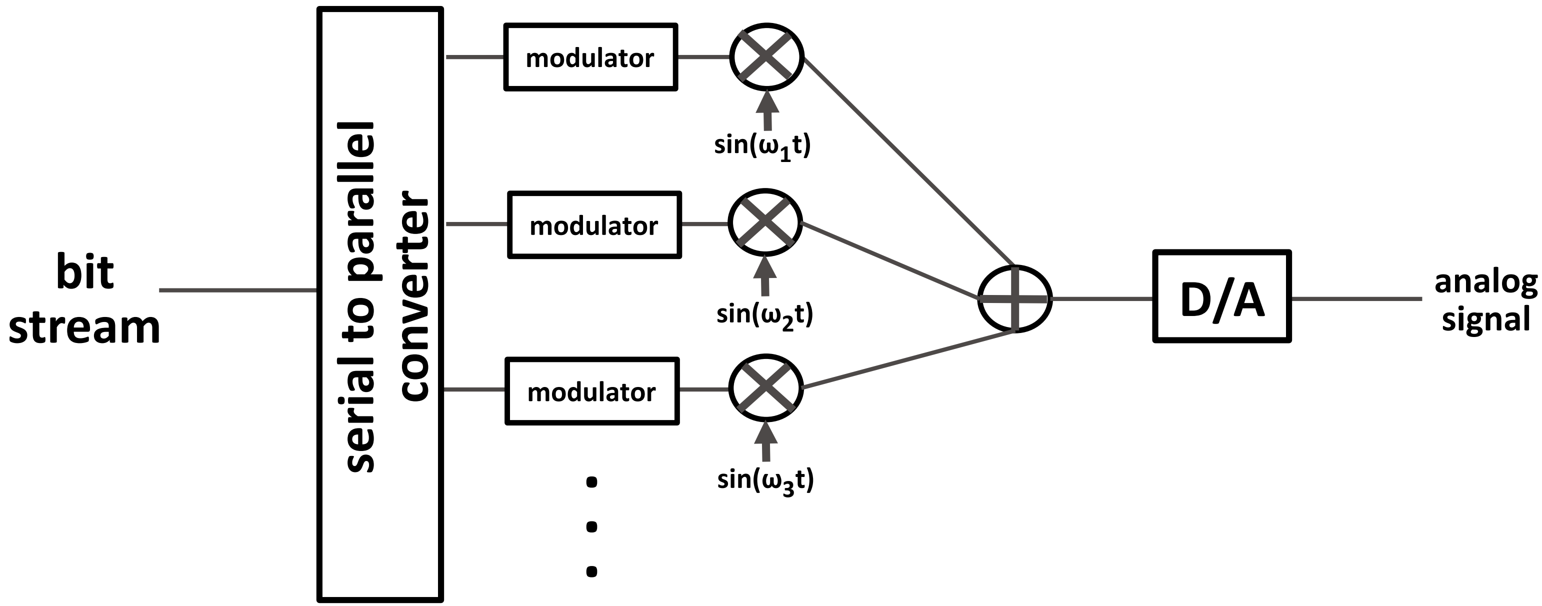
FFT

In order to transmit each sub-channel

we should modulate a baseband signal

up-mix each to the desired sub-carrier frequency (multiply by $e^{i\omega_c t}$)

and add the sub-channels together

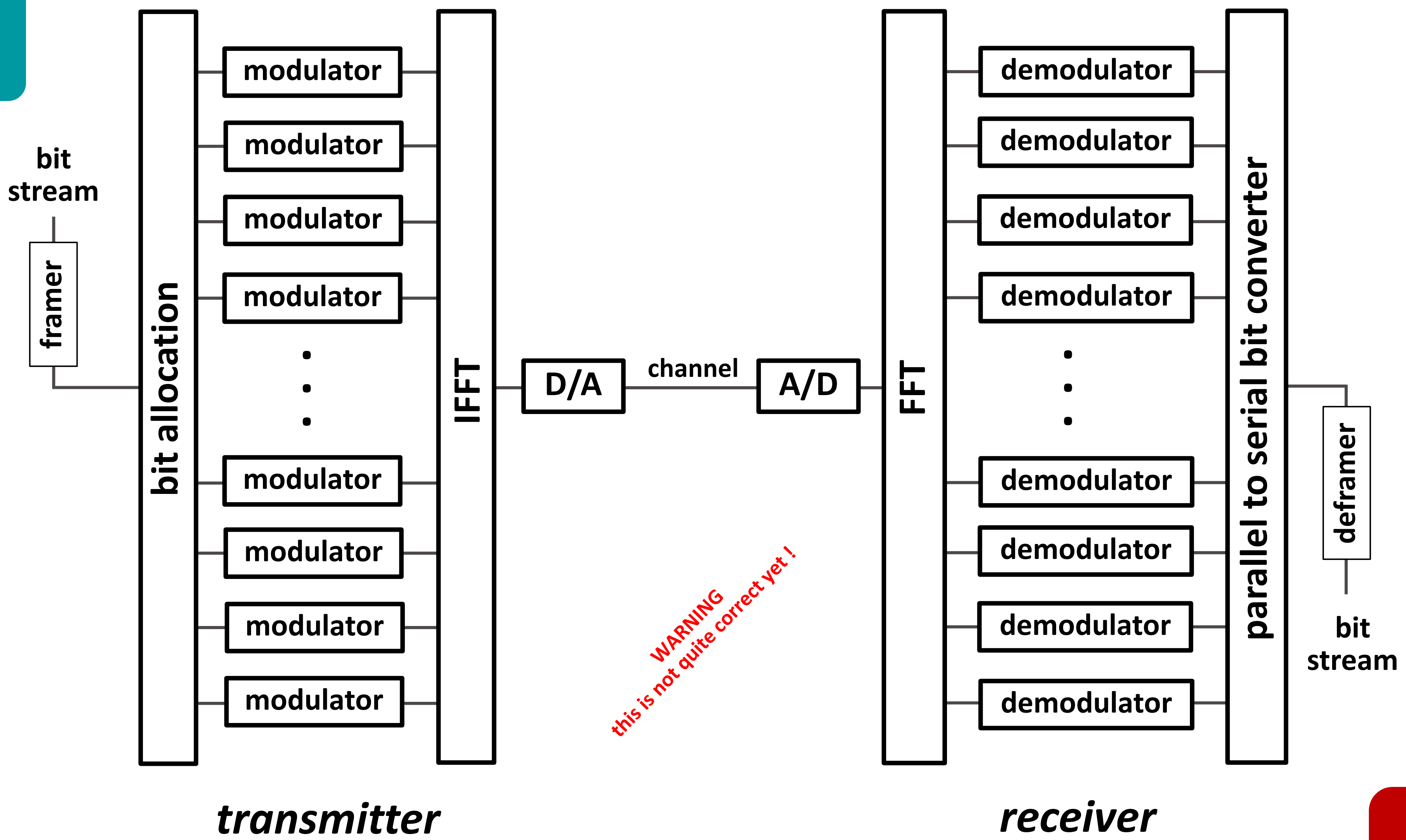


The upmixing can be performed in parallel for all sub-carriers

by the inverse Fourier transform (Zimmerman and Kirsch 1967)

and the FFT does it quickly (Weinstein and Ebert 1969)

OFDM modem paradigm



Cyclic prefix

What we previously derived *almost* works

For **analog** **digital** signal processing

linear **cyclic** convolution in the time domain **y = h * x**

is equivalent to multiplication in the frequency domain

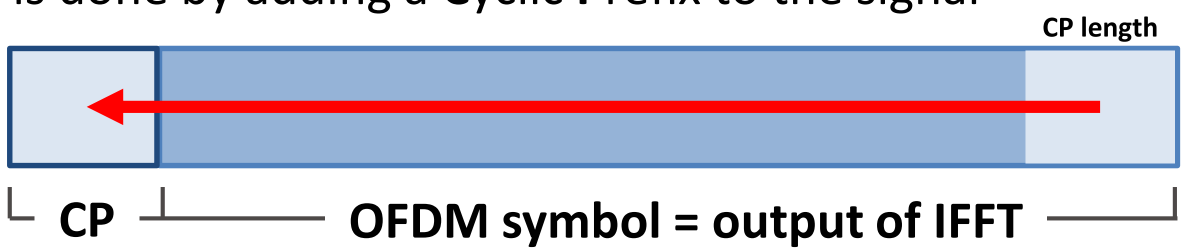
$$Y(\omega) = H(\omega)\, X(\omega)$$

$$Y_k = H_k\, X_k$$

So, the linear convolution in the analog channel

has to be converted into cyclic convolution for the digital channel

This is done by adding a **C**yclic **P**refix to the signal



which basically acts as a *guard interval* to eliminate ISI

The CO duration tends to be from 1/32 up to 1/4 of the symbol duration

and has to be long enough to include the maximum ISI spread

For example, if ISI is due to multipath

then the CP must be long enough to incorporate the longest delayed path

CP at work

OFDM splits the signal in the time domain into *frames*

In order for the signal processing of each frame to be independent
the CP ensures that delay spread is contained within the received frame

As a simple example, assume a frame of 8 samples
a 2-sample multipath $y_n = x_n + h_1 x_{n-1} + h_2 x_{n-2}$ and a CP of 2 samples

The OFDM frame before CP insertion is $x_0\, x_1\, x_2\, x_3\, x_4\, x_5\, x_6\, x_7$

After CP insertion the transmitted OFDM frame is $x_6\, x_7\, x_0\, x_1\, x_2\, x_3\, x_4\, x_5\, x_6\, x_7$

The received OFDM signal is $y_6\, y_7\, y_0\, y_1\, y_2\, y_3\, y_4\, y_5\, y_6\, y_7$

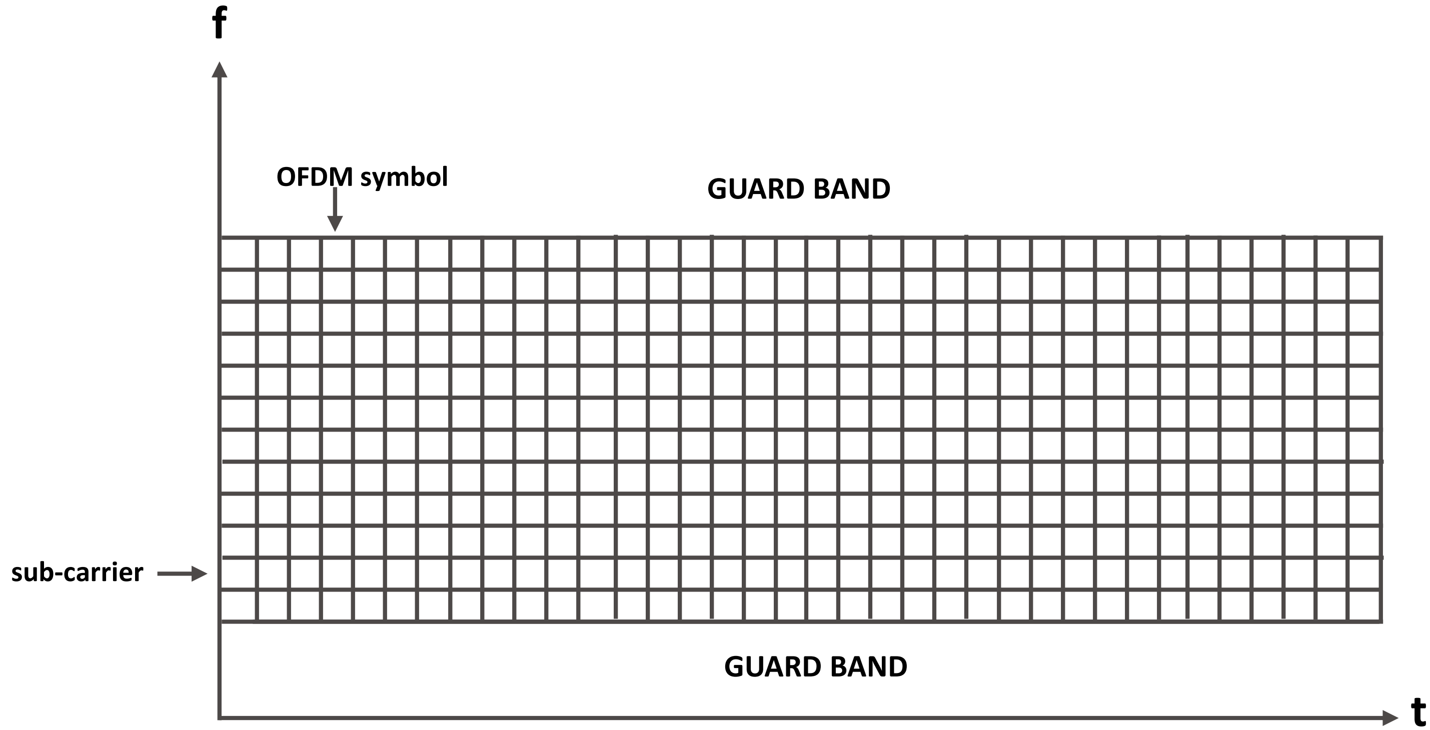
where y_6 and y_7 contain ISI from the previous frame and are *discarded*

and

$$\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & h_2 & h_1 \\ h_1 & 1 & 0 & 0 & 0 & 0 & 0 & h_2 \\ h_2 & h_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & h_2 & h_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & h_2 & h_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & h_2 & h_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & h_2 & h_1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & h_2 & h_1 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}$$

All the information is present for equalization to recover the original $x_0 \cdots x_7$
and the matrix is *circulant* – which is solved in the frequency domain!

OFDM signal structure



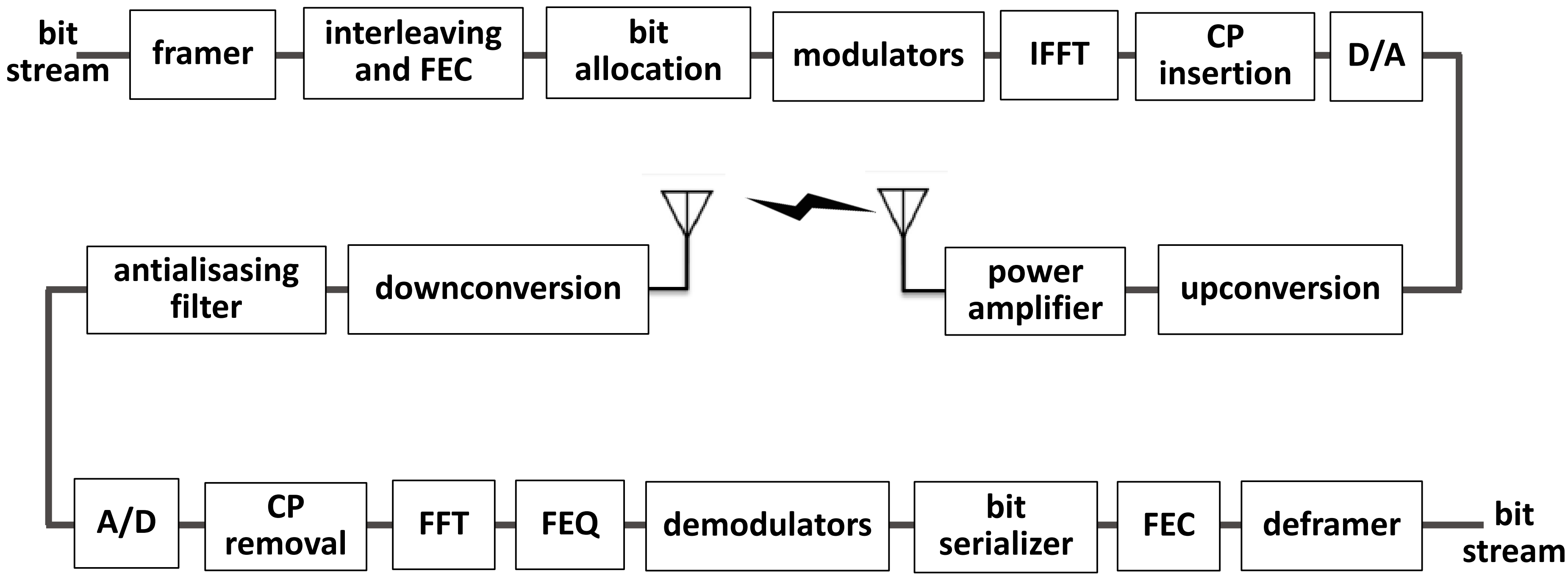
For LTE:

- **channel bandwidth ..., 5, 10, 15, or 20 MHz**
- **guard band overhead is 10%**
- **sub-carrier spacing = 15 kHz**
- **OFDM symbol duration = 1/15KHz = 66.67 μsec**
  - **short CP = 4.7 μsec so total duration = 71.367 μsec**
    - **1 *slot* = 7 symbols ≈ ½ msec***
  - **long CP = 16.7 μsec so total duration = 83.367 μsec**
    - **1 *slot* = 6 symbols ≈ ½ msec***

BW (MHz)	usable BW (MHz)	subchannels	FFT
5	4.5	300	512
10	9	600	1024
15	13.5	900	1536
20	18	1200	2048

* CP durations are *adjusted* so that the slot is precisely ½ msec

OFDM modem


bit stream
framer
interleaving and FEC
bit allocation
modulators
IFFT
CP insertion
D/A
antialisasing filter
downconversion
power amplifier
upconversion
A/D
CP removal
FFT
FEQ
demodulators
bit serializer
FEC
deframer
bit stream

PAR

The **P**eak to **A**verage **R**atio (AKA crest factor) of a signal
- is the ratio between its maximum peak-to-peak to its root mean square

For example, for a single the p2p is 2 while the rms is √2, so PAR= √2

When adding N independent sine values (as is done in OFDM)
- the PAR is multiplied by N (the Gaussian tail)

This creates a major problem for OFDM transmitter's power amplifier

Power amplifiers are only linear in a limited operating range
- if the input signal value is too large, the output saturates
  - leading to distortion that breaks orthogonality between sub-carriers

This is especially problematic for the cellular uplink
- where the UE amplifier can not be sophisticated/expensive

Many PAR reduction techniques have been developed
- but it may be better to avoid using multi-carrier methods in the UL

SC-FDM

To drastically reduce the PAR we can use **S**ingle **C**arrier FDM
sometimes called linearly precoded OFDMA (LP-OFDMA)

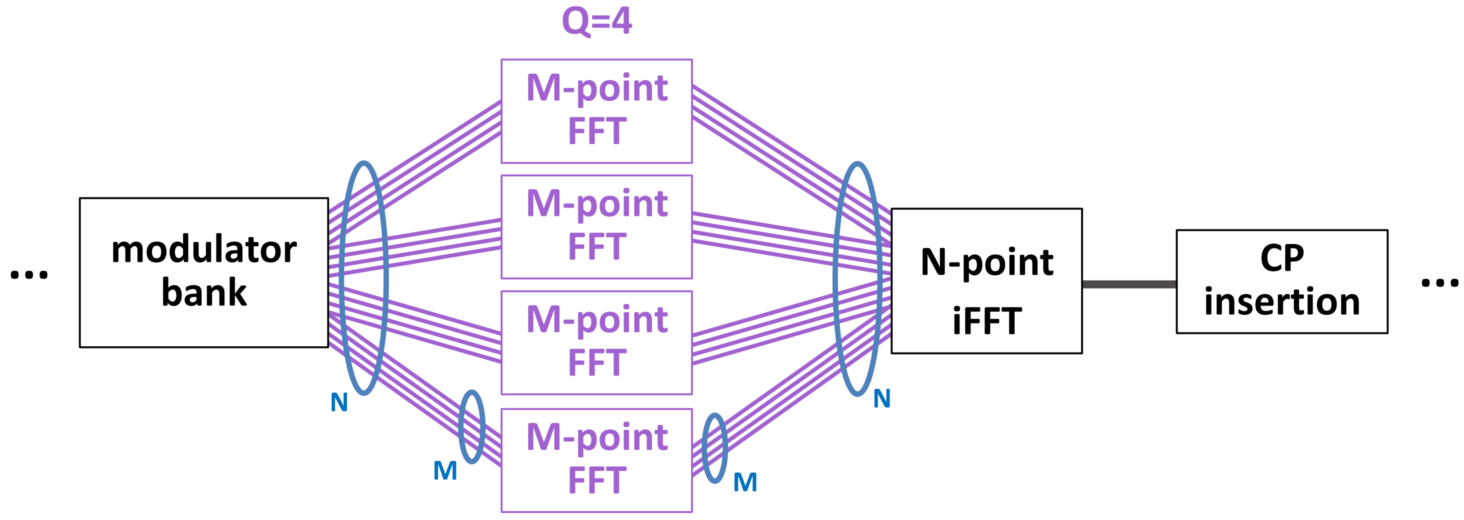
To transmit SC-FDM instead of OFDM
instead of inputting N QAM symbols into the iFFT
we group them into blocks of M (M = N/Q)
and perform a M-point FFT on each of these before the N-point iFFT

This results in Q orthogonal sub-carriers
if Q=1 (M=N) FFT and iFFT cancel out, resulting in a single carrier
if Q>1 the PAR is Q times that of a single carrier signal

So, why if we want a single carrier – why use OFDM at all?

- we can still perform the equalization in the frequency domain
- we can still flexibly allocate channels

SC-FDM transmitter detail



SC-FDM only