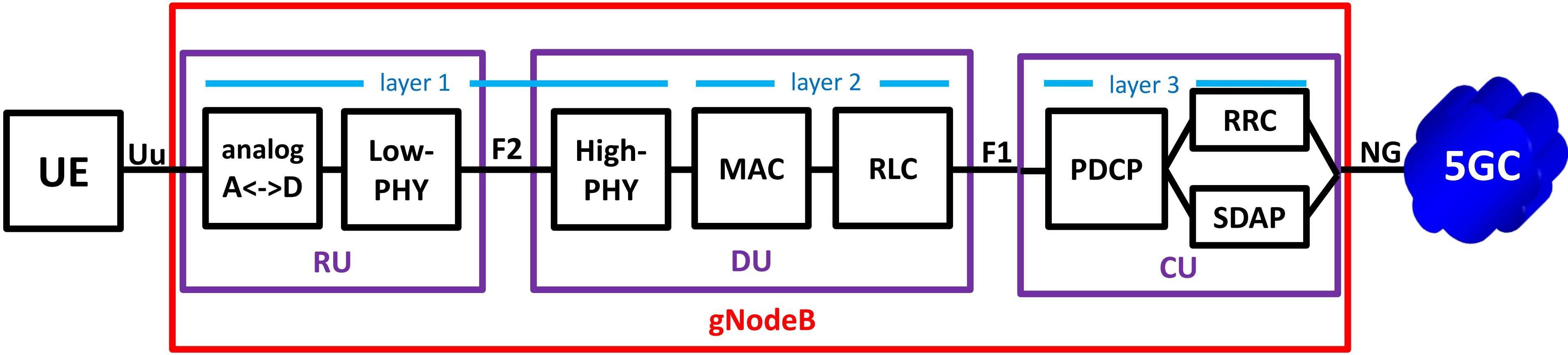
Air Interface – Part III
more NR-specific details

5G NR air interface architecture

We saw that the gNodeB may be decomposed into 3 parts (RU, DU, and CU)
We can now fully understand the decomposition



PHYsical layer
Media **A**ccess **C**ontrol
Radio **L**ink **C**ontrol
Packet **D**ata **C**onvergence **P**rotocol
Radio **R**esource **C**ontrol
Service **D**ata **A**daptation **P**rotocol

Control User Plane Separation

User plane – control plane separation is important for independent scaling

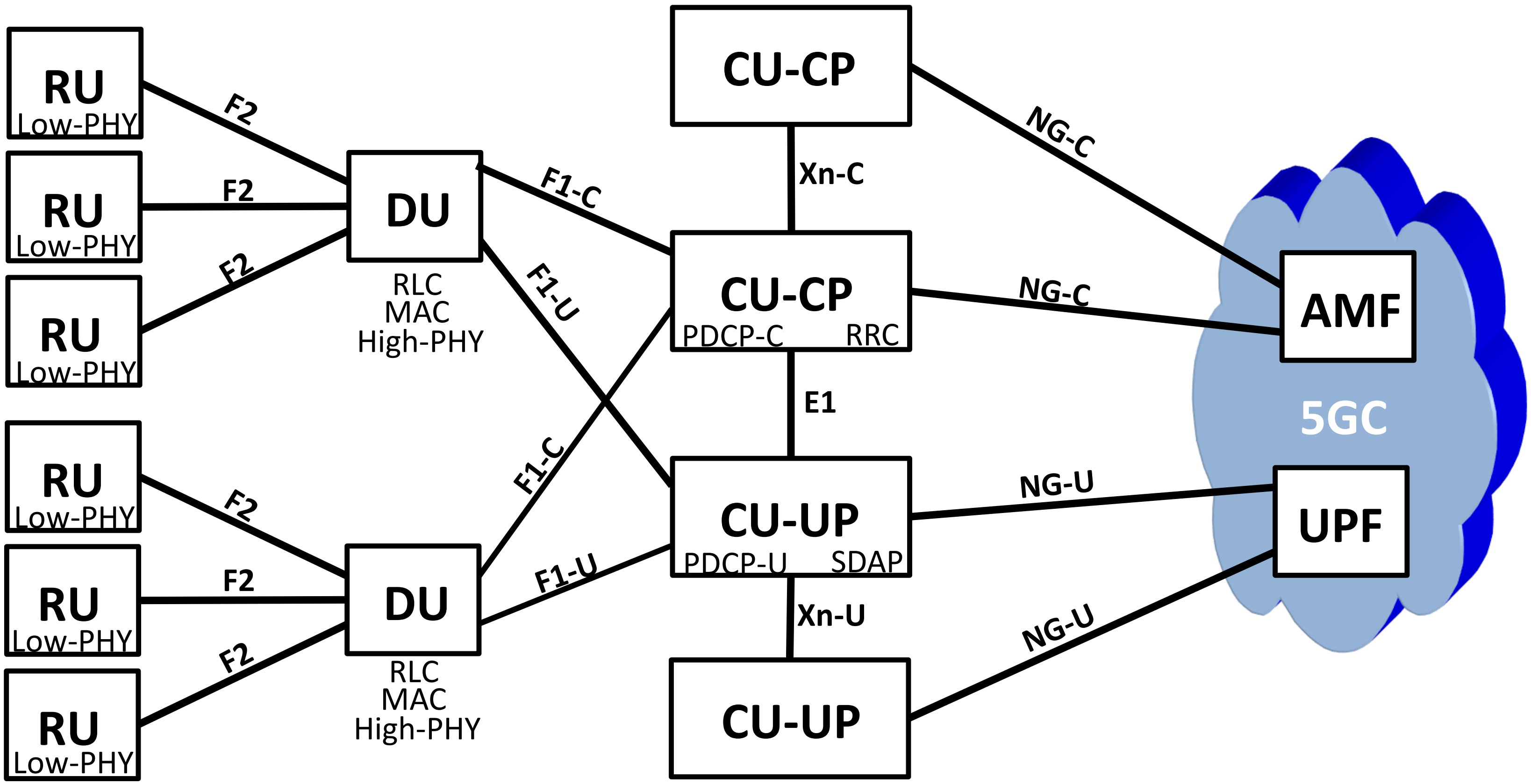
Example:
The 2009 iPhone AT&T meltdown was caused
by CP servers not keeping up with iPhone frequent state transitions
while there *was* sufficient UP capacity

4G Release 14 introduced EPC **C**ontrol and **U**ser **P**lane **S**eparation, in order to

- enable adding user plane NEs w/o adding control servers or vice versa
- independently locate UP and CP core nodes
- independent evolve UP and CP functions
- enable **S**oftware **D**efined **N**etworking

5G continues and further extends this direction

5G RAN control/user plane separation



5G NR

5G's new air interface is based on LTE's OFDMA
but features several improvements

- new RF bands (including mmWaves)
- wider bandwidth (up to 400 MHz) and larger FFT sizes (up to 4k)
- multiple subcarrier spacings (15 kHz ... 480 kHz)
- scalable numerology
- higher order modulations (up to 256 QAM) and so higher spectral efficiencies
- new error correcting codes
- more flexible frame formats, including self-contained integrated subframe
- carrier aggregation
- massive MIMO (64+ antennas) and beam-forming

Spectrum

5G needs bandwidth, and can use anything it can get

RF bands are divided into :

- low (< 1 GHz) and medium frequencies (<6 GHz)
  for macro cells
- mm waves (> 6GHz) where the range is expected to be < 500 meters
  so will be used for small cells (LoS 300m, urban multipath 150m)

Operating bands being considered:

	low (MHz)	medium (GHz)	high (GHz)
US	600	3.5	28, 37-40, 64-71
EU	700	3.4-3.8	24.25-27.5
China		3.4-3.6	24.25-27.5
Japan		3.6-4.2, 4.4-4.9	27.25-29.5
South Korea			28, 37.5-40

Each band has a standardized name n# (LTE used just #)

when the bands are the same, the numerical value is the same

R15 operating bands

FR1 (< 6GHz)

NR operating band	Uplink (UL) *operating band* BS receive / UE transmit F_{UL_low} – F_{UL_high}	Downlink (DL) *operating band* BS transmit / UE receive F_{DL_low} – F_{DL_high}	Duplex Mode
n1	1920 MHz – 1980 MHz	2110 MHz – 2170 MHz	FDD
n2	1850 MHz – 1910 MHz	1930 MHz – 1990 MHz	FDD
n3	1710 MHz – 1785 MHz	1805 MHz – 1880 MHz	FDD
n5	824 MHz – 849 MHz	869 MHz – 894 MHz	FDD
n7	2500 MHz – 2570 MHz	2620 MHz – 2690 MHz	FDD
n8	880 MHz – 915 MHz	925 MHz – 960 MHz	FDD
n20	832 MHz – 862 MHz	791 MHz – 821 MHz	FDD
n28	703 MHz – 748 MHz	758 MHz – 803 MHz	FDD
n38	2570 MHz – 2620 MHz	2570 MHz – 2620 MHz	TDD
n41	2496 MHz – 2690 MHz	2496 MHz – 2690 MHz	TDD
n50	1432 MHz – 1517 MHz	1432 MHz – 1517 MHz	TDD
n51	1427 MHz – 1432 MHz	1427 MHz – 1432 MHz	TDD
n66	1710 MHz – 1780 MHz	2110 MHz – 2200 MHz	FDD
n70	1695 MHz – 1710 MHz	1995 MHz – 2020 MHz	FDD
n71	663 MHz – 698 MHz	617 MHz – 652 MHz	FDD
n74	1427 MHz – 1470 MHz	1475 MHz – 1518 MHz	FDD
n75	N/A	1432 MHz – 1517 MHz	SDL
n76	N/A	1427 MHz – 1432 MHz	SDL
n77	3300 MHz – 4200 MHz	3300 MHz – 4200 MHz	TDD
n78	3300 MHz – 3800 MHz	3300 MHz – 3800 MHz	TDD
n79	4400 MHz – 5000 MHz	4400 MHz – 5000 MHz	TDD
n80	1710 MHz – 1785 MHz	N/A	SUL
n81	880 MHz – 915 MHz	N/A	SUL
n82	832 MHz – 862 MHz	N/A	SUL
n83	703 MHz – 748 MHz	N/A	SUL
n84	1920 MHz – 1980 MHz	N/A	SUL

FR2 (>6GHz)

NR Operating Band	Uplink (UL) operating band BS receive UE transmit F_{UL_low} – F_{UL_high}	Downlink (DL) operating band BS transmit UE receive F_{DL_low} – F_{DL_high}	Duplex Mode
n257	26500 MHz – 29500 MHz	26500 MHz – 29500 MHz	TDD
n258	24250 MHz – 27500 MHz	24250 MHz – 27500 MHz	TDD
n260	37000 MHz – 40000 MHz	37000 MHz – 40000 MHz	TDD

mmWave challenges (1)

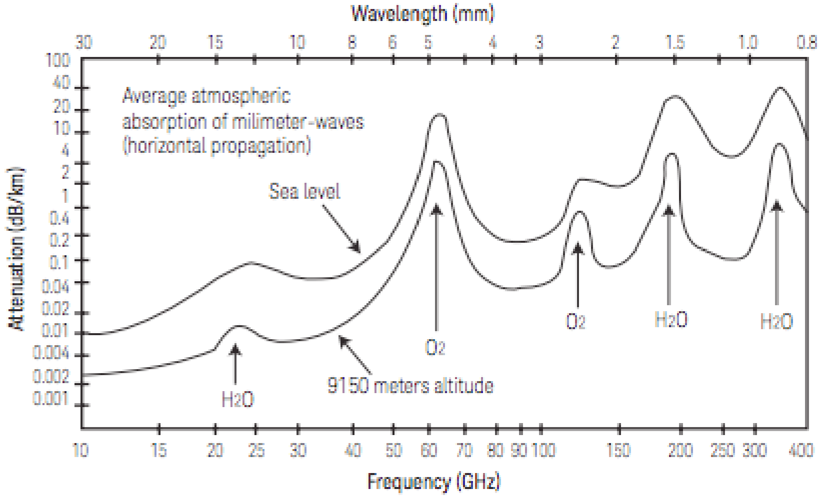
30 GHz corresponds to 1 mm wavelength

so the 24, 27, 28 GHz bands are informally called mm waves

At these high RF frequencies there is plenty of real estate for 5G

but there are multiple challenges to exploiting these frequencies!

1. signal attenuation
   - free space signal attenuation scales like f^2 due to antenna aperture size
     so increasing from 3 GHz to 30 GHz increases loss by 20 dB = factor of 100 !
   - dry oxygen, water vapor, etc. have absorption lines



SOLUTION: much smaller cell sizes – 150-200 m

mmWave challenges (2)

2. concrete and glass strongly absorb mm waves - so need clean line of sight
   but 80% of expected usage is indoors
   SOLUTION: indoor coverage

3. much shorter channel coherence time (time over which can assume channel stationarity)
   for mobile user the coherence time scales like 1/f
   increasing from 3 GHz to 30 GHz decreases the coherence time by a factor of 10!
   SOLUTON: much more frequent channel probing

4. energy efficiency (and carbon footprint)
   scaling BW by 10 usually equates to 10 times more power
   and scaling data rate by 10-100 produces the same scaling in the backhaul
   SOLUTION: more directive antennas, energy-efficient system design, sleep modes

5. implementation issues
   – power amplifiers are very inefficient at mmWaves (< 10% is common)
   – need low loss/low dielectric constant materials for multi-layer PCBs
   – copper traces have high resistance due to skin effect
   – need temperature stable high-Q dielectrics
     • various ceramics and polymers being investigated
   – need new magnetic materials for high-frequency oscillators
     • Yttrbium Iron garnet, Ni-Zn ferrites, Li-based spinels being investigated

mmWave (3) health issues

6. health issues
   - there is no data on the long-term safety of continuous mmwave exposure
   - direct effect will be on external 2 mm of skin or eyes
     - 30-40% of power will be reflected
   - thermal effects known for radiation above 5 mW/cm^2
     so FCC guidelines limit public to 1 mW/cm^2 over 30 minutes fro 1.5-100GHz
   - non-thermal effects reported for mmWaves:
     - long term exposure may be carcinogenic
     - altering structural and functional properties of plasma membranes
       by modifying ion channel activity or modifying the phospholipid bilayer
     - skin nerve ending damage
     - activation of the immune system and peripheral neural system
     - changes to blood and bone marrow composition
     - changes to activity of various enzymes
     - modification of microbes (growth depression and property alteration)
       may lead to various effects including antibiotic resistance,

SOLUTION: none yet

Slots

In NR (like LTE) the frame is always 10 ms = 10 subframes of 1 ms

For normal CP there are 14 symbols per subframe (for SCS < 240 kHz)
and this is defined as the basic slot (not 7 symbols = ½ ms like in LTE!)
(also called the **T**ransmission **T**ime **I**nterval)

So, the slot duration = 1 ms / 2^{μ}

- for SCS=15 kHz slot = 1 ms 1 slot per subframe
- for SCS=30 kHz slot = ½ ms 2 slots per subframe
- for SCS=60 kHz slot = ¼ ms 4 slots per subframe etc.

We saw that the slots can *nest* across numerologies

However, NR is even more flexible, allowing:

- slot aggregation
  transmission occupies 2 or more slots, in order to reduce overhead
- allocating mini-slots (non-slot-based scheduling)
  transmission occupies less than a slot (2, 4, or 7 OFDM symbols)
  in order to reduce latency
- flexible (mixed DL/UL) slots for TDD operation

Cyclic Prefix

NR like LTE has regular and long CP (long CP is only defined for SCS = 60 kHz)
but the CP duration depends on numerology

The CP overhead stays about the same (about 7%)

For all symbols except the last in each slot the duration is $144 / 2^{\mu} T_s$
to the last symbol in the slot we add an additional $1024\ T_s$
so that it works out exactly 500 μs.

μ	SCS (kHz)	CP (μs)	CP (μs) for last symbol	Nsymbols
0	15	4.69	5.2	7
1	30	2.34	2.86	14
2	60	1.17	1.69	28
2	60	4.17	4.17	28
3	120	0.59	1.11	56
4	240	0.29	0.81	112

Flexible slots

A TDD slot can contain

- all DL symbols
- all UL symbols
- both DL and UL symbols, with 1 or 2 switching points per slot

The type is indicated by the Slot Format Indication (configured by PDCCH or RRC)

The *self-contained integrated* subframe (2 switching points)

always starts with a DL control burst and end with a UL control burst

- DL-centric subframe

  DL control, DL data (single user), guard, UL control

- UL-centric subframe

  DL control, guard, UL data (to multiple users), UL control

This enables (e.g., for DL-centric subframes)

- lower latency, since the UL HARQ ACK is in the same subframe (self-contained)

  note: LTE assumes HARQ processing time of 3 ms, NR requires DL < 1 ms and UL < 0.3-0.8 ms
- massive MIMO tracking, since the UL link quality is in the same subframe
- use of shared spectrum, via Listen-Before-Talk indications from network

Some more technical details

NR supports 1008 cell identities (twice as many as LTE!) (N_1=0...335)

PSS, SSS, frame number, PBCH/MIB, PDCCH, and PDSCH are similar to LTE
although details may vary, especially for > 6 GHz

In particular, there are added indications for SCS and the other new features

PDSCH and PUSCH support QPSK, 16-QAM, 64-QAM, and 256-QAM
(PBCH and PDCCH use QPSK, PUCCH uses BPSK or QPSK)

Instead of LTE's *turbo* and *tail-biting convolutional* codes, NR uses:

- *LDPC* codes for PDSCH and PUSCH
- *polar* codes for PDCCH and PBCH (up to 80 bit blocks)
- *polar*, *repetition*, *simplex*, and/or *Reed-Muller* codes
  for PUCCH and for control information in the PUSCH

LDPC was chosen for

- flexibility under block length scaling
- efficiency gain over LTE turbo codes
- parallelization for lower complexity and latency

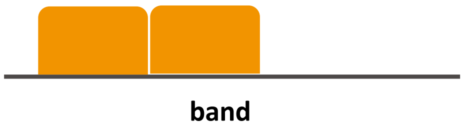
and there are many many more minor technical advances ...

Carrier aggregation

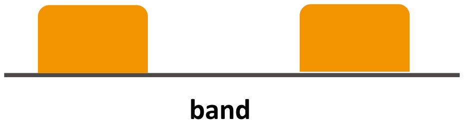
CA was introduced in LTE-A to achieve higher data rates
by assigning multiple subcarriers to the same user

We differentiate between:

- intra-band contiguous aggregation
  (RB aggregation in PHY)



- intra-band non-contiguous aggregation
  (integration in MAC)



- inter-band aggregation
  (integration in MAC or L3)

While LTE-A allowed up to 5 **C**omponent **Carriers** up to 20 MHz bandwidth
5G currently allows up to 16 CCs with up to 1 GHz aggregated bandwidth
including **D**ual **C**onnectivity with 4G and 5G CCs (**E**-UTRA-**N**R-**DC**)

MIMO

While LTE supports 4*2 MIMO (4 BS and 2 UE antennas)
R15 supports 32*4 MIMO (and this will increase significantly!)

MIMO is critical for both < 6 GHz and > 6 GHz operating bands
but for different reasons

- for < 6 GHz
  - cells are large
  - there will be rich multipath
  - many users in cell
  - users will be highly mobile
  - so MIMO will use spatial mux to help achieve spectral efficiency goals
- for > 6 GHz
  - signal attenuation much higher (100 times higher?)
  - cells are small and little multipath
  - few users in cell
  - users are relatively static in the cell
  - so MIMO will use beamforming to overcome the high attenuation

Low frequency MIMO

For low frequencies MIMO exploits multipath

each receive antenna receives linear combination of channels

With MIMO, the Shannon capacity $C = N\ BW\ \log_2(SNR + 1)$

where N is the number of MIMO channels

MIMO works by increasing N

which is much more effective than increasing SNR!

Standard **S**ingle **U**ser MIMO requires the UE to invert the channel matrix

in order to recover the N independent data streams

which leads to substantial battery drain

An alternative is for the base station (which is less power constrained)

to perform channel precoding (pre-filter with the inverse channel matrix)

so that each UE antenna receives a clean channel

With **M**ulti **U**ser MIMO the BS always uses channel precoding

since each UE only handles some of the channels

But this solution hits power constraints (in order to guarantee high SNR for each UE)

Beamforming

The basic idea can be understood by considering (see figure)

2 transmitting antennas emitting the same sinusoid

spaced ½ wavelength apart

If there is no phase shift to either antenna

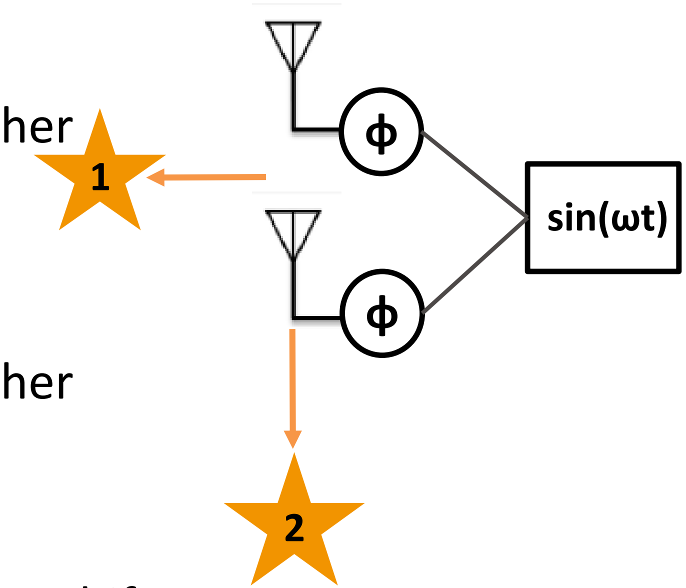
at point 1 the signals will re-enforce each other

at point 2 the signals will cancel out

If we apply a phase shift of π to one antenna

at point 1 the signals will cancel out

at point 2 the signals will re-enforce each other



More generally, we can achieve a focused beam

by using a linear array of antennas

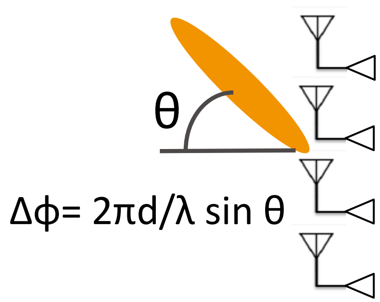
with amplifiers (to reduce side-lobes) and phase shifters

The structure is analogous to MA filter

but in the spatial axis rather than frequency

And similarly, an FFT structure (called the *Butler matrix*)

can be used to efficiently steer the array



MIMO steering mechanisms

For simple free-space geometries

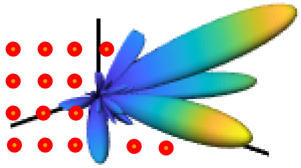
one can directly manipulate relative antenna phases

thus directing the MIMO beam in a certain direction or directions

(see beam patterns on the coming pages)

This technique can be extended to a 2-dimensional array

in order to beamform in three dimensions



Geometric beamforming is simple enough

to be performed in the analog domain

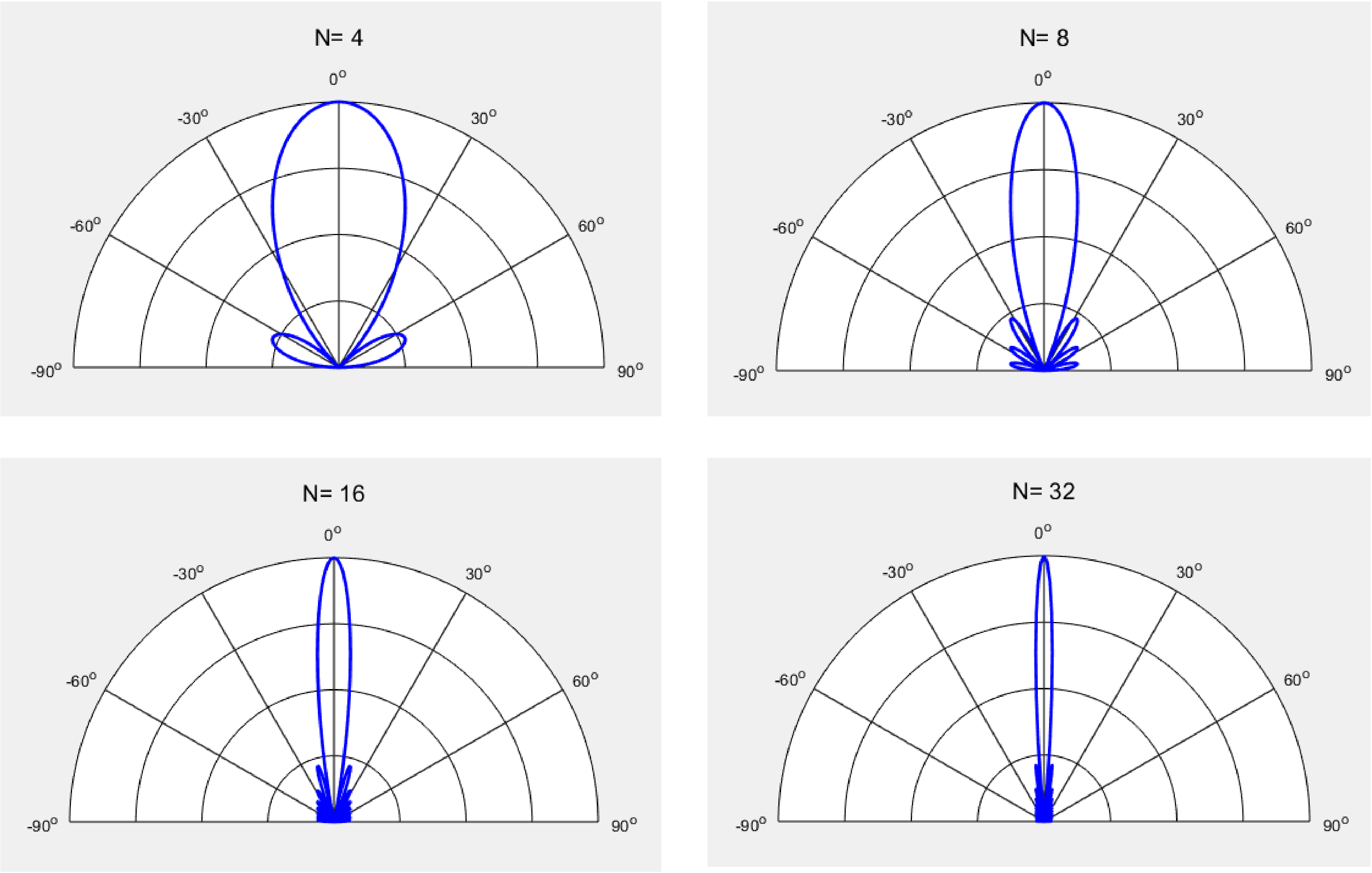
the antenna signals are combined to reduce the number of channels

Thus the F2 interface we needn't contain channels for every antenna

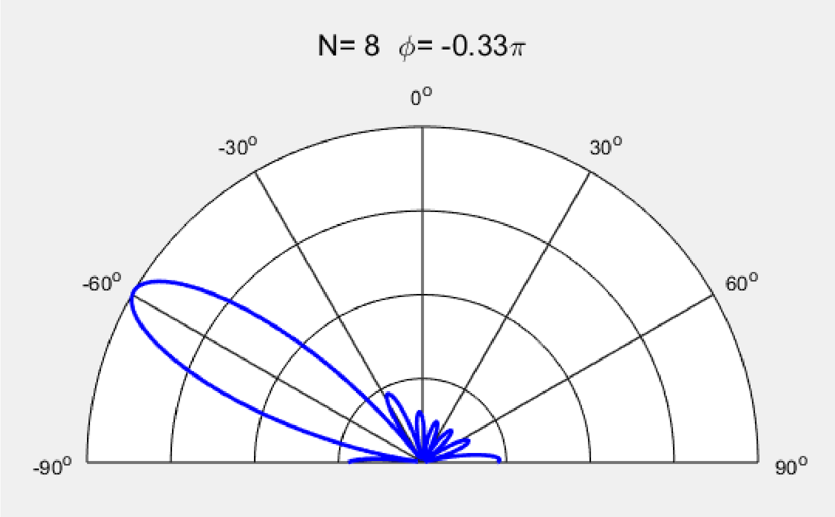
For complex scattering geometries

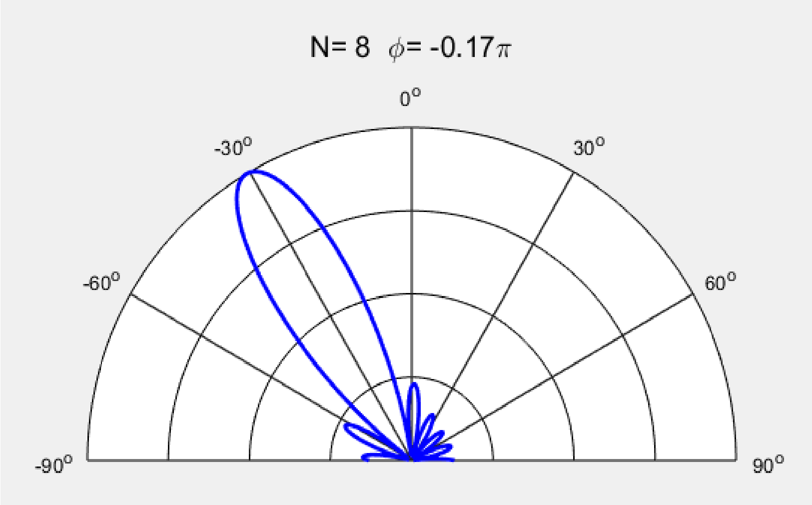
we need to use (analog or digital) *precoding* (to be discussed later)

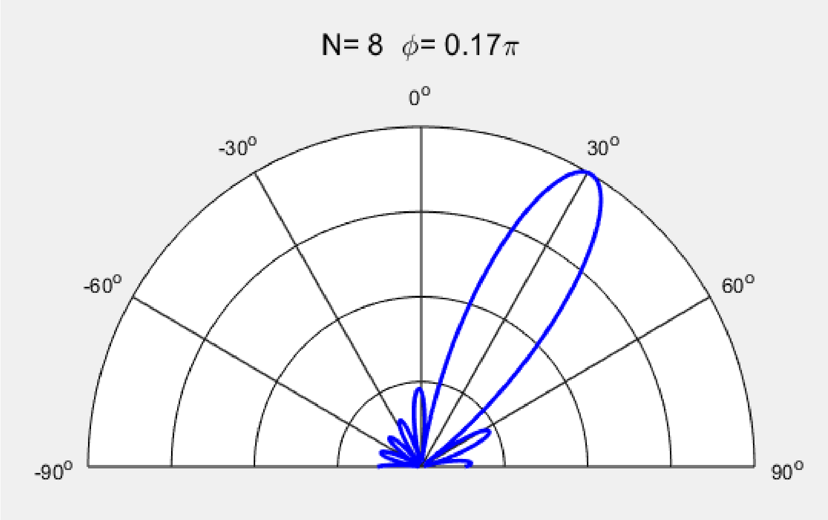
Geometric beamforming - beamwidth

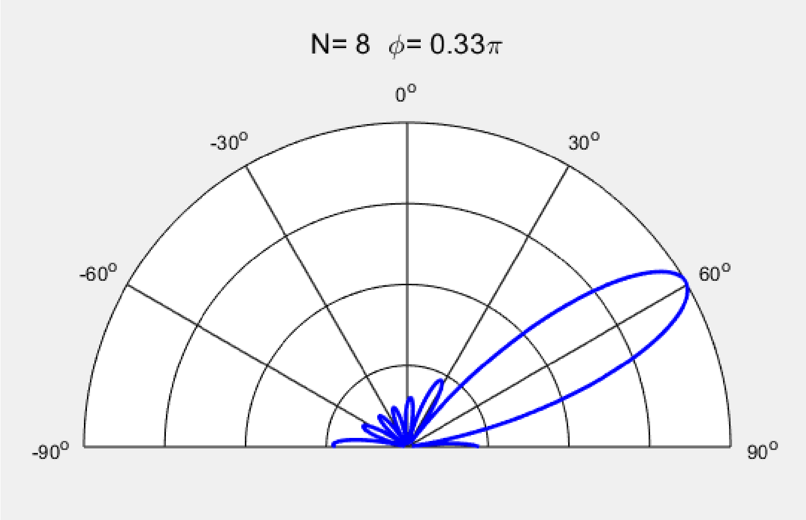


Geometric beamforming - directivity



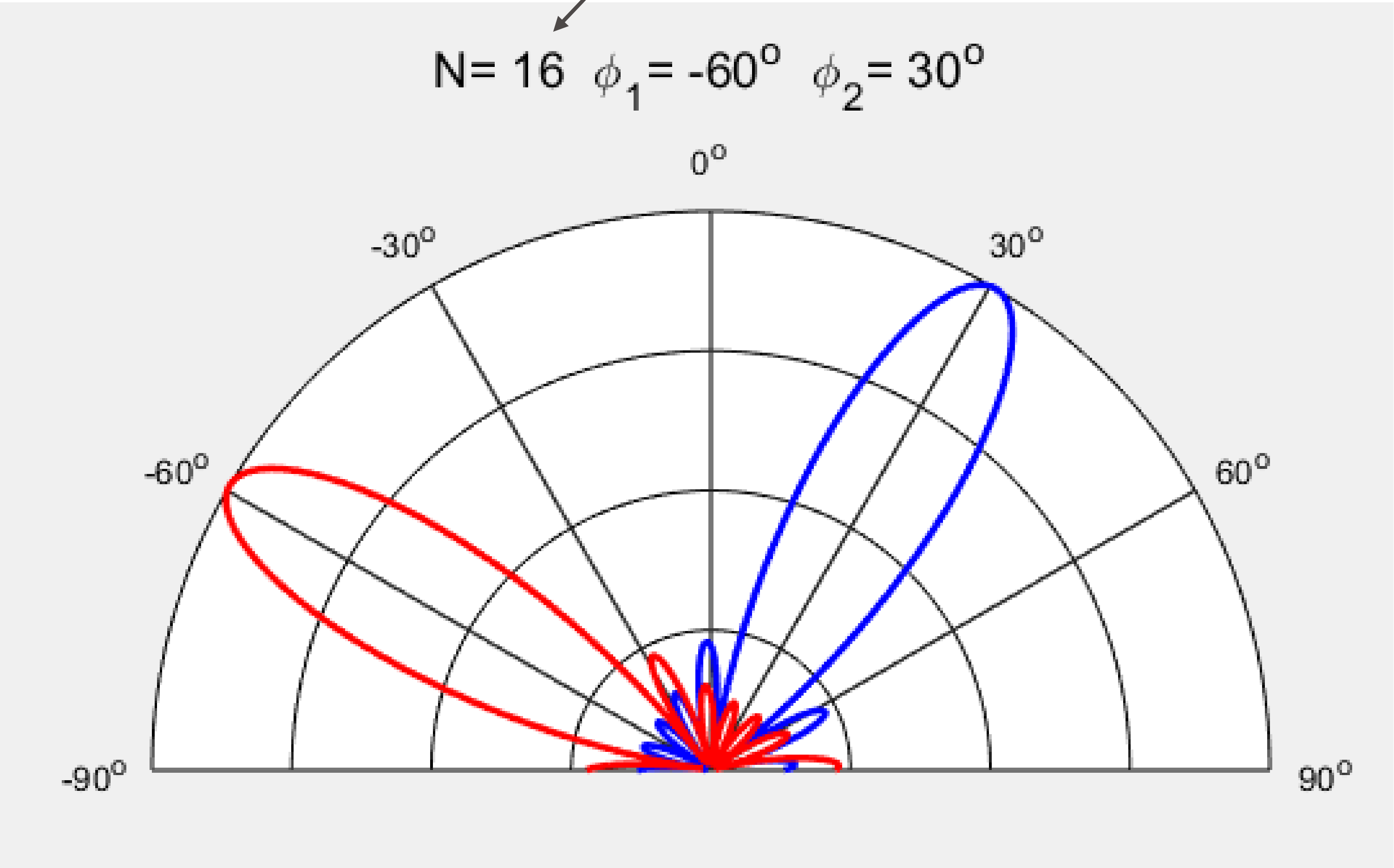






MU-MIMO

8 channels for each beam



Massive MIMO

Spatial diversity MIMO increases Shannon capacity (to 1 or multiple users)

$C = N_{MIMO}\ BW \log_2(SNR + 1)$

by effectively utilizing additional channels N_{MIMO}

Beamforming MIMO increases Shannon capacity by increasing received SNR

since with a highly directional beam

transmitted power is focused in the intended direction

These two types of MIMO actually opposites

- spatial diversity sends energy in many directions to get *different* channels
- beamforming focuses energy in one direction to improve SNR

Massive MIMO combines both mechanisms beamforming and MU MIMO

by using many more antennas than the number of UEs

we can focus energy on each UE

In practice, digital beamforming is first employed for coarse steering

followed by analog beamforming for fine tuning

Massive MIMO allows each UE to have different power and data rates

Precoding

For non-free-space geometries we have many multipath reflections
and simple geometric beamforming doesn't work

We can model the multipath channel by a multidimensional filter
the effects of which can be compensated by a *precoding* operation
(this is similar to *vectoring* used in DSL)

Let's start by considering a very small cell
in which we assume all paths have the same delay

Given N transmit antennas transmitting RF signals x_n
and K receive antennas receiving RF signals y_k
we can model $\underline{y} = \underline{\underline{h}}\ \underline{x} + \underline{n}$ where

- $\underline{n}$ are k noise signals (which we will neglect here)
- h is a k*n channel cross-talk matrix

In order to recover the k desired signals s_k (to within the noise) we precode
$\underline{x} = \underline{\underline{h}}^{-1}\ \underline{s}$ so that $\underline{y} = \underline{\underline{h}}\ \underline{x} = \underline{\underline{h}}\ \underline{\underline{h}}^{-1}\ \underline{s} = \underline{s}$ recovering the desired signal

Note that for TDD with channel reciprocity the transmitter has direct access to $\underline{\underline{h}}$

Even more precoding

Reflections are frequency dependent and thus so is $\underline{\underline{h}}$

so the above calculation needs to be performed for each frequency

For mobile UEs the geometry changes with time (coherence time)

so the above calculation needs to be continuously updated

We unrealistically assumed that all paths have the delay

In a more realistic model the matrix $\underline{\underline{h}}$ needs to perform convolutions

$\underline{y} = \underline{\underline{h}} * \underline{x} + \underline{n}$ (i.e., each matrix element is a set of filter coefficients)

Instead of considering a 3-dimensional tensor, we can remain with matrices

Given a maximum delay T we define

$x_n = (x_{10}, x_{11}, \dots x_{1T}; x_{20}, x_{21}, \dots x_{2T}; \dots ; x_{N0}, x_{N1}, \dots x_{NT})$

$y_k = (y_{10}, y_{11}, \dots y_{1T}; y_{20}, y_{21}, \dots y_{2T}; \dots ; y_{K0}, y_{K1}, \dots y_{KT})$

and we can model (neglecting noise) $\underline{y} = \underline{\underline{h}}\ \underline{x}$ where h is a kT*nT matrix

Once again we can precode with the inverse matrix

but the inversion requires much more computational power

In practice $\underline{\underline{h}}$ may be relatively sparse and various approximations used

Yet even more beamforming

Until now we have assumed that although the UE may be mobile
the scatterers are stationary
If this is not the case Doppler shifting causes mixing of different frequencies
and we need a 4d tensor or an even larger composite matrix

For beamforming to work in practice, various mechanisms are needed

- beam quality measurement and reporting
- beam assignment
- recovery if quality is low
- mobility tracking

Quality measurement is based on accurate **C**hannel **S**tatus **I**nformation

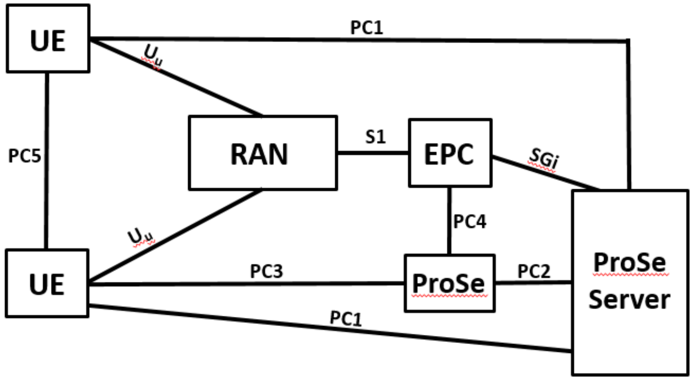
- for UL the **S**ounding **R**eference **S**ignal is used
- for DL TDD (with channel reciprocity!) the UL signals themselves can be used
- for DL FDD a new 2-stage high resolution spatial precoding is used
  to approximate the channel eigenvector

PC5 interface

4G R12 introduced a device-to-device (D2D) air interface
called PC5, *sidelink*, **Pro**ximity **Se**rvices (ProSe)
with which one UE can communicate directly with a second UE
without going through a base-station

PC5 was introduced in LTE for cases such as
first responders and **P**ush-**T**o-**T**alk (walky-talky)

The R12 architecture allows direct and EPC-aided discovery



5G will extend the PC5 interface for the V2V use case