HEXA-X-II

A holistic flagship towards the 6G network platform and system, to inspire digital transformation, for the world to act together in meeting needs in society and ecosystems with novel 6G services

Deliverable D5.3
Initial design and validation of technologies and architecture of 6G devices and infrastructure

Co-funded by the European Union

6GSNS

Hexa-X-II project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101095759.

Date of delivery:	29/02/2024	Version:	1.0
Project reference:	101095759	Call:	HORIZON-JU-SNS-2022
Start date of project:	01/01/2023	Duration:	30 months

Document properties:

Document Number:	D5.3
Document Title:	Initial design and validation of technologies and architecture of 6G devices and infrastructure
Editor(s):	Claude Desset (IMEC)
Authors:	IMEC: Claude Desset, Meng Li, Jeroen Famaey, Ritesh Kumar Singh, Adnan Sabovic, Priyesh Pappinisseri Puluckul, LMF: Bikramjit Singh, Hamza Khan, SEQ: Efstathios Katranaras, NDE: George Hotopan, WIN: Pavlos Alexias, Xygkos Christos, ORA: Philippe Ratajczak, Dinh-Thuy Phan Huy, SAT: Lukas W. Mayer, Martin Schiefer, OUL: Onel L. A. López, David E. Ruiz-Guirola, Amirhossein Azarbahram, Osmel Martínez Rosabal, Prasoon Raghuwanshi, Rafael Valente da Silva, Nuutti Tervo, Dedar Rashid, EAB: Muris Sarajlic, Rickard Ljung, AAU: Riku Jäntti, Liao Jingyi, Azzan Al-Nahari, QLC: Nicolas Cornillet, SON: Nafiseh Mazloum, BI: Sebastian Haas, TUD: Emil Matus, Viktor Razilov, Simon Friedrich
Contractual Date of Delivery:	29/02/2024
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	Hexa-X-II_D5.3_v1.0

Revision History

Revision	Date	Issued by	Description
v0.1	28.09.2023	Hexa-X-II WP5	Document creation with initial ToC
v0.2	21.12.2023	“	First draft ready for review
v0.3	11.01.2024	"	Ready for external (X-WP) review
v0.4	01.02.2024	"	Ready for GA approval
v1.0	29.02.2024	"	Published version

[1] SEN = Sensitive, only members of the consortium (including the Commission Services). Limited under the conditions of the Grant Agreement

PU = Public

Abstract

Various technologies are expected to support 6G. The first design results are illustrated in four main directions. First, sub-THz architectures are considered. Dimensioning for energy efficiency is illustrated and hardware non-idealities are reviewed, with a special focus on phase noise. Alternative architectures based on resonant tunneling diodes or lens antennas are also explored. Second, reflective intelligent surfaces are considered. Radiation models are created, validated against prototypes, and the main features and control functions are identified. Third, system-on-chip architectures are explored. They combine DSP cores such as RISC-V and AI accelerators. Energy harvesting and power management aspects are also considered. Fourth, ultra-low cost and power IoT devices are investigated. They combine many technologies such as power-saving protocols, energy harvesting from multiple sources, tiny ML engines. Backscattering demonstration is progressing for zero-energy devices.

Keywords

sub-THz, energy efficiency, hardware non-idealities, phase noise, RIS, resonant tunneling diodes, lens antennas, SoC, RISC-V, AI accelerators, energy harvesting, ultra-low-power, IoT, power saving protocols, backscattering

Disclaimer

Funded by the European Union. The views and opinions expressed are however those of the author(s) only and do not necessarily reflect the views of Hexa-X-II Consortium nor those of the European Union or Horizon Europe SNS JU. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

This deliverable investigates various technologies expected to support 6G devices and infrastructure and presents the first related design results. Those technologies are clustered in four main directions.

First, transceiver architectures for sub-THz frequency bands are investigated on different aspects. Architecture dimensioning for maximal energy efficiency is illustrated for several hotspot scenarios. A proper choice of hybrid MIMO architecture and related parameters such as number of antennas and output power are essential. Hardware non-idealities are revisited for those high frequency bands. Key extensions needed concern PA models with memory, frequency-selective models for I/Q imbalance and per-antenna decorrelated non-idealities. Another key problem is phase noise. Its limitations are investigated more specifically and the related role of LO distribution is illustrated when the architecture is not perfectly symmetric, i.e., when the connection to each antenna of the array leads to delay or phase differences. Alternative architectures based on resonant tunneling diodes are considered. They could offer a relevant alternative towards higher frequencies, based on simpler architectures and sufficiently large antenna arrays to compensate for output power limitations. Another alternative is to build on switched beam antenna lenses. They offer promising alternatives to traditional phased arrays, saving a lot on power consumption and area, assuming the related integration challenges can be solved.

Second, reflective intelligent surfaces (RIS) are considered. They offer a useful diversity solution to extend the range of line-of-sight dominated propagation at high frequencies. Models of their impact are created to simulate the 3D radiation pattern they can create and related reflection parameters. A connection to two prototype designs is also made, validating models, and illustrating flexible candidate architectures. Concerning important system integration aspects, the main RIS properties and control functions and protocols are identified.

Third, system-on-chip architectures are investigated. Advanced DSP and AI capabilities are required for future systems. Especially, AI is expected to play an important role in resource allocation optimization and network management. Specific architectures and accelerators are considered. A RISC-V core processor is considered to support the main DSP operations and the related execution complexity is simulated. Additionally, a scalable design approach integrating cores and accelerators is proposed, together with security aspects. Energy harvesting and power management aspects are also investigated. This study focuses on the combination of different energy sources.

Finally, ultra-low cost and power devices are considered. Battery-operated or energy-neutral IoT devices face challenging trade-offs and require thorough optimization. Various techniques supporting reduced power consumption are reviewed, such as duty-cycling and power saving modes. Another optimization is channel coding gain vs. complexity. To achieve such efficient IoT designs, many enabling technologies are considered. They include energy harvesting, specific energy-aware protocols, tiny machine learning (tinyML) solutions or wake-up radios. Connectionless dedicated modes with the 6G RAN are expected. In order to move towards demonstration of such solutions, the progress on zero-energy proof-of-concept is described with lab and field trials, validating an approach based on backscattering of reference signals.

Table of Contents

List of Tables

List of Figures

Acronyms and abbreviations

Term	Description
6G	Sixth Generation
ADC	Analogue-to-Digital Converter
AIoT	Ambient Internet of Things
AM	Amplitude Modulation
ANN	Artificial Neural Network
ARM	Advanced Risc Machine
ARVTDNN	Augmented Real-Valued Time-Delay Neural Network
ASK	Amplitude Shift Keying
ASM	Accelerator Support Module
BCB	Benzocyclobutene
BD	Backscatter Device
BER	Bit Error Rate
BiCMOS	Bipolar Complementary Metal-Oxide Semiconductor
BLE	Bluetooth Low Energy
BLN	Battery-Less Node
BPSK	Binary Phase Shift Keying
BRAM	Block Random Access Memory
BS	Base Station
CFO	Carrier Frequency Offset
CMOS	Complementary Metal-Oxide Semiconductor
CN	Core Network
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CRC	Cyclic Redundancy Check
CRS	Cell specific Reference Signal
CW	Combiner Weight computation
DAC	Digital-to-Analogue Converter
DC	Direct Current
DL	Downlink
DMA	Direct Memory Access
DMAA	Dynamic Metasurface Antenna Array

DNN	Deep Neural Network
DPD	Digital Pre-Distortion
DRAM	Dynamic Random Access Memory
DSP	Digital Signal Processor, Digital Signal Processing
EC	Energy Combiner
EH	Energy Harvesting
EIRP	Effective Isotropic Radiated Power
EN	Energy Neutral
EVM	Error Vector Magnitude
FPGA	Field-Programmable Gate Array
FSK	Frequency Shift Keying
GaN	Gallium Nitride
GPU	Graphic Processing Unit
HW	Hardware
ICI	Inter-Carrier Interference
IF	Intermediate Frequency
IMPATT	Impact avalanche and transit-time
InP	Indium Phosphide
IoE	Internet of Everything
IoT	Internet of Things
I/Q	In-phase / Quadrature
ISA	Instruction Set Architecture
ISI	Inter-Symbol Interference
LI	Local Inference
LNA	Low Noise Amplifier
LO	Local Oscillator
LOS	Line-Of-Sight
LPF	Low-Pass Filter
LPWA	Low-Power Wide Area
LTE	Long Term Evolution
LTI	Linear Time-Invariant
LUT	Look-Up Table
M³	Microkernel-based operating system for heterogeneous manycores
MAC	Medium Access Control

MAD	Multiply Add Operation
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
ML	Machine Learning
mMTC	Massive Machine Type Communication
MMU	Memory Management Unit
MTC	Machine Type Communications
MU-MIMO	Multi-User Multiple Input Multiple Output
NAS	Non-Access Stratum
NB-IoT	NarrowBand IoT
NDC	Negative Differential Conductance
NLOS	Non Line-Of-Sight
NoC	Network-on-Chip
NPU	Neural Processing Unit
OFDM	Orthogonal Frequency Division Multiplexing
OOK	On-Off Keying
OS	Operating System
OSF	Over-Sampling Factor
PA	Power Amplifier
PAPR	Peak-to-Average Power Ratio
PE	Processing Element
PHY	Physical layer
PLL	Phase-Locked Loop
PM	Phase Modulation
PMIC	Power Management Integrated Circuit
PMP	Physical Memory Protection
Pout	Output Power
PPG	Pulse Pattern Generator
ppm	Part Per Million
Psat	Saturation Power
PSD	Power Spectral Density
QAM	Quadrature Amplitude Modulation
QoS	Quality-of-Service
QPSK	Quadrature Phase Shift Keying

RACH	Random Access Channel
RAN	Radio Access Network
RDMA	Remote Direct Memory Access
RedCap	Reduced Capability
RF	Radio Frequency
RF-WPT	Radio Frequency Wireless Power Transfer
RI	Remote Inference
RIS	Reflective Intelligent Surface
RISC	Reduced Instruction Set Computer
RNN	Recurrent Neural Network
RSoC	Relative State of Charge
RTD	Resonant Tunneling Diode
RVTDNN	Real-Valued Time-Delay Neural Network
RVV	RISC-V Vector extension
RX	Receiver
SBD	Schottky Barrier Diode
SC	Single Carrier
SC-FDE	Single Carrier with Frequency-Domain Equalization
SCO	Sampling Clock Offset
SF	Subframe
SG	Signal Generator
SiGe	Silicon Germanium
SISO	Single Input Single Output
SNR	Signal-to-Noise Ratio
SoC	System-on-Chip
SPF	Signal Processing Function
SRAM	Static Random Access Memory
SS	Synchronization Sequence
SSB	Single Side-Band
SVE	Scalable Vector Extension
SW	Software
TCU	Trusted Communication Unit
TLB	Translation Look-aside Buffer
TRx	Transceiver

TUNNETT	Tunnel Transit Time
TX	Transmitter
UE	User Equipment
UL	Uplink
UP	User Plane
UPF	User Plane Function
UTC-PD	Uni-Traveling Carrier Photodetector
VR	Virtual Reality
WLAN	Wireless Local Area Network
WPT	Wireless Power Transfer
ZED	Zero-Energy Device
ZF	Zero-Forcing

1 Introduction

6G promises new applications and will accordingly require new technologies to enhance the related key performance indicators (KPIs) such as throughput and reliability. In this deliverable, various technologies supporting 6G are investigated and first designs are presented.

A first example is the extension of frequency bands towards sub-THz, investigated in Section 2. While low frequencies are well suited for wide-area cell coverage, they only offer a limited capacity. On the contrary, higher frequencies are more limited in propagation range, but can offer a huge capacity in places with high demand. In order to benefit from the related wide bandwidth and flexible integration of a large number of antennas, specific hardware (HW) design challenges need to be tackled and various architecture options are possible.

Sub-THz architectures can be significantly different from transceivers at lower frequencies. The large number of antennas, wide-bandwidth digital signal processing (DSP) requirements and power efficiency constraints on analogue components lead to specific trade-offs when dimensioning sub-THz systems. Those aspects are investigated based on link budget considerations and power consumption modelling.

Additionally analogue HW non-idealities become more prominent at higher frequencies due to design challenges and semiconductor limitations. Various non-idealities can have an increased impact on the signal quality, which limits the system throughput due to distortion or interference terms. Several non-idealities are reviewed for sub-THz architectures. In particular, the impact of phase noise for distributed antenna arrays, receives a special look due to its potential to strongly degrade the system performance.

Alternative architectures are also investigated. One option is to use resonant tunnelling diodes (RTDs) which are especially appealing for sub-THz frequencies, where traditional active devices become less effective. Related challenges include output power limitations, antenna array integration and support for various modulations. Another option is to use switched beams antenna lenses. Unlike phased arrays, lens-based solutions activate only one antenna front-end chain per beam and achieve beamforming gain via the lens optical properties. Such architecture can reduce the power consumption and area of the antenna arrays.

Reflective intelligent surfaces (RIS) have recently emerged as a solution to enhance the coverage of line of sight (LOS) dominated propagation at high frequencies. By adaptively refocussing the incoming signal to a specific direction, they can enhance the system reliability by improved propagation diversity and offer a solution to link blockage problems. RIS-based solutions are investigated in Section 3.

A simulation environment is used to model the RIS impact on the system performance. It computes the received power in the target direction as well as non-wanted directions. Simulation results are validated by comparison to a measured prototype. It includes both active and passive RIS configurations. A second prototype based on varactors is investigated. It enables extra flexibility in the RIS control. S parameters and radiation pattern are simulated.

Additionally, RIS integration challenges are considered. Control aspects are essential. Control can be performed from infrastructure, UEs, or other devices. The key properties of a RIS solution and related control protocols are described and quantified, such that RIS components can be incorporated into the network.

The new 6G services and the diversity of deployment and operating scenarios will further increase the challenge of system-on-chip (SoC) architecture design for 6G devices. This complexity results from the wide range of performance and cost expectations. In addition, key factors such as scalable and efficient signal processing, security aspects, efficient energy management and the integration of AI functions will significantly shape the design landscape. In Section 4, the aim is to research energy-efficient 6G SoC architectures and particularly specific aspects regarding scalable signal and AI processing, trusted hardware-software (HW/SW) SoC concepts to protect against potential attacks and misuse and approaches for multi-source energy harvesting and management.

One of the focuses is to explore the applicability of RISC-V (Reduced Instruction Set Computer, 5th generation) and its vector instruction extensions for 6G signal processing. The study covers the benchmark definition, i.e., representative signal processing algorithms and fundamental signal processing kernels, followed by the investigation of benchmark vectorization using RISC-V processor. A comprehensive

performance analysis enable identify critical kernels and gaps for further algorithm-hardware-software optimizations.

Another research aspect involves the design and integration of AI components into SoC. Initially, a flexible and scalable 3D array accelerator architecture is introduced, employing bit-serial processing to dynamically adapt to precision application requirements. Subsequently, integrating the AI accelerator into a secure SoC platform becomes a significant technical challenge. An assessment of the impact of hardware/software overhead on performance and cost provides insights into the efficiency of proposed approach.

Substantial advancements in processing capabilities of 6G devices necessitates addressing surrounding security and privacy concerns at the System-on-Chip (SoC) level. The primary challenge lies in establishing integrated trustworthiness spanning hardware, runtime operations, and the operating system (OS). Part of this work investigates a secure and scalable SoC architecture adopting a tiled structure wherein physically distinct tiles interconnect through a network-on-chip (NoC). Each tile integrates a trusted communication unit (TCU) responsible for enforcing isolation among tiles. The OS microkernel takes charge of managing communication channels and configuring TCUs, ensuring secure and reliable communication throughout the platform. To analyse the impact of hardware/software overhead associated with trustworthiness, the field-programmable gate array (FPGA) prototype is designed, featuring multiple processing tiles equipped with RISC-V cores.

Energy harvesting (EH) is another important aspect to enable the continuous operation of IoT devices. This involves the conversion of ambient energy into electrical energy, followed by the regulation, storage and management of the generated electrical energy. The energy generated from ambient sources requires a power management integrated circuit (PMIC), which serves as an interface between the energy source and the load to optimize energy harvesting and storage and ensure seamless energy transfer. The aim of this research is to develop a PMIC that can work seamlessly with multiple sources simultaneously. For this purpose, an energy combiner (EC), is needed to efficiently combine energy from different sources into a single buffer. In addition, low-power ML algorithms are investigated to gain insight into energy availability and consumption, enabling predictive energy-aware schedulers and reconfigurable energy buffers.

In the ever-evolving landscape of the Internet of Things (IoT), the demand for energy-efficient and cost-effective devices has become increasingly paramount [LRR+23]. This imperative is fuelled by the widespread integration of IoT into diverse domains, ranging from smart cities to industrial applications, where longevity, minimal maintenance, and sustainability are critical considerations. Indeed, although the fifth-generation (5G) network brought new features to support IoT connectivity more natively [VAK+22], [MBA+22], it falls short of meeting the most stringent requirements in terms of low cost and power consumption as discussed in [HEX223-D52].

Note that using 5G IoT-supporting techniques/technologies in extreme and inaccessible conditions is nowadays problematic due to the need for frequent battery replacement and recharging plugins, which is costly, difficult, and hazardous [LRR+23], [MBA+22]. Furthermore, ultra-small size devices are vital to facilitate the implementation of numerous applications including wearables. Finally, the use of a massive number of devices to ensure efficient and accurate management requires developing low-cost devices to reduce the overall economic cost [NBM+23].

In Section 5, we aim to advance the state-of-the-art research on ultra-low power/cost IoT devices. Therein, we analyze the power consumption of such devices, while addressing the crucial challenge of optimizing energy usage to extend device lifespan and enhance overall efficiency. Also, we explore enabling technologies like EH, including potential energy sources and related technologies/techniques, radio frequency (RF) wireless power transfer (WPT), energy-aware and lightweight protocols and signal processing techniques, TinyML, and intelligent wake-up. In addition, a recent zero-energy proof of concept (PoC), wherein devices harness ambient RF cellular signals for backscatter communication, is revised. Notably, this paradigm shift minimizes complexity and power consumption to magnitudes below existing low-power wide area (LPWA) network technologies. Our insights not only underscore the necessity for energy-efficient IoT solutions but also spotlight the transformative potential of cutting-edge technologies that pave the way for a more sustainable and interconnected future.

2 Sub-THz transceiver design

In order to sustain ever-increasing throughput requests, moving to higher frequencies is an essential ingredient to benefit from a wider bandwidth and flexible integration of a large number of antennas. Sub-THz waves essentially propagate in line-of-sight, unlike lower frequencies. This makes those frequencies less suited to large area coverage, where blockage from the environment requires the use of reflected or diffracted paths. However, they are particularly suited for dense environments where high-capacity short-range hotspots are required. In order to benefit from sub-THz capacity benefits, specific HW design challenges need to be tackled and various architecture options are possible.

First, sub-THz transceiver architectures can be significantly different from transceivers at lower frequencies. The large number of antennas, wide-bandwidth DSP requirements and power efficiency constraints on analogue components lead to specific trade-offs when dimensioning sub-THz systems. In Section 2.1, those aspects are investigated based on link budget considerations and power consumption modelling for different transceiver architectures. A special focus is put on multi-user hotspot scenarios, considering both infrastructure side and user equipment side of the communication link, with various architecture options.

Secondly, Sections 2.2 and 2.3 consider another important aspect: analogue HW non-idealities. Indeed, designing analogue components for higher frequency or covering several GHz of bandwidth is very challenging. With the increase in frequency, various non-idealities can have an increased impact on the signal quality, which limits the system throughput due to distortion or interference. Section 2.2 reviews a number of different HW non-idealities and the way they are expected to scale when extending into sub-THz bands. Section 2.3 zooms in on one of the most important effects: phase noise. Especially, for large antenna arrays, the local oscillator distribution from a central phase-locked loop (PLL) to the different antenna chains can be asymmetrical. This leads to differential phase noise effects which degrade the system performance, especially over a wide bandwidth.

Thirdly, Section 2.4 considers alternative architectures based on RTDs. RTDs are especially appealing for sub-THz frequencies, where traditional active devices become less effective. Output power limitations and antenna array integration are investigated for RTD-based transceivers. Possible modulations supported on such architectures are also explored.

Finally, Section 2.5 investigates switched beams antenna lenses. Alternatively, to traditional phased arrays, where beams are generated based on adaptive phase shifting of the different antennas, lens-based solutions activate only one antenna chain per beam, and achieves beamforming gain via the lens optical properties. Potential benefits in power consumption as well as area are presented.

2.1 Dimensioning of sub-THz architectures

2.1.1 Introduction

Sub-THz communication systems benefit from a wide available bandwidth and as such, promise much larger throughputs compared to more constrained bands. They also come with specific HW implementation constraints, due to the specific constraints on high-frequency electronic circuits, the need for high-speed DSP, and the choice of specific architecture exploiting many antennas in order to compensate for the limited aperture of antenna elements as they scale with the wavelength.

Sub-THz bands mainly target short-range communication links, as they rely mostly on LOS communications. It is hence important to analyse corresponding use cases focusing on large throughputs at short range, in order to model and optimize the corresponding power consumption. Hence, short-range multi-user hotspots are selected as target scenarios. Based on Deliverable D5.2 [HEX223-D52], the following general approach is used:

- Parameterize each scenario and possible architectures
- Compute link budget requirements
- Assess power consumption and optimize trade-offs by sweeping architecture parameters

In this section we dimension architectures by adapting the number of antennas, the output power level of each PA, and the choice of architecture (fully-digital, hybrid partially connected or hybrid fully connected). Modulation and coding rates are adapted accordingly within the bounds from the link budget, selecting the most energy-efficient options providing the required data rate. Other architecture dimensioning parameters such as quantization levels are left for further optimization.

2.1.2 Scenario definition

Three different types of hotspots are specified. A short-range hotspot where an access point communicates with a limited number of devices in a small room, a medium-range hotspot with more users, and a large hotspot where range and number of users are further increased, representing a wider environment such as airport, shopping mall, or outdoor link on a densely populated square. The small and medium hotspot assume a single omnidirectional access point, e.g., an access point from the ceiling where we assume an isotropic antenna radiation pattern covering 360° in azimuth and the necessary elevation range to target all users. The large hotspot assumes three sectors of 120°, similarly to a small base station.

Based on Deliverable D5.2 [HEX223-D52], several parameters need to be specified properly in order to enable modelling and optimization. Some of those parameters are common to the three scenarios and are listed in Table 1.

Table 1. Common parameters to the three hotspot scenarios.

Parameter	Spec	Value
PHY performance	Active (minimal and peak) throughput per UE (DL, UL)	1 Gbps DL, 0.3 Gbps UL
	Target BER (DL, UL)	1e-5, coded
Band of operation and regulations	Carrier frequency	140 GHz
	RF bandwidth	5 GHz
	Possible waveforms, constellations, coding rate/options	OFDM (11ad MCS)
Channel	Type of propagation (LOS, NLOS)	LOS
Multi-antenna / multi-user correlation	Channel correlation between BS antennas	1
	Channel correlation between UE antennas	1
	Channel correlation between UEs	0

Other parameters are specific to each of the three hotspot scenarios and are listed in Table 2. The parameters in both tables will enable system analysis based on link budget and power consumption modelling. Some additional parameters specified in Deliverable D5.2 [HEX223-D52] are not listed here, as they mainly concern end-to-end physical layer (PHY) link simulation. Those relate to specific studies such as interference modelling, channel tap-delay response impact on performance, or mobility. Such studies are not part of the high-level architecture dimensioning. Concerning user separation, this is partially done in spatial dimension (based on *Max simultaneous spatially multiplexed users* in Table 2) and partially in time and frequency (for the higher number of users).

Table 2. Specific parameters for the three different hotspot scenarios and UE.

Parameter	Spec	Small	Medium	Large	UE
Geometry	Target range	10 m	30 m	100 m	
	Angular spread of UEs (seen from BS)	360°	360°	120°	
	Height/position of BS/UE devices	3 m (BS)	5 m (BS)	10 m (BS)	1 m (UE)

Multi-user	Number of simultaneously active users in same band	10	32	120 (40 per sector)	
	Max simultaneous spatially multiplexed users	2	4	8	
Architecture constraints	Max PA power (techno-dependent)	20 dBm	20 dBm	20 dBm	20 dBm
	Max number of antennas and array size	64	256	1024	16
	Antenna element gain, interconnect loss	+2	+2	+4	+2
	Noise figure	8 dB	8 dB	8 dB	8 dB

2.1.3 Link budget and power consumption modelling

Dimensioning of wireless architectures mainly focuses on trade-offs between performance (e.g., throughput, error rate) and power consumption. The performance is assessed based on a link budget computation, while power consumption uses a power modelling tool enabling to estimate the power consumption of various architectures and configurations. It includes the digital power consumption, estimated from the complexity.

The main parameters used by the link budget are listed in Table 1 and Table 2. Additionally, the following assumptions are taken:

- Time and frequency margins (guard band, cyclic prefix): 0.125
- Additional implementation loss margin (imperfect algorithms, estimation errors, distortion and interference): 5 dB extra on the SNR
- PA back-off: 0 dB for BPSK/QPSK, 3 dB for 16-QAM, 6 dB for 64-QAM (simulated on OFDM including oversampling filter and coding rate 3/4)
- Simulated SNR thresholds for the different modulation and coding options
- MU-MIMO digital precoding gain: (N - K + 1) assuming N digital chains serving K users, considering (K - 1) degrees of freedom being consumed for ZF interference cancellation

Based on that, the maximum supported throughput can be computed for various architectures and link parameters. Different architecture choices will lead to different maximum throughput values.

The power consumption model is based on [DCS+21] and [DWZ+20]. The model splits contributions from the PA, the analogue front-end, the digital signal processing and the general power supply. Zero-IF (direct conversion) analogue architectures are considered. Only support of the wireless physical layer processing is included in the model, i.e., higher layer protocols, application processes or additional connections (such as fronthaul/backhaul connection) are not part of the model. It can be applied to both base stations and user equipment. Default quantization options assume 5-bit for channel decoding and 10-bit for other operations [DWZ+20].

Various architecture types can be considered for multi-user communications: full digital (Figure 2-1), hybrid fully connected (Figure 2-2) or hybrid partially connected (Figure 2-3). The goal is to figure out which are realistic or have a prohibitive complexity. Another architecture type (analogue beamforming only) could be considered but it is not realistic as it leads in general to an excessive inter-user interference, making it not useable unless users are strongly separated in specific geometries. In complement to the spatial multiplexing offered by those architectures, TDMA can also be used. However, it only shares the available throughput over multiple users, rather than multiplying the total available throughput by adding extra layers.

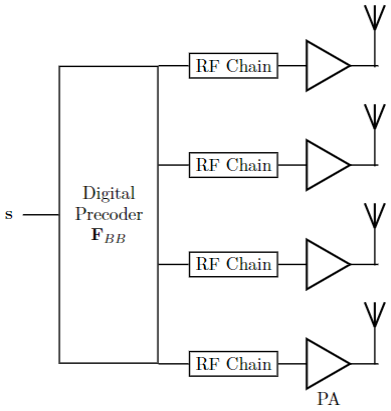


Figure 2-1: Block diagram of the fully digital architecture having four RF chains, connected to four PAs and antennas.

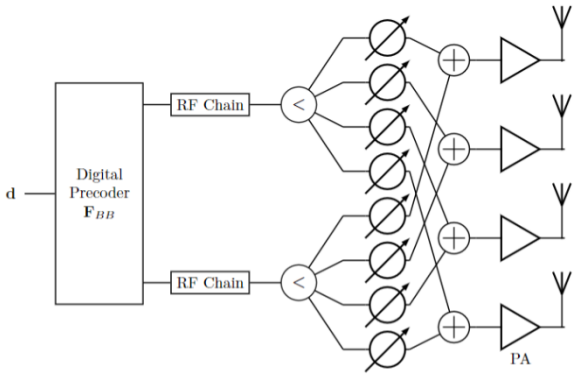


Figure 2-2: Block diagram of the hybrid fully connected architecture having two RF chains connected to four PAs and antennas where each RF chain is connected to each PA by utilizing a phase shifter.

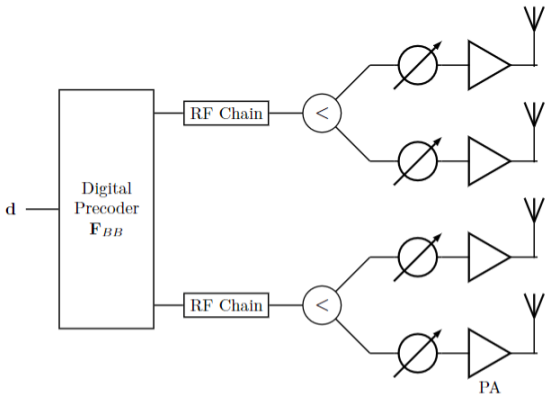


Figure 2-3: Block diagram of the hybrid partially connected architecture having two RF chains connected to four PAs and antennas where each RF chain is connected to a disjoint subset of PAs by utilizing a phase shifter.

2.1.4 Energy efficiency optimization results

Energy efficiency optimization assumes that duty-cycling can be exploited. This means that if the system requires, e.g., 1 Gbps while maximum allowed throughput is 10 Gbps, short bursts of data can be transmitted 10% of the time at 10 Gbps while the system can be put in sleep mode the rest of the time. This is generally more efficient in terms of system energy efficiency [DDL15], and if designed properly, all electronic

components are able to switch on/off in a relatively fast manner, such that the impact of this strategy on application latency is negligible.

Energy efficiency optimization results are presented considering the three use cases introduced in 2.1.2 and the architectures presented in 2.1.3. The optimization is performed by varying the PA output power and number of transmit antennas, considering suitable modulation and coding schemes for the given link budget. Specifically, PA output power is assumed to be between 10 dBm and 18 dBm considering an InP PA suitable for sub-THz wireless communication [DCS+21]. A brute force parameter search has been done to find the optimal architecture configuration in energy efficiency. The optimization minimized energy per bit while guaranteeing that the required throughput (in this example 1 Gbps per user) can be guaranteed.

Figure 2-4 illustrates the power consumption comparison for 2 spatially separated data streams of a suboptimal fully digital architecture where all parameters are maximized in order to achieve the peak throughput (64 antennas, 18 dBm output PA output power) with the optimized fully digital architecture having 4 antennas and 12 dBm PA output power, considering a small hotspot scenario. Although the optimal fully digital architecture supports a lower throughput (26 vs. 53 Gbps) as it only supports a lower modulation and coding due to reduced EIRP, it is much more energy-efficient compared to the suboptimal fully digital architecture. The reason is that the combination of a large number of antennas and high PA output power dominates the Tx power consumption of the suboptimal fully digital architecture. The energy efficiency of the optimal fully digital architecture is more than a factor 10 better than the suboptimal fully digital architecture. Both architectures support the target rate (2 users, 1 Gbps per user) and can hence be compared on energy efficiency after duty-cycling.

In Figure 2-4 and subsequent figures representing the system power consumption, three states are always represented for the base station side: the transmitting state (downlink), the receiving state (uplink) and the channel estimation (receiving uplink pilots). In each state, the instantaneous power is split into the four main categories (PA, analogue front-end, digital baseband and power supply), and some of those categories are further refined into sub-components.

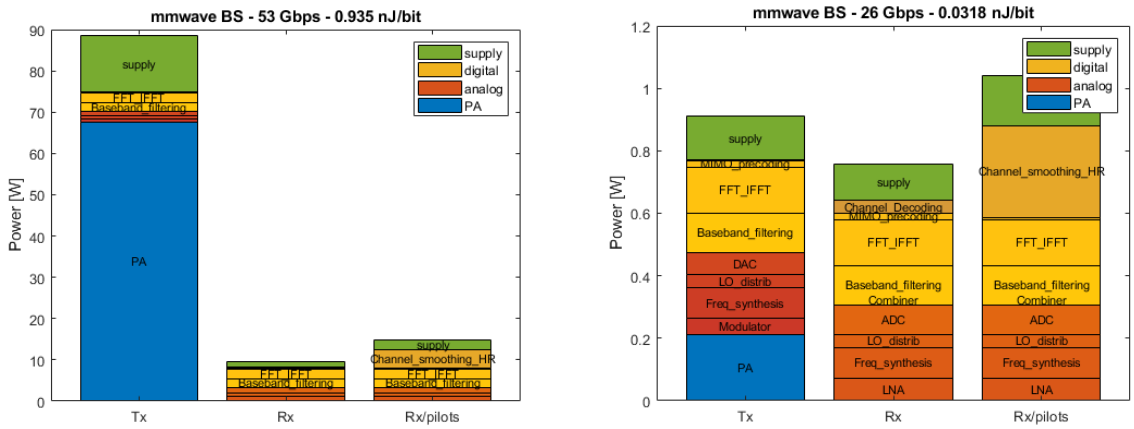


Figure 2-4: Power consumption comparison of a suboptimal (left) and optimal (right) fully digital architecture having 2 data streams in a small hotspot scenario.

Figure 2-5 illustrates the power consumption comparison of hybrid partially connected architectures having 2 data streams and 2 RF chains, with 64 antennas and 18 dBm PA output power for the sub-optimal (max-performance) case, and 4 antennas and 14 dBm PA output power for the optimal case, considering a small hotspot scenario.

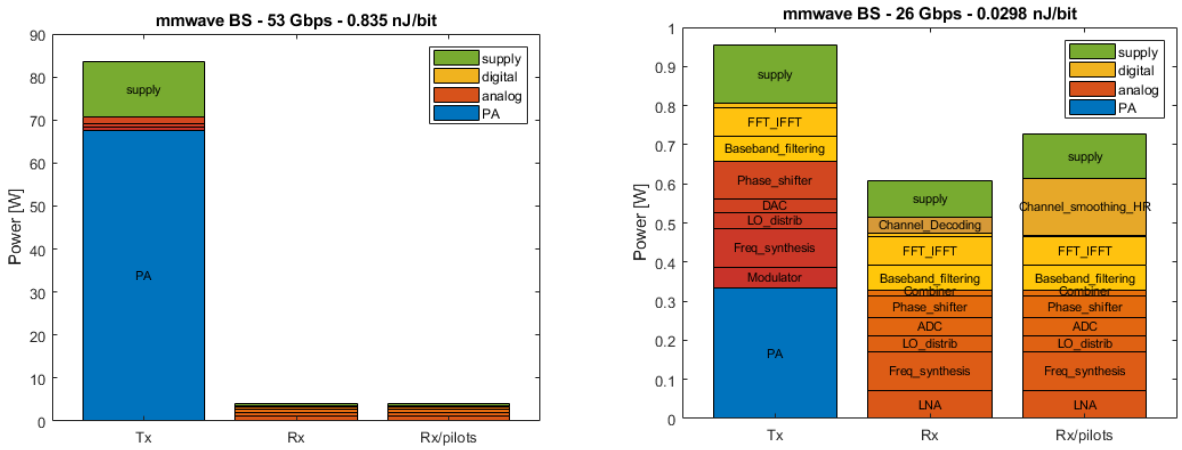


Figure 2-5: Power consumption comparison of a suboptimal (left) and optimal (right) hybrid partially connected architecture having 2 data streams in a small hotspot scenario.

Figure 2-6 illustrates the power consumption comparison of the suboptimal hybrid fully connected architecture having 2 data streams, 64 antennas, and 18 dBm PA output power with the optimized hybrid fully connected architecture having 2 data streams, 4 antennas, 12 dBm PA output power considering a small hotspot scenario.

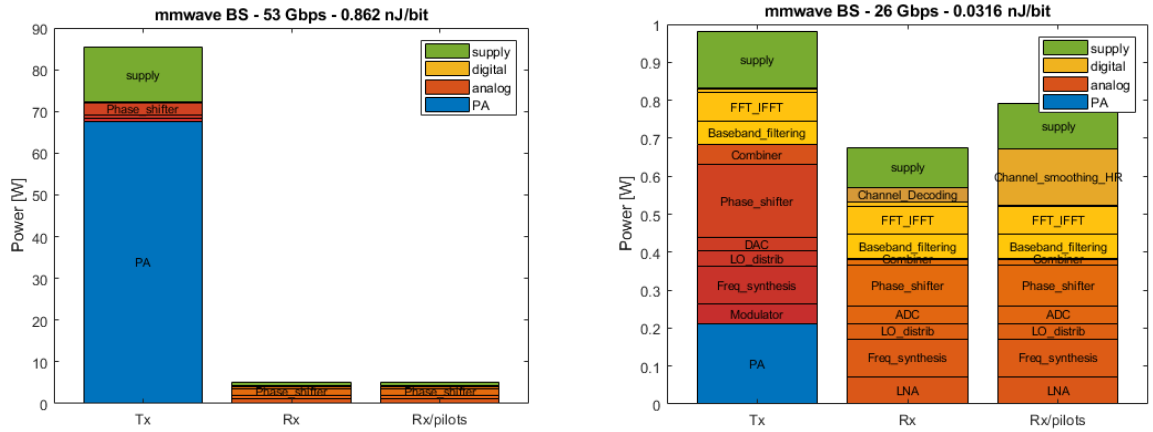


Figure 2-6: Power consumption comparison of a suboptimal (left) and optimal (right) hybrid fully connected architecture having 2 data streams in a small hotspot scenario.

All three optimized architectures in small hotspot scenario supports a throughput of 26 Gbps to serve two users simultaneously. The hybrid partially connected architecture is slightly more energy-efficient compared to hybrid fully connected and fully digital architectures. Although hybrid partially connected architecture utilizes a higher PA output power, it has simpler analogue front-end compared to the hybrid fully connected and reduced digital baseband compared to fully digital architecture, respectively.

Figure 2-7 illustrates the power consumption comparison of the suboptimal fully digital architecture having 4 data streams, 256 antennas, 18 dBm PA output power with the optimized fully digital architecture having 4 data streams, 32 antennas, 14 dBm PA output power considering a medium hotspot scenario.

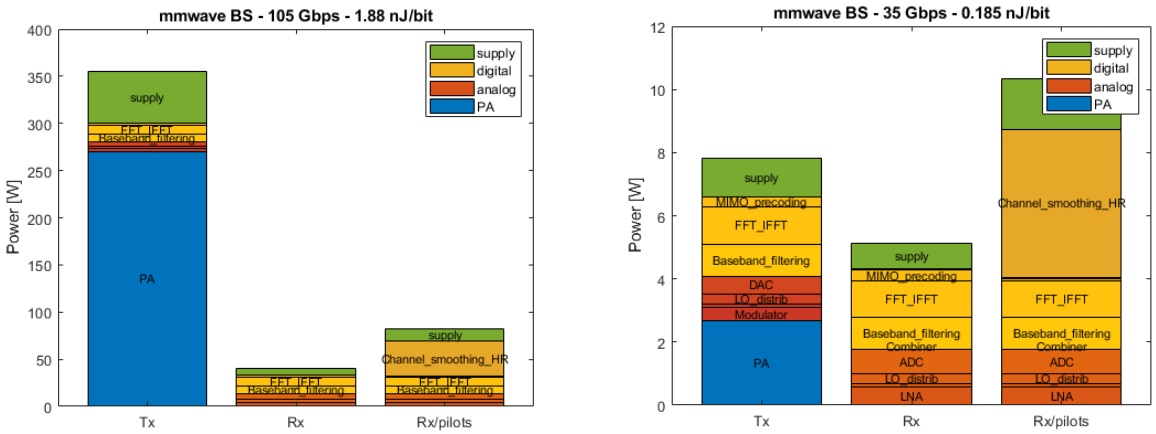


Figure 2-7: Power consumption comparison of a suboptimal (left) and optimal (right) fully digital architecture having 4 data streams in a medium hotspot scenario.

Figure 2-8 illustrates the power consumption comparison of the suboptimal hybrid partially connected architecture (left) having 4 data streams, 256 antennas, 18 dBm PA output power with the optimized hybrid partially connected architecture (right) having 4 data streams, 100 antennas, 12 dBm PA output power considering a medium hotspot scenario.

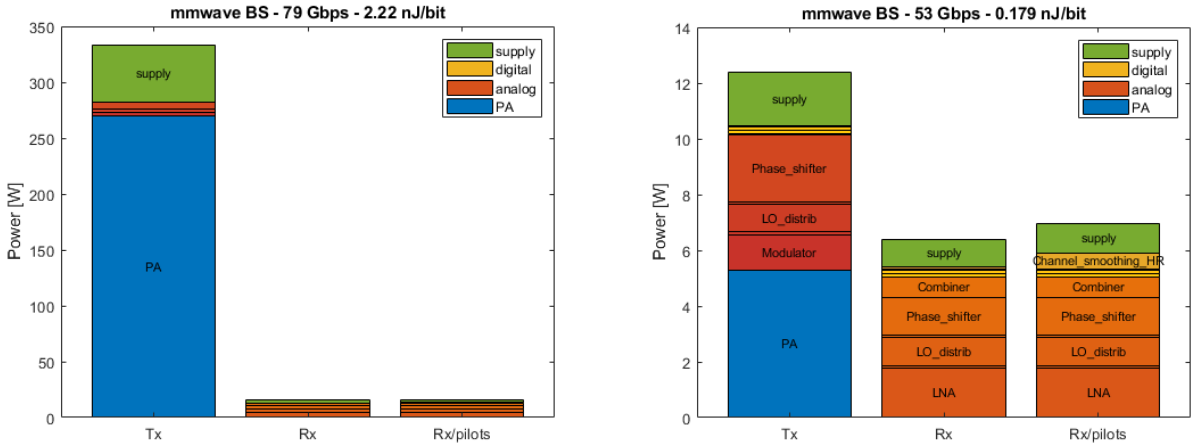


Figure 2-8: Power consumption comparison of a suboptimal (left) and optimal (right) hybrid partially connected architecture having 4 data streams in a medium hotspot scenario.

Figure 2-9 illustrates the power consumption comparison of the suboptimal hybrid fully connected architecture having 4 data streams, 256 antennas, 18 dBm PA output power with the optimized hybrid fully connected architecture having 4 data streams, 64 antennas, 14 dBm PA output power considering a medium hotspot scenario.

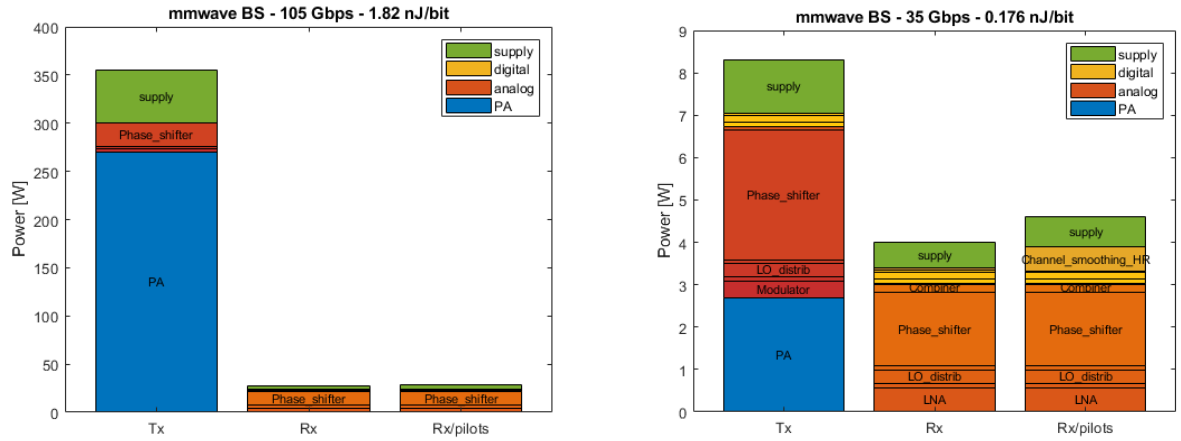


Figure 2-9: Power consumption comparison of a suboptimal (left) and optimal (right) hybrid fully connected architecture having 4 data streams in a medium hotspot scenario.

The hybrid fully connected architecture is the most energy-efficient architecture compared to fully digital and hybrid partially connected architecture considering a medium hotspot scenario. It benefits from a higher EIRP compared to the hybrid partially connected, as each steam benefits from the full array gain, allowing to relax the dimensioning in number of antennas or PA output power. On the other hand, it allows simplified digital baseband compared to fully digital architectures.

Figure 2-10 illustrates the power consumption comparison of the suboptimal fully digital architecture having 8 data streams, 1024 antennas, 18 dBm PA output power with the optimized fully digital architecture having 8 data streams, 64 antennas, 14 dBm PA output power considering a large hotspot scenario. The fully digital architecture has a simple analogue front-end circuit compared to other hybrid architectures at the expense of additional digital processing.

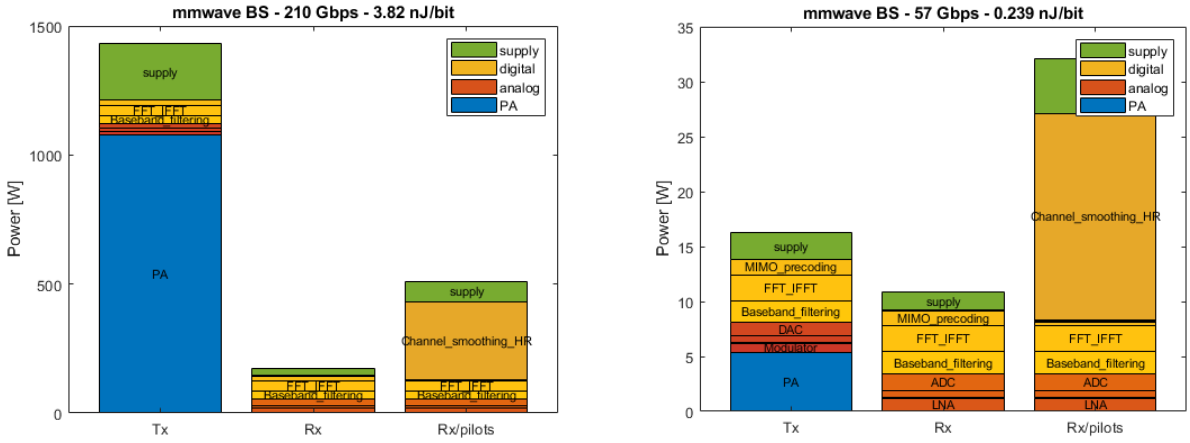


Figure 2-10: Power consumption comparison of the suboptimal (left) and optimal (right) fully digital architecture having 8 data streams in a large hotspot scenario.

Figure 2-11 illustrates the power consumption comparison of the suboptimal hybrid partially connected architecture having 8 data streams, 1024 antennas, 18 dBm PA output power with the optimal hybrid partially connected architecture having 8 data streams, 160 antennas, 14 dBm PA output power, considering a large hotspot scenario.

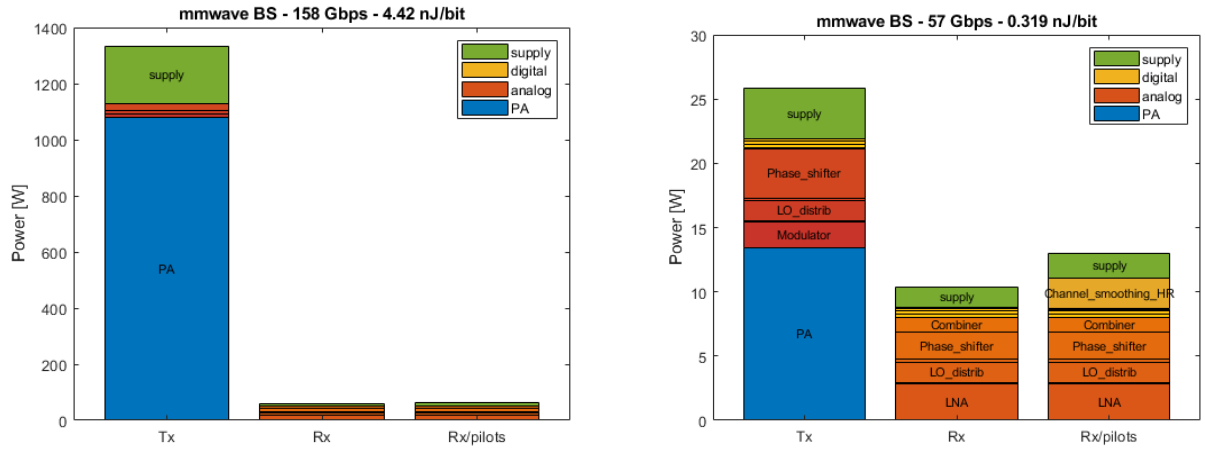


Figure 2-11: Power consumption comparison of the suboptimal (left) and optimal (right) hybrid partially connected architecture having 8 data streams in a large hotspot scenario.

Figure 2-12 illustrates the power consumption comparison of the suboptimal hybrid fully connected architecture having 8 data streams, 1024 antennas, 18 dBm PA output power with the optimized hybrid fully connected architecture having 8 data streams, 64 antennas, 14 dBm PA output power, considering a large hotspot scenario. The hybrid fully connected architecture exploits reduced digital power consumption compared to fully digital architecture at the expense of additional analogue components, i.e., phase shifters, splitters and combiners.

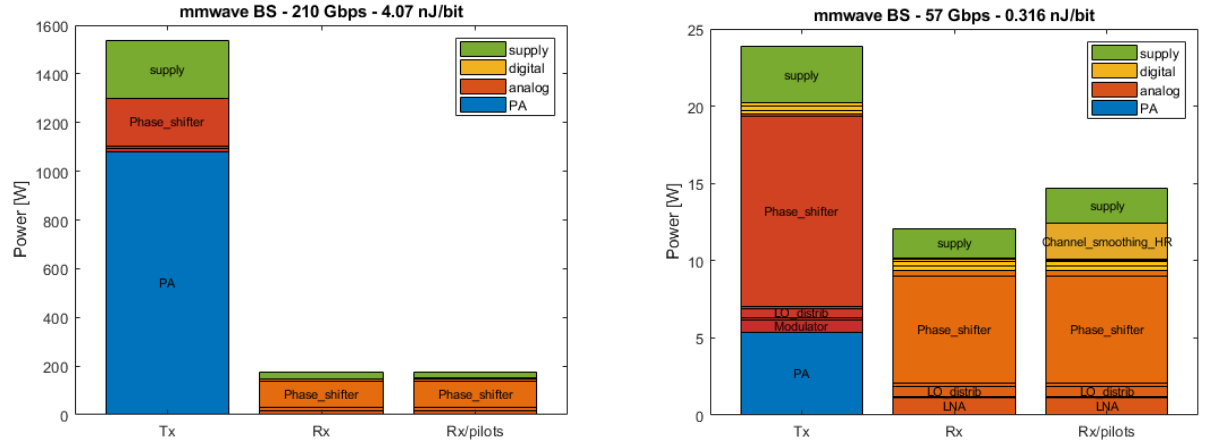


Figure 2-12: Power consumption comparison of the suboptimal (left) and optimal (right) hybrid fully connected architecture having 8 data streams in a large hotspot scenario.

Table 3 summarizes the energy efficiency values of the optimized fully digital, hybrid partially connected and hybrid fully connected architectures considering small, medium and large hotspot scenarios, respectively. Note that depending on the application scenario, different architecture options become favourable. In short range communications, hybrid partially connected is suitable while in long range communications, a fully digital architecture is favourable. The main reason is that in long range, the increased analogue and PA power necessary to bridge the link budget or achieve a fully-connected hybrid architecture, make the excess digital

power of a fully digital architecture more acceptable. In moderate range communications, hybrid fully connected is favourable due to balanced analogue front-end and digital baseband complexity.

Table 3: Summary of the most energy efficient architectures per scenario, based on 1 Gbps per user in downlink and scenario parameters defined in 2.1.2.

Architecture/Scenario	Small Hotspot	Medium Hotspot	Large Hotspot
Fully digital	0.0318 nJ/b	0.185 nJ/b	**0.239 nJ/b**
Hybrid partially connected	**0.0298 nJ/b**	0.179 nJ/b	0.319 nJ/b
Hybrid fully connected	0.0316 nJ/b	**0.176 nJ/b**	0.316 nJ/b

While optimized configurations lead to similar order of magnitude on the energy efficiency of the three optimized architectures, it is important to realize how critical it is to properly dimension the architecture. As shown in the different figures of Section 2.1.4, the difference between a baseline configuration (with maximum number of antennas and output power) and the optimum configuration is often a factor 10 to 100 in energy efficiency.

It is also worth noting that although the three architecture types lead to similar energy efficiency in Table 3, they correspond to different dimensioning parameters (number of antennas, output power, peak throughput). It means that different types of architectures can generally offer efficient solutions, provided they are dimensioned accordingly. Logically, moving from less demanding (small hotspot) to more demanding (large hotspot) configurations strongly impacts the energy efficiency, with a factor 10 in this example. Interestingly, the optimal configuration has a different split of the power consumption for each architecture type, with more digital dominance for the fully-digital, more PA dominance for the hybrid partially connected, and more analogue front-end dominance for the hybrid fully-connected case. This is especially visible for the large hotspot case, smaller cells lead a more balanced power between the three categories of components.

At a lower granularity level, the model also gives insights on sub-components dominating the power of specific architectures. For example, fully-digital architectures in medium to large hotspots are especially dominated by the channel smoothing when processing uplink pilots, as it performs matrix operations scaling with both the number of user streams and the number of antennas. Another example is the comparison of fully-connected and partially-connected hybrid architectures, showing the large impact of phase shifters in large configurations, as all RF chains are connected to all antennas.

2.2 Reviewing models of hardware non-idealities

2.2.1 Introduction

A wireless transceiver consists of two major building blocks namely the RF front-end and digital baseband. The front-end performs functions such as frequency conversion, filtering, and amplification, while the digital baseband performs functions such as synchronization, equalization, mapping/demapping, encoding/decoding. Front-end components are made of electronic devices such as resistors, inductors, capacitors, and transistors. Those electronic devices are not perfect and can behave differently from what is expected for many reasons such as materials differences, temperature variation, aging, mismatch between two identical devices, etc. HW impairments are unavoidable, and those impairments may increase as working frequency increases. While impairments in digital circuits are negligible due to the binary nature of digital signals, those hardware impairments lead to non-idealities of front-end components, introducing in-band and out-of-band interference terms and degrading the performance of wireless systems. In this section, we first review hardware non-idealities and state-of-the-art models, then we provide an assessment of necessary model upgrades towards sub-THz communication in which the carrier frequency, bandwidth and number of antennas increase.

We consider Zero-IF (Intermediate Frequency) or direct-conversion architectures converting the RF signal directly to baseband, using an LO tuned to the carrier frequency. At the receiver for instance, the received signal at the antenna is passed through a broadband preselection filter removing out-of-band energy and is then amplified by an LNA (Low Noise Amplifier). The signal is then down-converted to baseband by I/Q mixers which have a phase delay of 90º between them and then further processed by ADCs (Analogue-to-Digital Converters) and sent to digital baseband. Figure 2-13 illustrates the corresponding transmit/receive architectures and the non-idealities associated to the different components. Those are listed in Table 4 and the corresponding models are described in the following sub-sections.

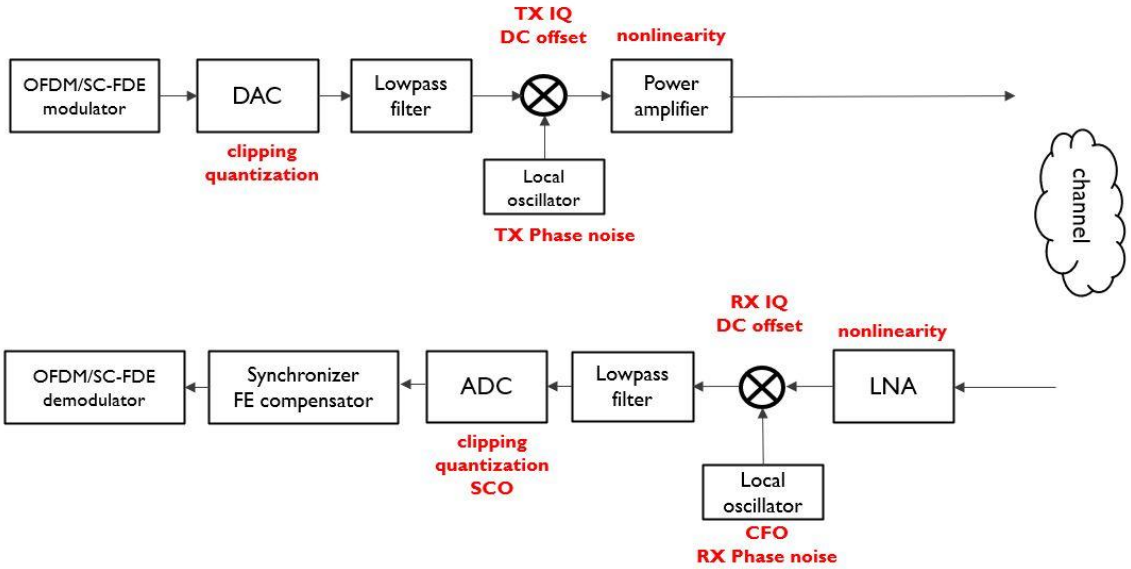


Figure 2-13: OFDM/SC-FDE system block diagram and the sources of different non-idealities in the transmitter (top) and receiver (bottom).

Table 4: List of non-idealities related to the different components.

Non-idealities	RF component (main)
PA non-linearity	Power Amplifier/LNA
Phase noise	Local Oscillator
IQ imbalance	Mixer, Low-Pass Filter
DC Offset	Mixer, Local Oscillator
Sampling Frequency Offset	ADC/DAC
Carrier Frequency Offset	Local Oscillator
DAC/ADC non-idealities, clipping/quantization	DAC/ADC
Phase shift error	Phase shifters

2.2.2 Power amplifier non-linearity

2.2.2.1 Introduction and cause of non-idealities of the PA

The power amplifier is an essential electronic component in front-end architecture, which is responsible for amplifying the electrical signals to be transmitted. However, the amplification is not always linear as the amplitude of input signal increases. When the input signal amplitude becomes too large, the amplifier's output

can reach its maximum limit, resulting in clipping or flattening of the output waveform. The memory effect is another non-ideality of the power amplifier. The memory effect means that the current output of the power amplifier is not only depending on the current input but on previous inputs as well. The major contributor to memory effects in many PAs is the change in drain (or collector) voltage due to non-zero bias/power supply impedance [Mau14]. Change in gate (or base) voltage can also be a significant contributor to short-term memory effects in ns time scale. Dynamic thermal effects and semiconductor trapping effects are other causes of long-term memory effects in ms time scale.

Those hardware impairments of power amplifiers result in gain compression, harmonic distortion, inter-modulation distortion, phase distortion, adjacent channel interference, etc. Those will increase the EVM (Error Vector Magnitude) for communications particularly for high PAPR (Peak-to-Average Power Ratio) waveforms like OFDM (Orthogonal Frequency-Division Multiplexing). They will also cause an increase in noise floor on the delay-Doppler profile for sensing applications.

2.2.2.2 Memoryless models

Based on analytical nonlinear models, the nonlinear behaviour of a PA can be described in a general manner by a power series expansion. The odd order nonlinear terms bring interference in band, which can't be filtered out. They are hence the most important terms in non-linear PA modelling. Even order harmonics may also cause some distortion, but generally are much lower thanks to design symmetry, and can be neglected. The cubic model $A(|x|) = ax(1 + |x|^2)$ is a simple analytical model, which only takes the third order nonlinear term into account. The modified Rapp [HH97], Ghorbani [GS91], modified Saleh [OML09], and White [WBJ03] are most popular memoryless models reported in Hexa-X deliverable D2.2 [HEX20-D23]. They are summarized in Table 5. Each of the parameters in the models has specific meaning available in the references and hence not discussed here in detail. Different memoryless models have been fitted to a 290 GHz SiGe amplifier based on measured and simulated AM-AM and AM-PM distortion datasets [HEX20-D23]. Among those, the modified RAPP model gives meaningful behaviour estimation outside the measurement range, which is suitable for system simulation [HEX20-D23]. The memoryless models are traditionally used for narrowband high-power nonlinear PA systems.

Table 5: Common memoryless PA distortion models.

Model	Mod. Rapp [HH97]	Ghorbani [GS91]	Mod. Saleh [OML09]	White [WBJ03]
$A(\|x\|)$	$\frac{g_r}{\left(1+\left(\frac{\|x\|}{x_{sat}}\right)^{2s}\right)^{\frac{1}{2s}}}$	$\frac{a_1\|x\|^{a_2}}{1+a_3\|x\|^{a_2}}+a_4\|x\|$	$\frac{a_s\|x\|}{1+b_s\|x\|^2}$	$a_w(1-\exp(-b_w\|x\|)$ $+c_w\|x\|\exp(-d_w\|x\|^2)$
$\varphi(\|x\|)$	$\frac{\alpha\|x\|^{q_1}}{1+\left(\frac{\|x\|}{\beta}\right)^{q_2}}$	$\frac{b_1\|x\|^{b_2}}{1+b_3\|x\|^{b_2}}+b_4\|x\|$	$\frac{c_s\|x\|}{1+d_s\|x\|^2}$	N/A

2.2.2.3 Volterra-based models with memory effect

The memoryless behavioural models mentioned above are developed based on simple single-tone simulations or measurements, which are traditionally used for narrowband. They do not exhibit the memory effects caused by the dependency of the single-tone characteristics on frequency.

The Volterra series is a universal mathematical tool to describe any nonlinear function with memory effects [Mau14]. However, due to its high complexity, a wide variety of analytical models have been developed to overcome the computational complexity when the system has finite memory. In general, Volterra series analysis usually consists of linear or nonlinear filters with finite bandwidth that determine the frequency selectivity of the system. In the time domain, these filters are represented by their kernel functions (impulse response) where the support of the kernel represents the memory in the system [MMK+06]. Different Volterra-based models are illustrated in Figure 2-14.

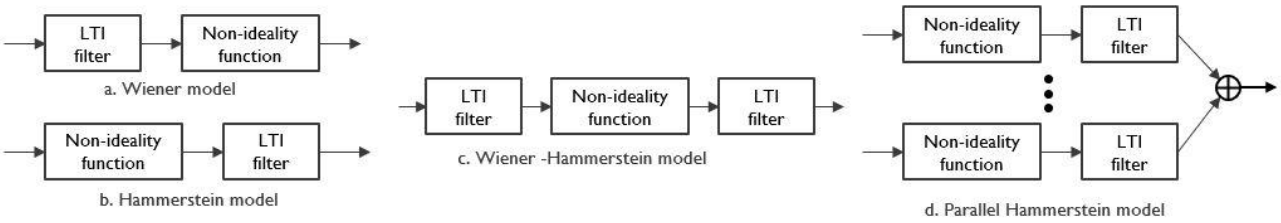


Figure 2-14 Summary of Volterra-based box-oriented models including filters

The Wiener [CBV01] [GH09] model is explained in (2-1), where a_k are the polynomial coefficients of the non-ideality with non-ideality order of K and $h(m)$ are the filters with memory depth of M. the Wiener model is formed by a linear filter followed by a nonlinearity functionality.

$$y_W(n) = \sum_{k=1}^{K} a_k \cdot \left[\sum_{m=0}^{M-1} h(m)x(n-m) \right]^k \tag{2-1}$$

The Hammerstein [NZ01] model is explained in (2-2), where a_k are the polynomial coefficients of the non-ideality with non-ideality order of K and $g(m)$ are the filters with memory depth of M. Different from Wiener, the Hammerstein model is formed by a nonlinearity functionality followed by a linear filter. The Parallel Hammerstein model is formed by combining the outputs from several Hammerstein models.

$$y_H(n) = \sum_{m=0}^{M} g(m) \cdot \sum_{k=1}^{K} a_k x^k(n-m) \tag{2-2}$$

The Wiener-Hammerstein [BCM01] model is explained in (2-3), where the Wiener model is cascaded with the Hammerstein model.

$$y_{WH}(n) = \sum_{m_2=0}^{M-1} g(m_2) \cdot \sum_{k=1}^{K} a_k \cdot \left[\sum_{m_1=0}^{M-1} h(m_1)x(n-m_1-m_2) \right]^k \tag{2-3}$$

The memory polynomials model [DZM+04] [KK03] further simplifies the Hammerstein model by replacing filter $g(m)$ with a single tap filter for different orders k. Combining the different filters and power series coefficients into a 2-D array a_{km}, the memory polynomials model is explained in (2-4). It represents a special simple case of the general Volterra model in discrete time as explained in Figure 2-15.

$$y_{MP}(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} a_{km} x(n-m) \cdot |x(n-m)|^k \tag{2-4}$$

The general memory polynomials model [MMK+06] is explained in (2-5), where delayed both positive and negative cross-terms are added to the memory polynomial model. The advantage of this model is that all the coefficients are in linear form, and they can be simply and robustly estimated using least-squares type of algorithms. The GMP models have been extensively used in power amplifier digital pre-distortion (DPD) domain and show best performance and complexity balance [MSD+21].

$$y(n) = \sum_{k=0}^{K_a} \sum_{l=0}^{L_a} a_{kl}\, x(n-l) \cdot |x(n-l)|^k + \sum_{k=1}^{K_b} \sum_{l=0}^{L_b-1} \sum_{m=1}^{M_b} b_{klm}\, x(n-l) \cdot |x(n-l-m)|^k + \sum_{k=1}^{K_c} \sum_{l=0}^{L_c-1} \sum_{m=1}^{M_c} c_{klm}\, x(n-l) \cdot |x(n-l+m)|^k \quad (2\text{-}5)$$

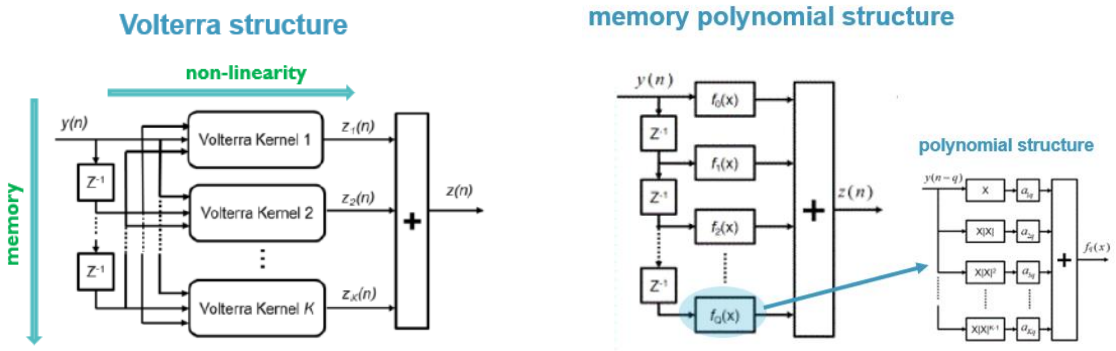


Figure 2-15: Volterra and simplified memory polynomial model.

2.2.2.4 Artificial neural network models

Alternatively, Artificial Neural networks (ANNs), including multi-layer perception, radial basis function based neural networks and recurrent neural networks (RNNs), have attracted increasing attention in behavioural modelling of PAs due to their ability to learn any arbitrary nonlinear function according to the universal approximation theorem.

The ANN models are able to model both short-term and long-term memory effects [YSW16] [CAA+18]. Figure 2-16 shows the block diagram of RVTDNN (Real-Valued Time-Delay Neural Network) [LBG04] [LYZ+08] and ARVTDNN (Augmented Real-Valued Time-Delay Neural Network) [WAH+19] models which are proven to model wideband PAs with memory effects.

Different types of ANN models for DPD have been explored recently: Convolutional Neural Networks (CNN) [HLY+21], Recurrent Neural Networks (RNN) [ZYZ+13], Long Short-Term Memory (LSTM) [WSL+22], which are suitable for PA modelling.

In contrast with the intrinsically local approximating properties of Volterra-based models, ANNs behave as global approximates, which is an important advantage when one is modelling strongly nonlinear systems. Also, since the sigmoidal functions used in ANNs are bounded in output amplitude, ANNs are, in principle, better than polynomials at extrapolating beyond the zone where the system was operated during parameter extraction. Moreover, the ANN models are able to characterize non-linearities not only from PA but also from DC offset and IQ imbalance [WAH+19] [TJD+19].

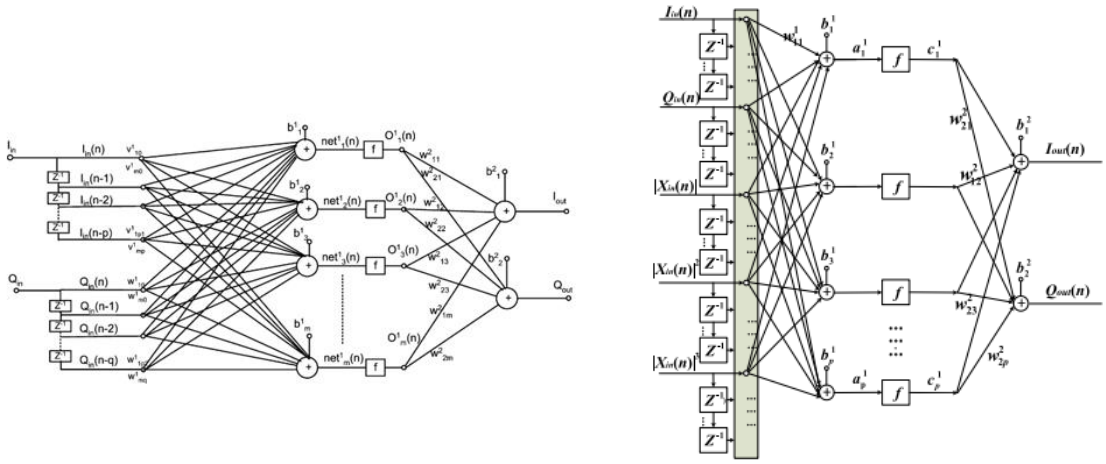


Figure 2-16: Block diagram of new two-layer RVTDNN and ARVTDNN PA behavioural models.

2.2.2.5 Model extension for sub-THz communication

Moving toward sub-THz communication, the power amplifier operates at high frequency (above 100 GHz) and typically operates over a wider bandwidth. The conventional CMOS power amplifiers do not easily provide large output power at those frequencies. Compound semi-conductors such as gallium nitride (GaN) or indium phosphide (InP) outperform CMOS power amplifiers at high operation frequency, providing higher required output power and efficiency [IME+23] [WWH+23]. However, it is well-known that GaN devices inherently suffer from low-frequency dispersive phenomena due to self-heating and charge trapping, which requires additional low frequency filter feedback loop for Volterra-based models and the long-term memory part is also requiring an ANN model, where LSTM ANN model is a good example [CAA+18]. Besides that, generally the wider bandwidth brings more frequency selective effects, for which models with memory effects are also required. The impacts of those short-term and long-term memory effects need to be simulated.

Given the limited output power of sub-THz transmitters and the low aperture of individual antennas at those frequencies, sub-THz systems demand a very high antenna gain to compensate for the link budget. Therefore, most experiments use highly directional antennas (e.g., horn antenna, parabolic antenna) at both Tx and Rx sides. Alternatively, massive antenna arrays can be utilized to steer the beam direction with sufficiently high array gain. For antenna array systems, the different PAs may simply be modelled as identical or include some variations of their main parameters (e.g., Psat, Pout), incorporated as random fluctuations into the model. In large arrays, coupling effects between the antennas can generate load pull impacting the PA operation. Such effects are not considered and are beyond the scope of mathematical models. They would need to be tested from real measurements in order to validate the accuracy of models.

2.2.3 Phase noise

2.2.3.1 Introduction and cause of phase noise

For an ideal oscillator where the whole power is concentrated at the central frequency, the power spectral density is a Dirac delta function. It can be described as $V(t) = A \cdot \cos(wt + \varphi)$. However, in the reality, the output is described as $V(t) = A(t) \cdot \cos\big(wt + \varphi(t)\big)$, which has both amplitude and phase offset variations as function of time in an unpredictable fashion [HL98].

Phase noise is described in frequency domain through its Power Spectral Density (PSD) in dBc/Hz, as illustrated in Figure 2-17.

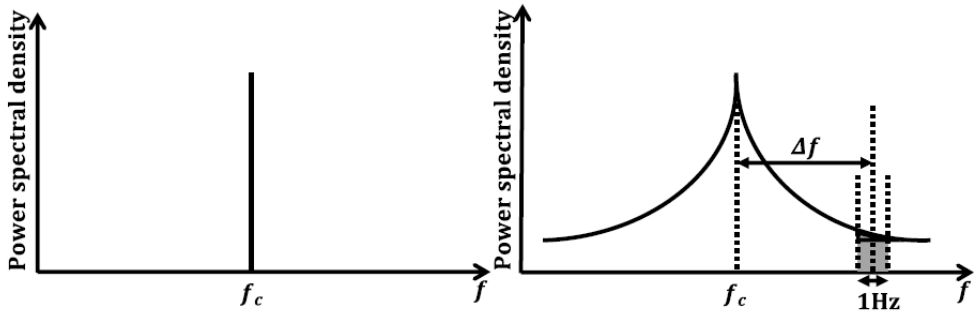


Figure 2-17: Spectral spreading due to phase noise (right) as compared to ideal oscillator (left).

Phase noise is mostly characterized in terms of single side baseband (SSB) of the PSD, where $P_{SSB}(f_c + f)$ is the single-sideband power of the oscillator within 1 Hz bandwidth around the offset frequency f from the central frequency f_c, and P_{total} is the total power of this oscillator. In most of the cases the effects of the phase variation are much larger and important than the effects of amplitude variation. Hence, in most application (2-6) includes the phase part only.

$$S(f) = \frac{P_{SSB}(f_c + f)}{P_{total}} \quad (2\text{-}6)$$

When targeting a high central frequency, the most economical method is to use a PLL at low frequency f_{basis} ,which has high quality factor, and further add additional multipliers circuits to boost the frequency to f_{target} [HEX20+D22]. Then the PSD of the PLL at the target frequency will be lifted up linearly by a factor of $20 \cdot \log(f_{target}/f_{basis})$.

$$S(f_{target}) = S(f) \cdot (f_{target}/f_{basis})^2 \quad (2\text{-}7)$$

Phase noise will cause spectral regrowth and decrease sensitivity or selectivity when the oscillator is used for mixing. Phase noise also introduces the phase rotation, ISI, and ICI that degrade the communication system's performance. Moreover, phase noise is a challenge for multicarrier signals such as OFDM because it destroys the orthogonality among subcarriers.

For multi-antenna systems using MIMO or beamforming, it is essential to characterize how correlated the phase noise is over the different antennas. In case of fully correlated phase noise, it will simply be seen as a common term to be simulated and compensated in the same way as in SISO systems. However, in case of non-correlated phase noise, it can destroy beamforming patterns, but also benefit from some kind of averaging effects over the different antennas. Non-correlated phase noise can also be more difficult to track in MIMO systems.

2.2.3.2 Phase noise model for free running oscillators and PLL

The oscillator phase noise originates from the noise inside the circuit, which can be categorized as white noise and coloured noise. The thermal noise and shot noise inside the devices are modelled as white. The substrate and supply-noise sources and low-frequency noise are modelled as coloured. The most significant part of the coloured noise inside the circuit can be modelled as flicker noise [Dem06]. The PSD of phase noise of a free running oscillator is modelled as a superposition of three independent processes as (2-8) [KKP+14]. The $\emptyset_3(t)$ and $\emptyset_2(t)$ originate from the integration of the flicker noise and white noise, with -30 and -20 dB/decade slope, respectively. The $\emptyset_0(t)$ is the noise floor which originates from the white noise with attenuation or amplification. A typical phase noise PSD for a free running oscillator is shown in Figure 2-18.a.

The phase lock loop (PLL) architecture is widely used in frequency synchronization to ensure a stabilized output, which includes a free running oscillator, a reference oscillator, a loop filter, a phase-frequency detector, and a frequency divider. All the elements contribute to the output of the phase noise of the PLL. The PLL has

the property of high-pass filtering the phase noise of the reference oscillators, which attenuates the oscillator's phase noise below a cutoff frequency. The cut-off frequency is determined by the bandwidth of the PLL architecture.

$$\emptyset(\mathbf{t}) = \emptyset_{\mathbf{3}}(\mathbf{t}) + \emptyset_{\mathbf{2}}(\mathbf{t}) + \emptyset_{\mathbf{0}}(\mathbf{t}) \quad (2\text{-}8)$$

$$S_{\emptyset_3}(f) = \frac{K_3}{f^3},\ S_{\emptyset_2}(f) = \frac{K_2}{f^2},\ S_{\emptyset_0}(f) = K_0 \quad (2\text{-}9)$$

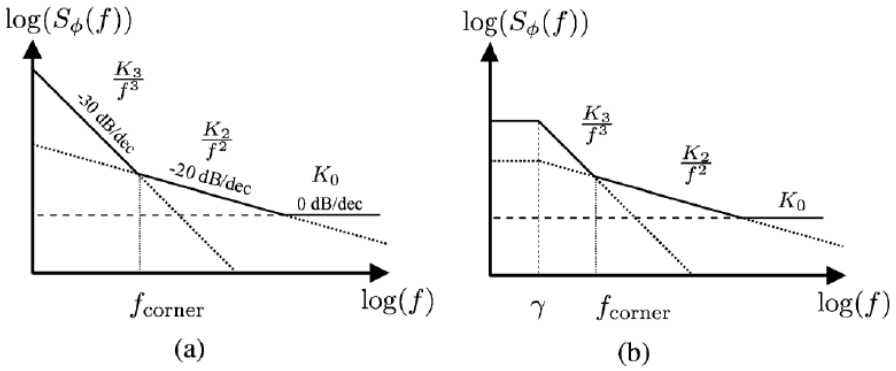


Figure 2-18: Phase noise PSD of a typical oscillator (a) PSD for free running oscillator (b) PSD for PLL.

2.2.3.3 SZSP (single zero/pole) and MZMP (multi zero/pole) models

The 802.11ad task group recommends the single zero and single pole phase noise model for millimetre-wave wireless systems at 60 GHz [IEEE06]. In order to provide more precision and better represent the PSD, this model is extended to multi zero and multi pole model as shown in (2-10). $S(f_0)$ is the phase noise level that is determined by the loop filter at low frequency, and f_{z_n} and f_{p_n} are the zero frequencies and pole frequencies, respectively. For 802.11ad, the recommended values of the above parameters are: $N = 1$, $S(f_0) = -90$ dBc/Hz, $f_z = 100$ MHz, $f_p = 1$ MHz and $S(f_\infty) = -130$ dBc/Hz.

$$\boldsymbol{S(f) = S(f_0) \cdot \prod_{n=1}^{N} \frac{1+\left(f/f_{z_n}\right)^2}{1+\left(f/f_{p_n}\right)^2}} \quad (2\text{-}10)$$

2.2.3.4 Piecewise linear frequency mask model

Differently from the analytical models, the piecewise linear frequency mask model uses the PSD data from real measurements. A typical shape of PSD is shown in Figure 2-18.b. The PSD measurement reports typically include the PSD at low frequency, 1 MHz, 10 MHz and 100 MHz. A piecewise linear connection of those data points form the approximate shape of the PSD of the PLL under test. Figure 2-19 provides a general shape definition of the frequency mask. The shape is defined by 3 PSD level parameters: P_1 , P_2 and P_3 , and 3 frequency parameters: F_{c_1} , F_{PLLBW} and F_{c_2}. Unlike the theory where the phase noise is almost constant within the PLL band, in the measurements there can be some variations modelled as a small slope. The P_1 and F_{c_1} parameters describe this slope at low frequency. The P_2 and F_{PLLBW} parameters describe the locked phase noise level within the bandwidth of PLL. After a slope of -20 dBc/decade, the P_3 and F_{c_2} values describe the white noise floor. The phase noise within the bandwidth of the tracking loop $f_{tracking}$ in baseband signal processing can be compensated, while the phase noise at frequency above $f_{tracking}$ cannot and it will degrade the communication system.

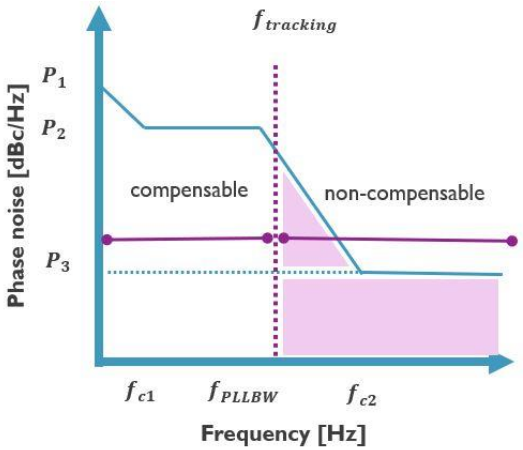


Figure 2-19: General shape of phase noise frequency mask.

2.2.3.5 Update of frequency mask based on the SOTA PLLs

In recent years, there has been many publications on PLL with frequency around D-Band [PCY+22], [BJK+23], [MLC+23], [BKJ+23], [AAK+21], [LL20], [CAL+22], [LIA+23] in CMOS or BiCMOS technology. The architectures fall into two categories: direct and indirect architectures. Direct architectures directly generate the target carrier frequency, while indirect architectures generate a lower frequency, which is later multiplied via multiplier circuits in order to generate the target carrier frequency. The direct architectures typically use cascaded sub-sampling or power-gating injection-locked frequency multiplier technology to create near D-Band frequency directly at the output of PLL. Figure 2-20 lists the measurement PSD values in recent publications at 1 MHz, 10 MHz and 100 MHz. We scale the DSP values at different PLL frequencies using (2-7) to 140 GHz and the results are shown in the right part of Figure 2-20. By averaging the SOTA publication data scaled at 140 GHz, we propose the following frequency mask of phase noise: -103 dBc/Hz at 1 MHz, -111 dBc/Hz at 10 MHz and -130 dBc/Hz at 100 MHz. The frequency mask at low frequency is extended with a slope of -20 dB/decade.

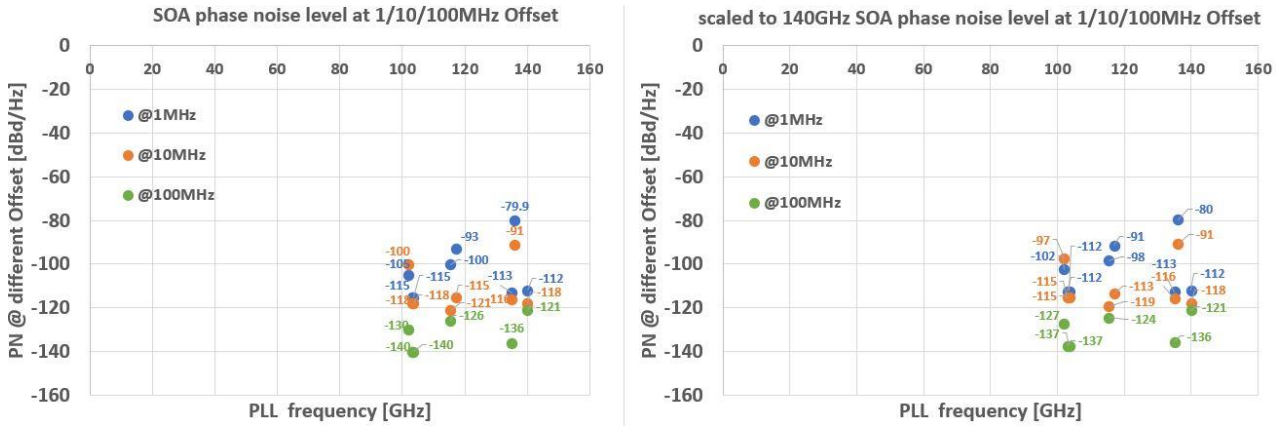


Figure 2-20: PSD measurement values from recent publications: (left) frequency not scaled, (right) scaled to 140 GHz.

Figure 2-21 shows the frequency mask at 140 GHz for [TZP+20], [IEEE+06], [BJK+23] and the proposed one above, respectively. The orange curve [TZP+20] is based on a 24 GHz PLL and assumes multipliers to scale up the frequency to 140 GHz. The blue curve [BJK+23] is based on a PSD lab measurement, from a design which provides very low phase noise, however with high complexity and large chip area. The black curve [IEEE+06] is based on phase noise model recommended for 802.11ad and scaled up to 140 GHz.

In order to compare their impact on communication systems, we use the integrated phase noise from $f_{tracking}$ to the signal bandwidth as the metric, which is the non-compensable phase noise impacting the system. Figure 2-22 shows the non-compensable integrated phase noise when the signal bandwidth is 5 GHz and $f_{tracking}$ changes, as well as the non-compensable integrated phase noise when $f_{tracking}$ is fixed at 1 MHz and signal bandwidth changes. The integrated phase noise is equal to the EVM degradation purely contributed from phase noise. To support 64-QAM, a -30 dB EVM is required based on a study at 100 GHz [BKJ+23]. From Figure 2-22, the PSD frequency mask based on 24 GHz and 60 GHz are not able to ensure 64-QAM communications even with a phase tracking loop of 10 MHz bandwidth. This means either improved PLL designs or channel coding to relax the spec are necessary.

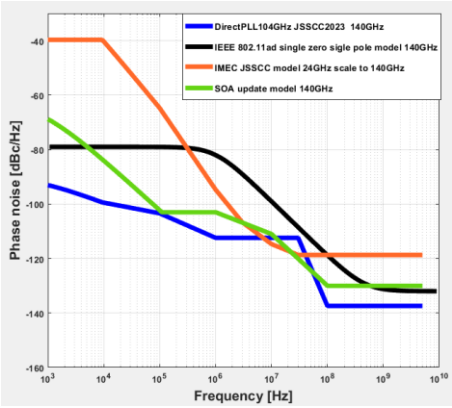


Figure 2-21: Comparison of SOA D-band phase noise frequency spectrum.

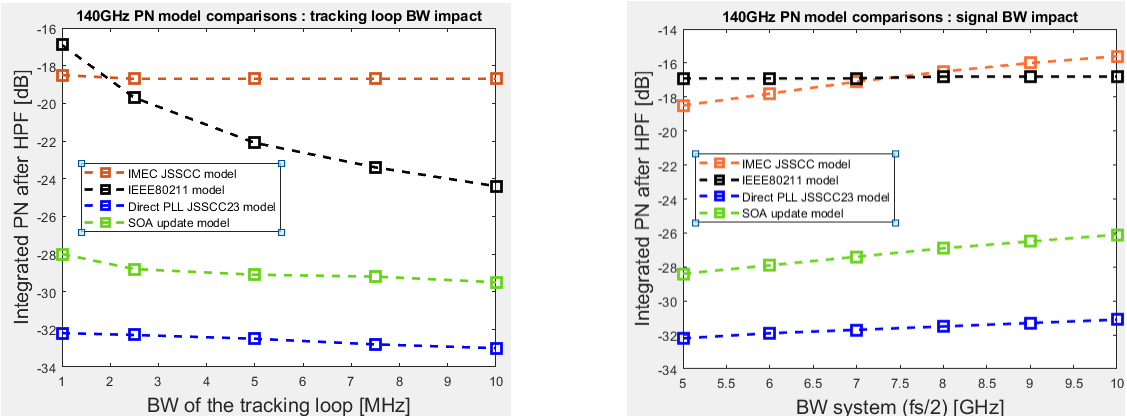


Figure 2-22: The impact of tracking loop and signal bandwidth on SOTA PN models: (left) bandwidth of the baseband tracking loop effect (signal bandwidth fixed to 5 GHz), (right) signal bandwidth effect (tracking loop bandwidth fixed to 1 MHz).

2.2.3.6 Model extension

6G connectivity will be able to operate in a wide range of frequency bands. This includes current frequencies in the sub-1 GHz range, 3.5-6 GHs mid-band range and millimeter-wave range, as well as new frequencies in the 7-15 GHz centimetric range and 90-300 GHz sub-THz range. The sub-THz systems are critically impacted by the strong phase noise of the frequency synthesizers at high frequency. In general, the PSD of phase noise will be lifted up by a factor of $20log_{10}(f_{target}/f_{basis})$ dB from f_{basis} to f_{target} [HPK+22]. The integrated phase noise that cannot be compensated will have an increased impact by the up-scaled phase noise floor and wider bandwidth, hence an accurate model of the phase noise floor is becoming critical for wider bandwidth.

The MIMO and phased array systems at mm-Wave or D-band may have multiple LOs instead of single LO. Figure 2-23 illustrates 4 different options of LO distribution architectures for multi-antenna systems. Different phased arrays can share the same high frequency LO or use independent high frequency LOs. If low frequency LO is used, then the different phased arrays share common low frequency LO and use independent multipliers or use independent low frequency LO plus independent multipliers. Among these options, the last option has lower RF circuit cost, and it maintains some correlation characteristic of the phase noise between different paths, which is helpful in MIMO detection. In such a case, the phase noise for different paths can be modelled as (2-11), where N is the multiplier factor and $n^{(m)}$ is the independent noise from different multipliers in different paths m. $n^{(m)}$ is the white noise with power at the noise floor of PSD of the common low frequency LO.

$$PN^{(m)} = PN_{LO} + 10\,log(N)^2 + n^{(m)} \quad (2\text{-}11)$$

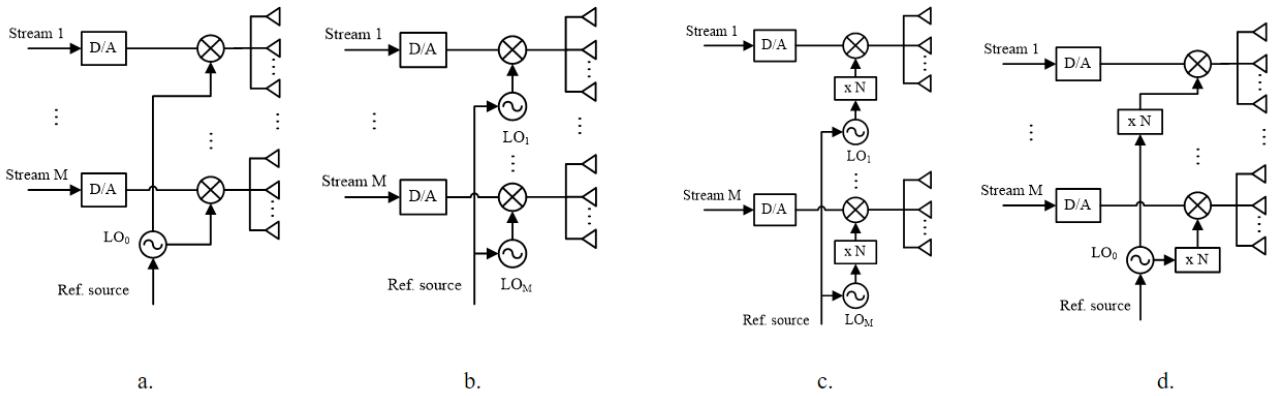


Figure 2-23: Different options for PN models for multi-antenna system.

2.2.4 I/Q imbalance

2.2.4.1 Introduction and cause of I/Q imbalance

Conversion with quadrature local oscillators in the analogue domain is often used in the transmitter and/or receiver. This convenient implementation of quadrature conversion is affected by phase and amplitude offsets in the two branches, which disturbs the perfect orthogonality, a problem referred to as I/Q imbalance. This imbalance can happen at the transmitter, receiver, or both [HB08]. The I/Q imbalance causes the spectral content of the upper sideband to mix with the lower sideband and vice versa, which is referred to as mirror interference (see Figure 2-24). I/Q imbalance in multicarrier systems introduces ICI terms on the mirror subcarriers. The harmful mirror interference degrades the system performance and can also impact channel estimation.

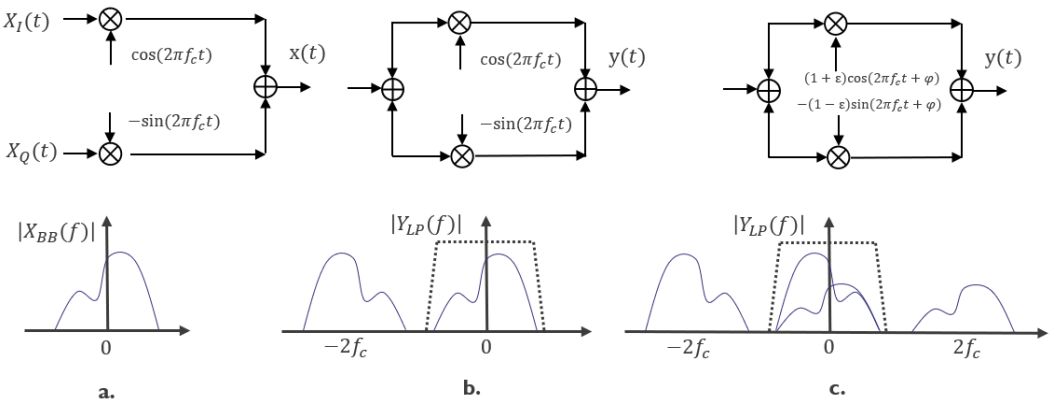


Figure 2-24: Non-frequency selective I/Q imbalance model and its impact in frequency.

2.2.4.2 Non frequency selective I/Q imbalance model

The frequency non-selective I/Q imbalance model introduces the amplitude mismatch ε and phase mismatch $\Delta\varphi$, where the amplitudes of I and Q branches are represented as $(1+\varepsilon)$ and $(1-\varepsilon)$ and the phases of I and Q branches as $+\Delta\varphi$ and $-\Delta\varphi$. The mismatch can happen at both transmitter and receiver side, as illustrated in Figure 2-25.

In Zero-IF front-end architectures, if we assume an ideal transmitter has been used to up-convert the complex baseband signal $x_L(t) = x_I(t) + x_Q(t)$, it would result in an ideal RF signal $x_{RF}(t)$. If the I/Q imbalance is introduced at the transmitter, the transmitted signal is represented as:

$$x_{RF}(t) = (1+\varepsilon_T)\cdot\cos(2\pi f_c t + \Delta\varphi_T)\cdot x_I(t) - j\cdot(1-\varepsilon_T)\cdot\sin(2\pi f_c t - \Delta\varphi_T)\cdot x_Q(t)$$

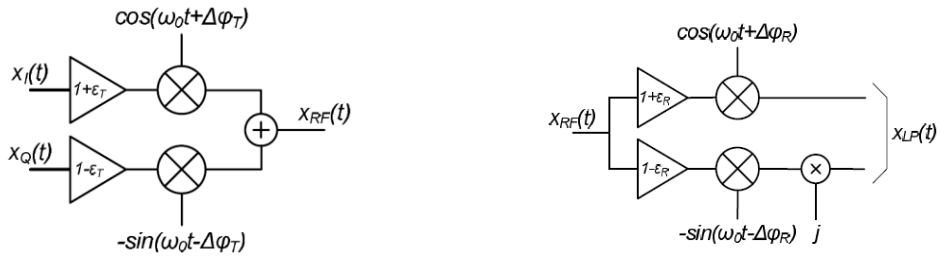
With an ideal receiver, the RF signal is down converted using an ideal LO signal $x_{LO}(t) = \exp(-j2\pi f_c t)$. The lowpass signal $x_{LP}(t)$ is modelled in (2-12). It consists of the ideal signal $x_L(t)$ plus a scaled version of its complex conjugate $x_L^*(t)$. Because of the complex conjugation, the positive part of the spectrum of $x_L(t)$ interferes with the negative part of its spectrum and vice versa.

$$\begin{aligned} x_{LP}(t) &= x_{RF}(t)\cdot x_{LO}(t) \\ &= \alpha_T \cdot x_L(t) + \beta_T \cdot x_L^*(t) \\ \alpha_T &= \cos(\Delta\varphi_T) + j\cdot\varepsilon_T\cdot\sin(\Delta\varphi_T) \\ \beta_T &= \varepsilon_T\cdot\cos(\Delta\varphi_T) + j\cdot\sin(\Delta\varphi_T) \end{aligned} \qquad (2\text{-}12)$$

If the I/Q imbalance is only present at the receiver, then the distorted signal after down-conversion mixing and low pass filtering to remove the mirror band is modelled as:

$$\begin{aligned} x_{LP}(t) &= x_{RF}(t)\cdot x_{LO}(t) \\ &= \alpha_R.x_L(t) + \beta_R\cdot x_L^*(t) \\ \alpha_R &= \cos(\Delta\varphi_R) - j\cdot\varepsilon_R\cdot\sin(\Delta\varphi_R) \\ \beta_R &= \varepsilon_R\cdot\cos(\Delta\varphi_R) + j\cdot\sin(\Delta\varphi_R) \end{aligned} \qquad (2\text{-}13)$$

The lowpass signal $x_{LP}(t)$ consists of the ideal signal $x_L(t)$ plus a scaled version of its complex conjugate $x_L^*(t)$. Because of the complex conjugation, the positive part of the spectrum of $x_L(t)$ interferes with the negative part of its spectrum and vice versa.



(a) Transimitter side (b) Receiver side

Figure 2-25: Frequency non-selective I/Q imbalance model.

2.2.4.3 Frequency selective model

The analogue lowpass filters in the I and Q branches may not be identical and have different gain, cutoff frequency, and group delay. With the increasing bandwidth of emerging air interfaces, the analogue lowpass filters in the I and Q branches can be very broadband (from several MHz up to tens of GHz) and it becomes increasingly difficult to match them. This results in the I/Q imbalance becoming more frequency selective. The frequency selective imbalance varies over the bandwidth and is caused by unbalanced frequency responses of the I and Q branches. Frequency-selective effects may also come from other components having non-ideal wideband response, such as mixers, I/Q LO generation and distribution, interconnect lines, etc.

The frequency selective model is based on the above non-frequency-selective model and appended with two frequency dependent filters modelling HW impairments of the filters in the I and Q paths, as shown in Figure 2-26.

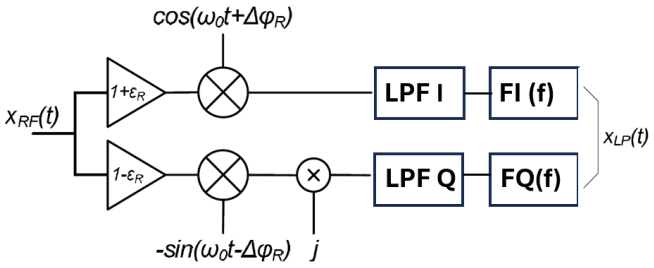


Figure 2-26: Frequency selective I/Q imbalance model, including frequency-dependent filters FI(f) and FQ(f).

2.2.4.4 Model extension for sub-THz communication

In sub-THz bands, I/Q imbalance may be more severe than that in lower frequency bands since making efficient and accurate components is more challenging than at lower frequencies. Moreover, the bandwidth is much larger than for lower frequency bands. In broadband communication, the amplitude and phase errors are variable as a function of the conversion frequency and the I/Q output bandwidth, in which the frequency selective effects need to be considered [SW21]. There are few publications [DHC+17] [SW21] in the literature that provide the properties of I/Q frequency selective filter models for wideband system. Hence although the model of section 2.2.4.3 is relevant, further study on how to model the frequency selective part is essential.

For MIMO and phased array systems, the IQ imbalance model can equal for each down-conversion path at RX and as equal for each up-conversion path at TX. Slight random differences between each path can be added to model the non-deterministic character of I/Q imbalance.

Additionally, some on-chip calibration circuits may be foreseen, especially when the expected I/Q imbalance to be too high for the target specs. The net effect is that only the residual (non-calibrated) I/Q imbalance will be visible.

2.2.5 Carrier frequency offset

2.2.5.1 Introduction and cause of CFO

The carrier frequency offset (CFO) occurs when the down-converting local oscillator at the receiver does not perfectly synchronize with the received signal's carrier. Local oscillator mismatch between TX and RX and Doppler shifts are two causes of the frequency mismatch. Local oscillator mismatch occurs because of the local oscillators' different physical properties and errors, such that they can't oscillate exactly at the desired frequency. The precision of the crystals is required to be better than +/- 20 parts per million (ppm) in IEEE 802.11a WLAN standard [IEEE99]. If the carrier frequency is 5 GHz, a difference of $2 \times 5 \times 10^9 \times 20 \times 10^{-6} = 200$ kHz may exist (assuming Tx and Rx oscillators both diverge from 20 ppm but in the opposite direction). The Doppler shift occurs when the relative motion between the transmitter and the receiver shifts the carrier frequency seen by the receiver. Although not negligible, CFO due to Doppler shifts is usually lower than the CFO due to crystal inaccuracies. CFO causes a rotation of transmit symbols and thus introduces inter symbol interference (ISI) and inter-carrier interference (ICI) in single-carrier (SC) and multicarrier systems.

2.2.5.2 State-of-the-art models of CFO

Based on the definition of carrier frequency offset, the model of CFO is straightforward and just applies a frequency shift Δf to received signal: $y(t) = x(t) \cdot e^{j\Delta f t}$. The successive samples will suffer an accumulative phase shift. For OFDM based system, normally the Δf is divided into two parts: integral and fractional of the frequency space between each subcarrier.

2.2.5.3 Model extension for sub-THz communication

CFO in absolute frequency becomes more severe when the carrier frequency increases, assuming the same relative CFO (in ppm). However, assuming the bandwidth also increases, which is the main motivation to move to sub-THz frequencies, the relative impact of CFO should not be fundamentally different than at lower frequencies. Hence no further model extension is foreseen when the carrier frequency or the bandwidth is increased towards to sub-THz communications.

2.2.6 Sampling clock offset

2.2.6.1 Introduction and cause of SCO

Sampling clock offset (SCO) occurs when the frequency inaccuracy of oscillators causes clocks of transmitter and receiver to drift with respect to each other. The SCO results in several problems in baseband signal processing, illustrated in Figure 2-27. First, because of the increasing sampling time error, the signal is no longer sampled at the optimum point in the eye diagram and degradation occurs. In modern systems (where there is not necessarily an eye diagram as in old single-carrier LOS links), the degradation will come due to a mismatch between the expected sampling moment (used in channel estimation and equalizer computation) and the actual sampling moment. Secondly, after certain time, the receiver will have a sampling time shifted by one complete sample. SCO is of particular concern for orthogonal frequency division multiplexing (OFDM) communication systems as clock drift can compromise the mutual orthogonality of the subcarriers, degrading performance.

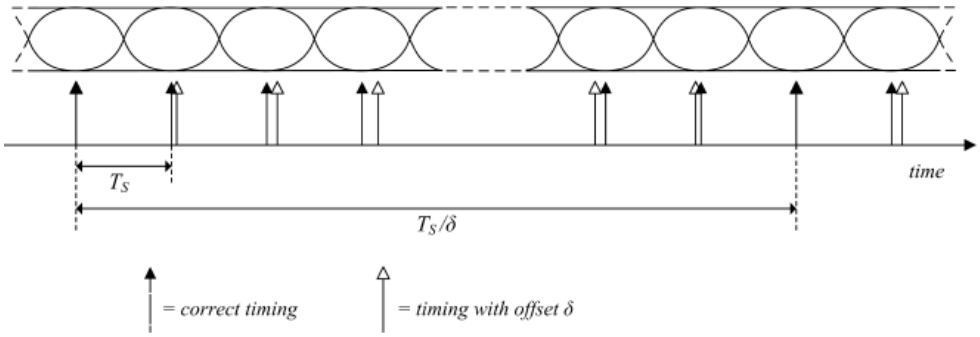


Figure 2-27: Sample point drift due to SCO and sampling position with respect to received symbols.

2.2.6.2 State-of-the-art models of SCO

Basically, if the sampling frequency $f_s = 1/T_s$ has a relative offset factor δ with respect to the ideal sampling frequency, the sampling points become $x[n(T_s + \delta T_s)]$ instead of $x[nT_s]$, which incurring error at each sample and the sample will drift to another correct sample time after T_s/δ samples.

To model this hardware impairment, the value sampled at $x[n(T_s + \delta T_s)]$ can be interpolated by Lagrange interpolation.

2.2.6.3 Model extension for sub-THz communications

Similar as CFO, no further model extension is foreseen when the carrier frequency or the bandwidth is increased towards to sub-THz communications, as the effect can be modelled in the same way based on transmit/receive frequency offset.

2.2.7 ADC/DAC related non-idealities

2.2.7.1 Introduction

The ADC and DAC converters are essential components at the bridge between the digital domain and analogue domain. Modelling them is essential for digital simulations, not only for non-idealities caused by those analogue components, but also in line with the resolution (quantization) for the digital part (see Figure 2-28).

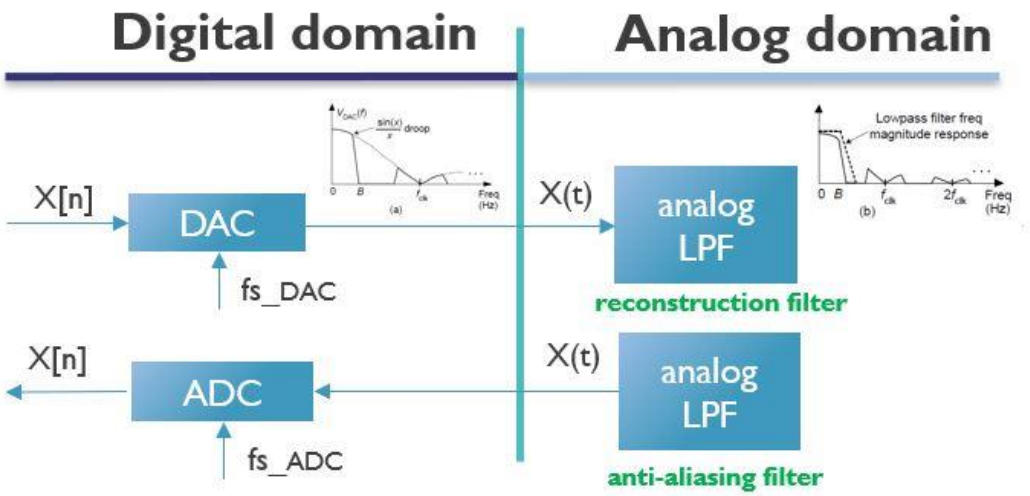


Figure 2-28: ADC and DAC basics.

The impairments caused from ADC/DAC mainly fall into the 3 categories: sampling and filtering, quantization and non-idealities.

The periodical Dirac delta pulse sampling in time domain creates spectral replicas in frequency domain. A real DAC samples the data and holds the data for a duration $\frac{1}{f_s}$ in the form of zero-order holding. The form of zero-order holding in time domain creates a signal shaping as $\frac{\sin(x)}{x}$ in frequency domain and leads to energy leaking, which is illustrated in Figure 2-28. A reconstruction filter in analogue domain is placed after the DAC to remove the replicas in frequency domain. In order to reduce the energy emitted to out of band and ease the complexity of the reconstruction filter, oversampling is used in practical implementations, leading to frequency-domain replicas further away from the desired signal. At the receiver side, an anti-aliasing analogue

filter is used before sampling to restrict the bandwidth of the signal in order to satisfy the Nyquist–Shannon sampling theorem over the band of interest while suppressing out-of-band noise and interferers.

The ADC model includes quantization, where the sampled analogue values are mapped to discrete digital values. The number of bits for quantization determines the quantization noise level and the resolution of the ADC. The ADC model may also include various non-idealities such as quantization noise, linearity errors, distortion, gain, offset errors, and other imperfections that real ADCs exhibit.

2.2.7.2 State-of-the-art models of ADC/DAC

Figure 2-29 illustrates the basic block diagram of DAC and ADC models in digital simulation chain. At the transmitter side, the signal is over-sampled and passes through a low pass filter. The up-sampling helps to properly capture the non-idealities of the signal in simulations, while in the real system over-sampling eases the design of low pass filter. The digital low pass filter follows the up-sampling process to mimic the reconstruction filter. A quantization model is required before the up-sampling process to generate a digital signal with $2^{N_{bit}}$ different quantization levels. At the receiver side (for ADC modelling), a digital low pass filter mimics the anti-aliasing filter and is followed by down-sampling. Then the signals are quantized based on clipping level and number of quantization bits.

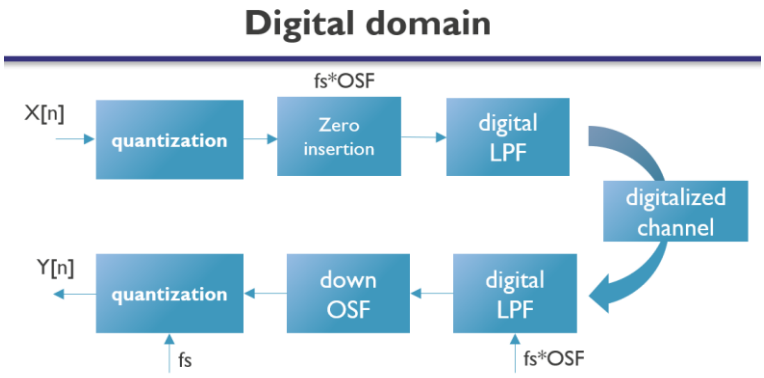


Figure 2-29: ADC and DAC high level simulation block diagram.

2.2.7.3 Model extension for sub-THz communications

As DAC/ADC operate at baseband, no specific model extensions are foreseen with respect to carrier frequency. However, the bandwidth increase may have some impact on the non-idealities. Different operating points may also be selected, e.g., a reduced oversampling due to the already wide bandwidth or a reduced resolution in order to reduce the digital complexity and related power consumption of wideband DSP.

2.2.8 DC offset

2.2.8.1 Introduction and cause of DC offset

The receiver with zero-IF front-end will not only suffer from IQ imbalance when down-converting the signal via mixing, but also from DC offset. Due to capacitive and substrate coupling, there will be imperfect isolation between ports of the mixer, and a small amount of signal from the LO port will feed through to the input port of the LNA and mixer. It can be seen as strong interferer and may leak through the mixer ports, appearing as an additive signal component at the mixer LO port and mixing with itself, leading to a phenomenon known as “Self-Mixing”. The DC offset limits the dynamic range of the receiver and interferes with the received signals in the middle of the band (close to DC). The degradation of the spectrum around the centre of the processed band caused by the presence of a DC component, can be solved by an AC coupling techniques implemented between the mixer’s output and the subsequent blocks of I/Q path leading to ADC [RK20]. However, this also attenuates the useful signal close to DC, which is why OFDM systems typically do not load the central subcarrier(s). Two types of DC offset are illustrated in Figure 2-30 [HB08].

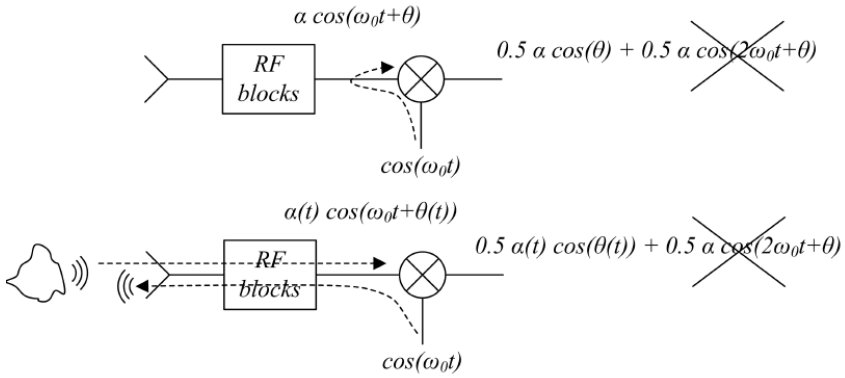


Figure 2-30: Static and dynamic DC offset in zero-IF architecture.

2.2.8.2 State-of-the-art models of DC offset

The model of DC offset is simply defined as static DC offset d or dynamic DC offset $d(t)$. This DC component is added separately for I and Q components after frequency down-conversion [ZHC+14] [LZH+15] [ZTY+22] [SJL02] [NS22], which is illustrated in Figure 2-31.

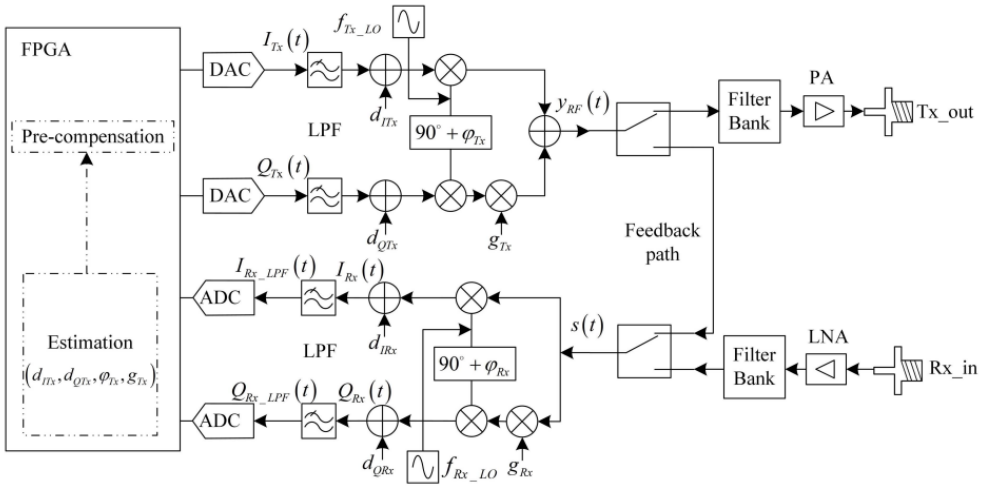


Figure 2-31: Zero-IF transmitter and receiver with DC offset and IQ imbalance.

2.2.8.3 Model extension for sub-THz communications

No further model extension is foreseen when the carrier frequency or the bandwidth is increased towards to sub-THz communications.

2.2.9 Phase shift error

2.2.9.1 Introduction and cause of phase shift error

With higher frequencies, communication systems rely more and more on large antenna arrays and analogue beamforming. This is generally implemented by using phase shifters in the different antenna chains as the ideal option - using delay lines - is impractical for implementation. Phase shifters can be implemented from different circuits and can be placed in order to affect the baseband signals, the LO signals or the RF signals. Depending on the implementation, various imperfections can be present.

2.2.9.2 State-of-the-art models of phase shift error

Similar to DAC and ADC components, phase shifters typically operate with quantized steps. Hence, a finite resolution is considered, typically using unit steps of 360° divided by a power of 2, depending on the quantization level. This is at least the case for polar implementations. Other phase shifting options rely on I/Q

cartesian implementations, where phase-shifted combinations on the unit circle are generated by combining the proper I and Q quantized signals.

Additionally to the finite resolution, phase shifter is not ideal and can generate a phase value different from the target. This can be modeled by adding a Gaussian random contribution on top of the target phase. The amount of residual non-ideality to consider can depend on the accuracy of chip calibration techniques.

2.2.9.3 Model extension for sub-THz communications

Given the growing sensitivity of signal phase to mismatches or line propagation delays when increasing the carrier frequency, we can expect to use the same models bit with increased amount of non-ideality when the carrier frequency is increased towards to sub-THz communications. Large bandwidth increase may also lead to frequency-selective inaccuracies.

2.2.10 Conclusions towards sub-THz architectures

As compared to communication systems at lower frequencies, sub-THz transceiver differ at least in the following elements:

- A higher carrier frequency
- A wider bandwidth
- A larger number of antennas
- Specific architectures, e.g., using more analogue/hybrid beamforming rather than fully-digital MIMO

In this section we have reviewed models for PA non-linearity, phase noise, I/Q imbalance, CFO, SCO, ADC/DAC, DC offset and phase shifters.

PAs are expected to move from memoryless models to models with memory. There are two reasons associated to that. First, short-term memory effects (ns time scale) become more visible due to the wider bandwidth. Second, the need for specific III/V semiconductors such as GaN or InP leads to specific device effects such as charge trapping, which creates long-term memory effects (ms time scale).

Phase noise scales quadratically with the increased carrier frequency, hence it becomes more and more problematic when moving to sub-THz bands. Additionally, the noise floor of the phase noise plays an increasingly important role due to the wider bandwidth on which it is integrated. Finally, LO distribution to wider antenna arrays creates potential decorrelated contributions over individual antennas.

I/Q imbalance can also become more problematic. First, increasing the carrier frequency makes the system more sensitive to mismatches in the design, as shorter wavelength translates into larger phase inaccuracies. Secondly, the increased bandwidth above 1 GHz is expected to create some frequency-selective behaviour of the I/Q imbalance.

ADC and DAC models are expected to evolve with the bandwidth. A reduced resolution and oversampling are expected to be used in order to keep the power consumption acceptable, especially for systems with a large number of digital chains. Specific non-idealities may also scale with the bandwidth: the clock jitter and thermal noise will bring more noise, reducing the effective number of bits. ADCs are expected to be more sensitive to design, due to the risk form out-of-band noise and interferers, and the lower signal levels as compared to DAC.

CFO, SCO and DC offset models are not expected to be significantly different for sub-THz systems.

Considering the higher number of antennas due to wider arrays, it is important to generalize models for SISO or few antenna systems towards this larger number. An important point is to check whether each individual antenna will suffer the same non-ideality or whether some randomness needs to be introduced over the different antennas. Specifically, the following antenna-specific elements could take place:

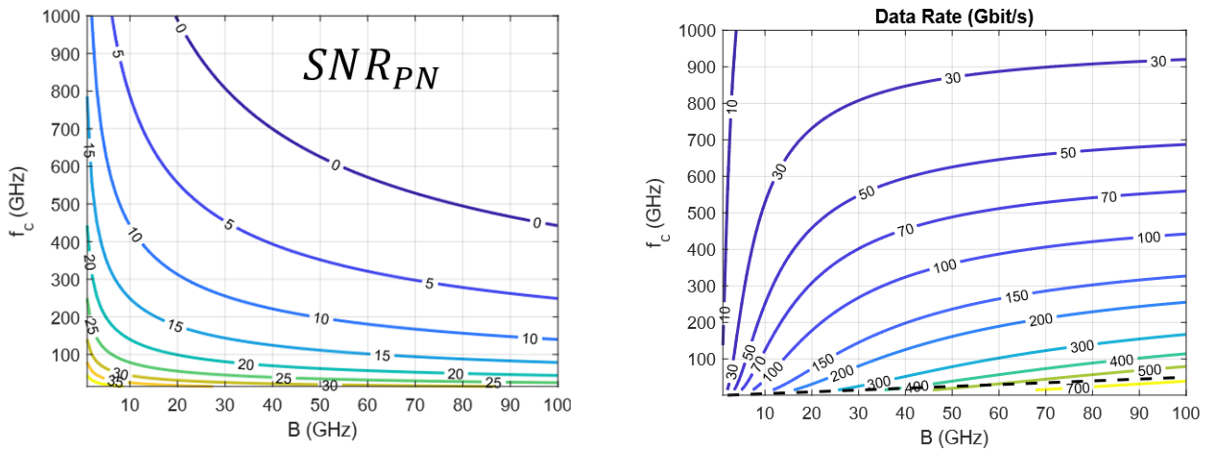
- Random fluctuations on PA output power and non-linear characteristic, e.g., +/- 1 dB
- LO distribution role in phase noise (see Section 2.3)
- Different I/Q imbalance values on different antenna chains

Future research in HW non-ideality modelling should focus on the following aspects:

- Decide on which memory PA model is most suited and parameterize it based on sub-THz PA designs.
- Agree on a phase noise power spectral density from averaging the main sources in state-of-the-art and consider improved PLL design options such that it can be kept acceptable.
- Check how far frequency-selective I/Q imbalance models need to be introduced and parameterize them.
- Further check ADC impairments for very wide bandwidth (above the basic quantization and clipping error).
- Introduce and parameterize phase shift errors based on analogue architecture design trade-offs

2.3 Wideband array phase noise analysis and role of LO routing

Wideband phase noise is identified as one of the key performance-limiting RF impairments in sub-THz systems with wide signal bandwidth [HEX23-D23, CHK+17, STP+23]. While the memory-dependent phase noise, i.e., slower phase variations, can be compensated by various compensation techniques [HEX21-D22], the flat region of the phase noise dominated by the noise floor of the phase locked loop (PLL) cannot be compensated by traditional techniques [STP+23]. Moreover, compared to thermal noise in the RF front end, phase noise cannot be typically compensated by beamforming and array gain. In [HEX23-D23] it has been observed that for a certain fixed-phase noise-limited SNR, phase noise limits the bandwidth at higher frequencies. Especially at frequencies of 150 GHz and above with even tens of GHz of signal bandwidth, phase noise may significantly limit the spectral efficiency and so the achievable data. To showcase this, the analysis performed in [HEX23-D23] for phase-noise-limited SNR is extended to an achievable data rate for a single link. The result is illustrated in Figure 2-32 (a) as phase noise limited SNR and in Figure 2-32 (b) as corresponding data rate. Each curve in the figure shows a different target possible bandwidth and centre frequency combination to achieve a certain fixed SNR and data rate, respectively. Note that this is the limitation only given by the phase noise with the assumption that all the memory-dependent phase noise is compensated ideally and only the noise floor of the LO is left. In the example, the noise floor of the PLL is -150 dBc/Hz at 15 GHz and follows model 2 of [HEX23-D23] for the centre-frequency dependency. To achieve 100 Gbit/s data rate, higher centre frequencies require potentially more bandwidth than the lower frequency systems. The results of the analysis depend highly on the level of the LO noise floor.



(a) (b)

Figure 2-32. (a) Different bandwidth-entre frequency combinations for a fixed phase-noise-limited SNR and (b) the corresponding phase-noise-limited data-rates.

2.3.1 Theoretical impact of asymmetrical LO routing for wideband phase noise

In mmW and above frequencies, larger antenna panels are typically divided into smaller subarrays, each performing analogueue beamforming using the phased array principle [ATS+21, ZGL+17]. Each subarray typically has a mixer driven by an LO signal for up and down conversion. Each of the mixer may have individual PLL. However, in many cases, the LO for more than one subarray is shared from a single PLL via dedicated LO routing [ALJ19, LPA+19] as shown in Figure 2-33. The benefit of shared LO is that the subarrays are phase-coherent constantly and can be used for creating more directive combined beams, or multiple wider beams simultaneously. For performing higher frequency LO signals, typically lower frequency signal is multiplied by several frequency multipliers [HEX23-D23, HEX21-D22]. Thus, the signal can be divided into subarrays also in different frequencies before or after, or in between the multipliers. Assuming that phase noise is dominated by the PLL output, the impact of multipliers is also to multiply the phase noise (increasing the level by 20 dB/decade) [DPS20].

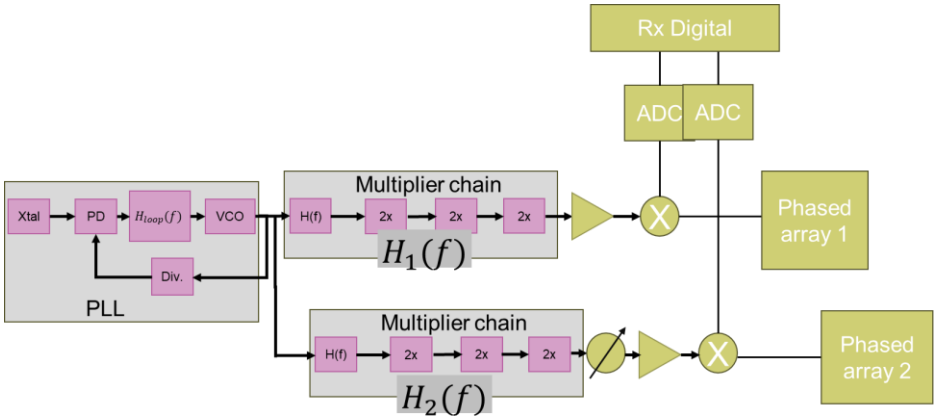


Figure 2-33: Description of the architecture to divide single LO signal to multiple mixers.

When multiple subarrays/mixers share the same PLL, it is often preferable to route the signals to subarrays in an asymmetric, meaning that the length of the signal path from PLL output to subarrays and their mixers is different. For example, the so-called BUS approach used for RF and IF signals in [SAH+23] can be used to enable highly reconfigurable and modular architecture also for LO sharing. This means that, in addition to the desired LO signal, also the phase noise, have delay differences between the paths. In the following, we study the impact of the routing delay for the combined phase noise over two mixer paths in receiver.

Consider a scenario depicted in Figure 2-33 with two sub-array-based receivers. For simplicity, we denote $H_1(f) = 1$ and $H_2(f) = \exp(-2\pi f\tau)$, meaning that the delay difference of the LO paths is τ. Let as denote time domain RF signal in both receiver paths (after phase shifting) as

$$x_{RF}(t) = sin\,(2\pi f_{RF} t) \tag{2-14}$$

where t is time and f_{RF} is RF frequency. Let us further write the LO signal of paths 1 and 2 as

$$x_{LO1} = \sin\big(2\pi f_{LO} t + \varphi(t)\big) \tag{2-15}$$

and

$$x_{LO2} = \sin(2\pi f_{LO} t - 2\pi f_{LO}\tau + \varphi(t-\tau) + \phi_{LO}) \tag{2-16}$$

where f_{LO} is the LO frequency, $\varphi(t)$ is the phase noise, and ϕ_{LO} is the phase value of the phase shifter in the LO path. Note that the phase shifter is shown in LO path for simplicity of the example, but the required phase shift to combine array paths coherently can happen also in IF or RF domain as typically done. Now if we choose $\phi_{LO} = 2\pi f_{LO}\tau$, LO signal in path 2 can be written as

$$x_{LO2} = \sin\big(2\pi f_{LO} t + \varphi(t-\tau)\big) \quad (2\text{-}17)$$

Now, the signals in the two IF paths with phase noise can be written as

$$x_{IF1}(t) = \sin(2\pi f_{RF} t)\sin\left(2\pi f_{LO} t + \varphi(t)\right) \quad (2\text{-}18)$$

$$x_{IF2}(t) = \sin(2\pi f_{RF} t)\sin\left(2\pi f_{LO} t + \varphi(t-\tau)\right) \quad (2\text{-}19)$$

By summing the IF paths coherently, using some basic trigonometric identities, denoting $f_{IF} = f_{RF} - f_{LO}$, and removing unwanted frequencies $f_{RF} + f_{LO}$, we can write

$$x_{IF}(t) = \tfrac{1}{2}\cos(2\pi f_{IF}\, t - \varphi(t)) + \tfrac{1}{2}\cos\left(2\pi f_{IF}(t) - \varphi(t-\tau)\right. \quad (2\text{-}20)$$

This can be further rewritten as

$$x_{IF}(t) = \cos\left(\frac{\varphi(t-\tau)-\varphi(t)}{2}\right)\cos\left(2\pi f_{IF} t - \frac{\phi(t)+\phi(t-\tau)}{2}\right) = A_c(t,\tau)\cos\big(2\pi f_{IF} t - \Delta(t,\tau)\big) \quad (2\text{-}21)$$

where $\Delta(t,\tau)$ is the combined phase noise of the paths and $A_c(t,\tau)$ represents the amplitude noise caused by the noncoherent combining due to the differences in the phase noise of the paths (i.e., delay of the LO). With $\tau = 0$, the phase noise is $\Delta(t,\tau) = \phi(t)$ and amplitude noise component $A_c(t,\tau) = 1$. Hence, it is evident that having any phase differences in the phase noise over the paths produces amplitude noise. The derivation can be easily generalized for more than two receiver paths/arrays. In the following sections, simulation results are given with different LO delay differences to showcase the impact of LO delay differences.

2.3.2 Simulation results on combined phase noise with asymmetrical LO routing

In this section, a simulation example with two receiver paths having a shared LO with delay differences between the LO paths from PLL to the mixers is shown. The phase noise is assumed to follow a model 2 of [HEX23-D23]. The model is based on PLL operating at 15 GHz with -150 dBc/Hz noise floor which is scaled to 300 GHz based on frequency multipliers. The phase noise is generated in frequency domain and the delay of the LO path is modelled as a constant slope in the phase response. Two different sets of simulations are done: one set with a continuous wave (CW) signal to show the combined phase noise spectra in the receiver and the other set is performed with a 64-QAM modulated signal (root raised cosine pulse shaping with roll-off factor of 0.35) to see the impact in the error vector magnitude (EVM). The simulation parameters are depicted in Table 6. Before combining the array paths in the baseband, a simple phase noise compensation is applied after the combining using in total of around 0.6% of the symbols for estimating slow phase drift.

Table 6: Simulation parameters

Scenario	PN model	f_0	BW	τ
CW	Model 2 of [HEX23-D23]	300 GHz	10 GHz (to observe spectrum)	{0.1, 0.35, 0.6, 1} ns
Modulated	Model 2 of [HEX23-D23]	300 GHz	4 GHz	{0.1, 0.35, 0.6, 1} ns

The simulation results are shown in Figure 2-34 (a-d) with CW signals and in Figure 2-35 (a-b). with modulated signals. Different sub-figures show the effect of different delay differences. From the spectrum plots, it can be seen that different delays produce nulls in the achieved phase noise spectra in different frequencies. In other words, the phase noise is effectively reduced in certain frequencies. For longer delays, more nulls can be observed over the band. One should note that a delay of 0.1 ns corresponds to 3 cm in free space, and in typical

PCB/IC material much less due to the permittivity of the substrate material. Hence, the used delays are practically possible, and even larger ones may be used. Based on the spectrums, equivalent phase noise-limited SNRs are calculated. It can be observed that when two LO paths have a delay difference, the SNR improves by 3 dB compared to when the LO paths have no delay differences. Furthermore, the benefit of SNR is not limited to only two LO paths. Hence, this corresponds to the same gain as one could get by combining two independent LO signals. It is expected that having more than two LO paths the results may be improved. This improvement in SNR directly translates to better EVM performance as can be seen in Figure 2-35. Hence, it can be seen that having delay differences in LO paths may help to improve the phase noise performance at the array level even when a single PLL is shared among the paths. This can be used as a method to reduce the flat region of the phase noise spectrum in wideband systems. The future work will look more at the modelling part of the LO routing and seek ways to combine the LO delay-based compensation with the traditional symbol-level compensation techniques. Also, we look at the impacts of some coarse analogueue filtering of the LO.

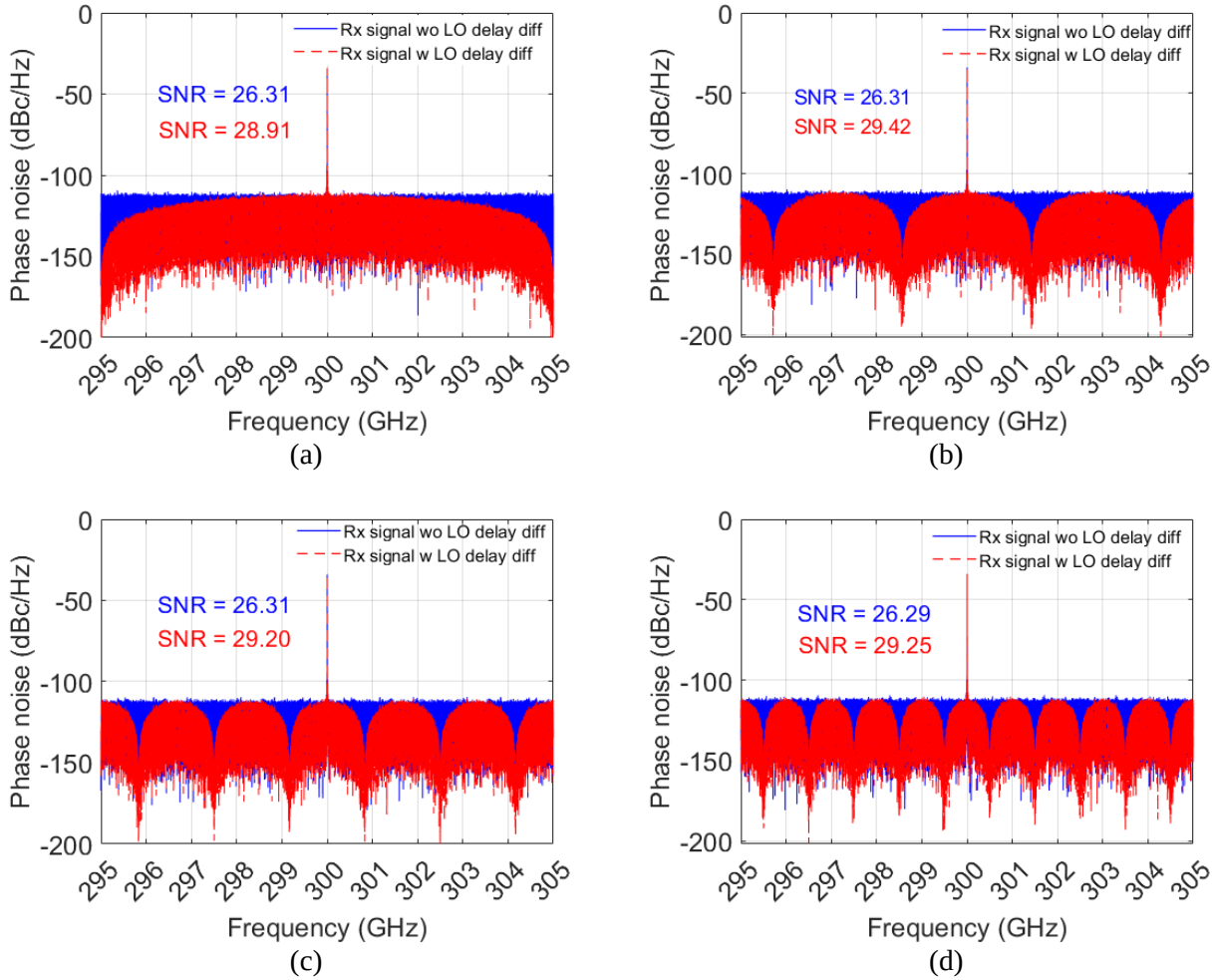
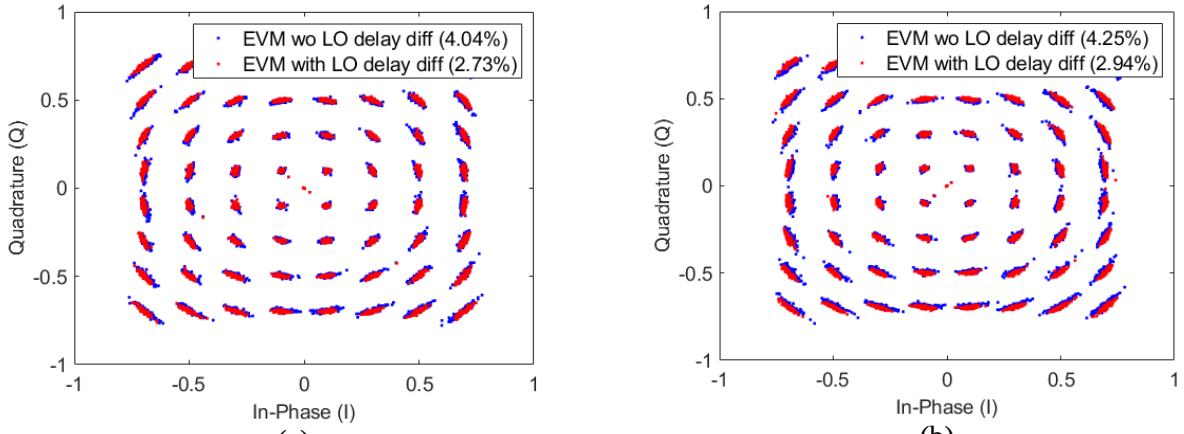


Figure 2-34: The PSD of phase noise after combining in receiver with (a) 0.1 ns (b) 0.35 ns (c) 0.6 ns and (d) 1 ns LO delay differences, respectively.



(a) (b)

Figure 2-35: Example received constellations of 64-QAM signal after combining. The blue points are drawn without LO delay difference, while the red points are with LO delay difference of (a) 0.35 ns (b) 1 ns. In both cases the EVM is improved due to the LO delay difference.

2.4 Transceiver architectures building on RTD devices

Resonant tunnelling diode (RTD)-based technology can be a promising candidate for future 6G communications, by leveraging its high frequency operation capability at room temperature with easy circuit realization, increased versatility, and energy-efficient implementation of the RTD-based device. This part addresses some of the benefits and limitations of using RTD-based devices for future 6G transceivers, with a focus on aspects such as output power, coherent operation based on array structures, antenna integration solutions and modulation aspects. The focus at this stage is on the analogue front-end part of the RTD TRx and no baseband/digital blocks are being considered.

Amongst the different III-V compound semiconductor technologies, InP is the most common material system used to implement RTD devices. The RTD is a nonlinear device which uses the quantum propagation effects of the electrons. The result is an N-shaped IV characteristic, which is also known as negative differential conductance (NDC). Oscillation occurs when the NDC compensates for the loss of the resonant circuit and the bias is within the NDC region over the IV curve. Currently the highest fundamental oscillation frequency achieved from a RTD oscillator sits at 1.98 THz [ISA17], which is also the fastest existing purely solid-state device. In addition, RTDs can operate as high sensitivity receivers due to the strong nonlinearity in the IV characteristic. This is achieved when the RTD is biased close to the peak current value on the IV curve, where the sensitivity of the device is highest. Therefore, the same RTD-based device can operate both as a transmitter and receiver depending on the selection of the bias voltage. Current SOTA RTD-based transmitters include 2 mW, 15 Gb/s at W-band [WAW+18] and 1 mW with 110 GHz modulation bandwidth at 260 GHz [AWW18]. The fastest reported wireless data rates using RTD based transmitters include 34 Gb/s at 500 GHz using single channel and 56 Gb/s using dual channel links [OHS+16], [OHS+17]. On the other hand, the fastest RTD based detector was reported recently with error free data rates of up to 27 Gb/s at 300 GHz [NND+18]. This is the best detector performance in any technology.

The promising potential of RTD-based technology for future 6G communications resides in their simplicity, e.g., a 1 mW 300 GHz source requires only a single RTD device realized using photolithography, while transistor-based technologies such as CMOS require an array of devices, sub 100 nm high-resolution lithography, and advanced circuit design techniques. Moreover, due to RTDs increased versatility (can operate both as a transmitter or receiver) and the fact that most of the times additional TRx building blocks such as PAs or LNAs can be omitted, the RTD-based front-end architecture is simpler than the more conventional, transistor-based approach. Figure 2-36 presents a simplified front-end RTD-based TRx architecture in a common communications setup, where both the Tx as well as the Rx incorporate RTD devices. Though not specifically illustrated, the RTD block incorporates the antenna solution as well. With respect to the shown architecture, an LNA block may be used or not in the receiver, choice dependant on the power level of the received signal. Moreover, based on the receiver required sensitivity for a given application, the RTD in the

Rx can be replaced by a Schottky Barrier Diode (SBD) that acts as an envelope detector whose output voltage is proportional to the e.g., amplitude modulated signal. Another approach to improve sensitivity while still using an RTD-based Rx approach is presented in [NND+19], where coherent detection is employed by setting the bias point of the RTD in the NDC region (oscillation region). This can improve the sensitivity of the receiver by over 20 dB compared to the direct detection case.

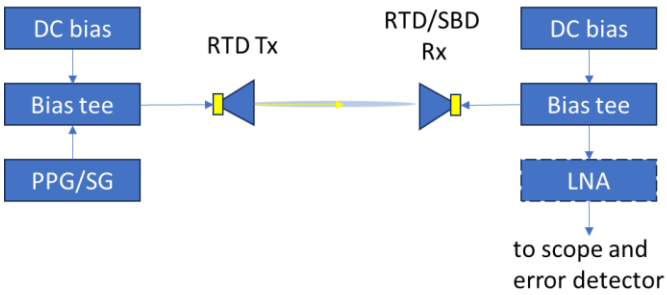


Figure 2-36: Simplified analogue front-end of RTD-based TRx architecture in a common communication setup.

2.4.1 Output power of RTDs

Regarding transmitters, two main approaches for THz signal generation are currently employed: photonic and electronic techniques [NDR16], [SNM18]. The photonic approach has proven to be effective to achieve data rates of up to 100 Gb/s along several meters long links [YJH+16], [JPO+22]. However, this approach is far too complex to be implemented in mobile/terminal devices.

Therefore, the electronic approach looks more feasible from this perspective. There are several solid-state technologies available, including impact avalanche and transit-time (IMPATT) diodes [AB14], tunnel transit time (TUNNETT) diodes [NPK+08], Gunn diodes [Eis10], Schottky barrier diodes (SBD) [MSL+17], transistors [PJG+18], and resonant tunnelling diodes (RTD) [AS21]. Figure 2-37 presents the SOTA in terms of output power generation in the sub-THz/THz frequency bands for some of these main solid-state technologies. With respect to the RTD, in case more output power is required, several RTDs can be combined in the form of an array and operated coherently. Other means of improving the radiated power would rely on increasing the span of the NDC region (by tuning the RTD epitaxial layer) and designing optimally loaded sub-THz/THz oscillators through the use of high-efficiency bias stabilization methods.

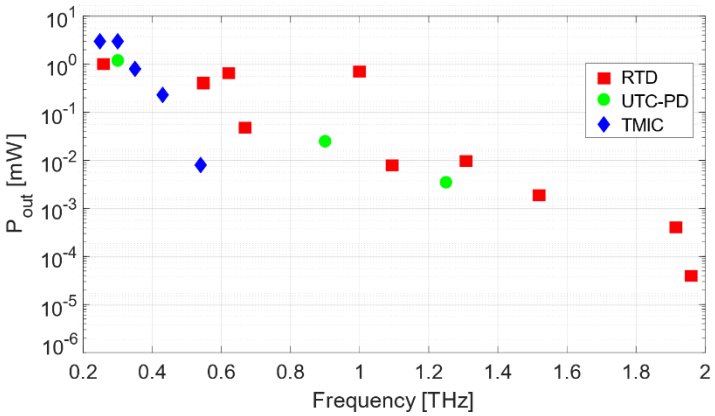


Figure 2-37: Output power of sub-THz/THz solid-state sources (UTC-PD: uni-travelling carrier photodetector; TMIC: transistor-based THz monolithic IC).

2.4.2 RTD-antenna integration

Due to the moderate output power of single RTD-based sources, antenna integration becomes crucial such that the generated THz radiation is properly coupled and radiated into free space. Therefore, a high-performance

antenna with high gain and directivity is required to radiate most of the RF power and support a larger link distance.

The most important challenge when integrating an antenna with the RTD is the large dielectric constant (i.e. 12.56 for InP) of the substrate on which the RTDs are manufactured. Therefore, most of the radiated energy will go into and through the substrate. As a result, any antenna implementation needs to account for this effect.

2.4.2.1 RTD-slot antenna

The combination RTD-slot antenna is probably the most common choice for designing RTD-based oscillators. In this case the antenna acts not only as a radiator but as a resonator element as well. To circumvent the effects of the higher dielectric constant of the substrate, different solutions can be implemented such as: lens antenna on the backside of the structure (for backside radiation) [AS16], tapered slot –Vivaldi (for horizontal radiation) [FSC+11] or placing a patch antenna on a BCB (benzocyclobutene) layer (for upward radiation) [OKO+15]

2.4.2.2 RTD-patch antenna

In terms of RTD-patch antenna, the solution mainly consists of integrating the antenna and the RTD on the same chip. In general, the approach consists of burring the RTD structure within a layer of BCB, which is sandwiched between the patch antenna and the ground plane, on top of an InP substrate. Therefore, radiation of the sub-THz/THz signal occurs upwards and into the air due to the antenna´s ground plane. Such a solution has been successfully demonstrated in [KSO13]. Moreover, same authors reported the successful integration of a 6x6 RTD-patch antenna array for high-power, and high directivity surface-emitting THz sources [KKY+22],

2.4.2.3 RTD bow-tie antenna

A bow-tie antenna is a simpler version of the planar slot antenna, with the added benefit of larger operating bandwidth. A proposed solution to compensate for the InP substrate effects consists in removing the substrate around the antenna conductor, while the ground plane underneath the diced substrate acts as a reflector and the antenna radiates to the air-side direction. This method has been demonstrated in [AKO+16], where a 230–325 GHz BW was achieved with a 11 dBi antenna gain at 280 GHz.

The previously summarized approaches toward RTD-antenna integration show that, despite the side effects of the InP`s higher dielectric constant on which RTDs are generally constructed, there are solutions to compensate these effects and allow a suitable RTD-antenna integration.

2.4.3 RTD-based arrays

2.4.3.1 *Coherent operation*

As RTDs can be used as signal sources without the need to add additional and more complex blocks such as power amplifiers (PAs), its power generation is not class leading, especially below 300GHz (limitation primarily determined by the geometrical limitations of the mesa size). An efficient way of circumventing this problem is to combine several of these devices in the form of an array and operate it coherently. The issue then relies on the extremely precise fabrication (e.g., photolithography) required, as the slightest mesa size variation (micrometer range) or antenna offset will lead to oscillation frequency drifts. This effect can however be partially compensated by tuning the operating point of the device or when an external reference signal is applied to injection-lock the devices (if the devices are within locking-range limits).

The effectiveness of the injection-locking approach is shown through the results in Figure 2-39. The results are derived from a previous work at Nokia, focused on the design of a 1x4 RTD-based array with patch antennas, operating at 28 GHz. The architecture of the array is based on 4 independent line-ups, each with its own RTD-based oscillator and patch antenna (design and prototype of the array are presented in Figure 2-38). The coherent operation is achieved by applying a reference 28 GHz signal to injection-lock the line-ups. This not only improves the radiated power of a single element, but also the phase noise characteristic of the devices. Note that the over-the-air results show a received signal power of -37 dBm including the 28 GHz pathloss at 50 cm and the Tx and Rx losses (mainly the two-stage power divider +cables).

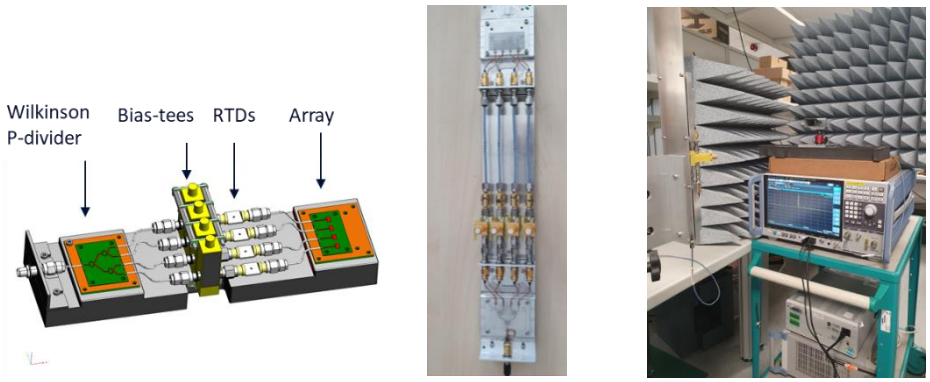
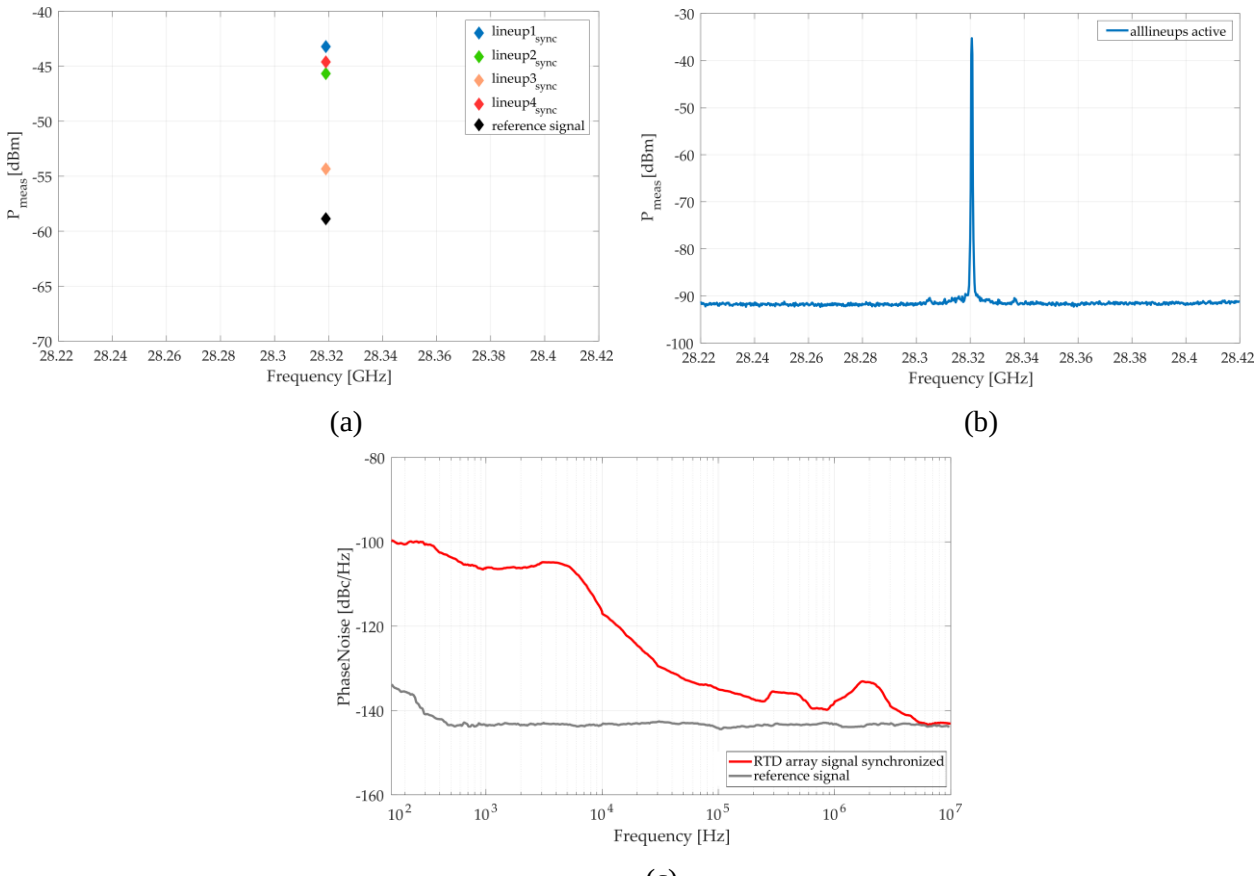


Figure 2-38: Linear 1x4 28GHz RTD-based array: design and prototype



(a) (b)

(c)

Figure 2-39: Linear 1x4 28GHz RTD-based array – Measurements: a.) power level of individual line-ups under synchronization; b.) power level when all line-ups are coherently operated; c.) phase noise

A 6x6 RTD array with integrated patch antennas operating coherently at 0.45GHz is presented in [KKY+22], which delivers 11.8dBm. Compared to the previous example, the RTD/patch antenna combination and complete array are manufactured over the same wafer die, providing a leaner solution. This is possible as coherent operation is achieved in this case through mutual injection-locking, with the antennas connected through coupled lines.

The previous two examples show that RTD-based array configurations can significantly improve the radiated output power vs. a single element under coherence operation when injection-locked (oscillation frequency of all elements is equal to the reference signal frequency, with the phase delta constant over time), and the technology can satisfy the extrapolation to virtually an unlimited number of elements.

Table 7 presents the relation between the number of elements, output power and antenna array gain. For the output power of a single RTD-based source 0 dBm is assumed, while for the antenna gain (e.g., bow-tie slot antenna) a typical 10 dBi value is used.

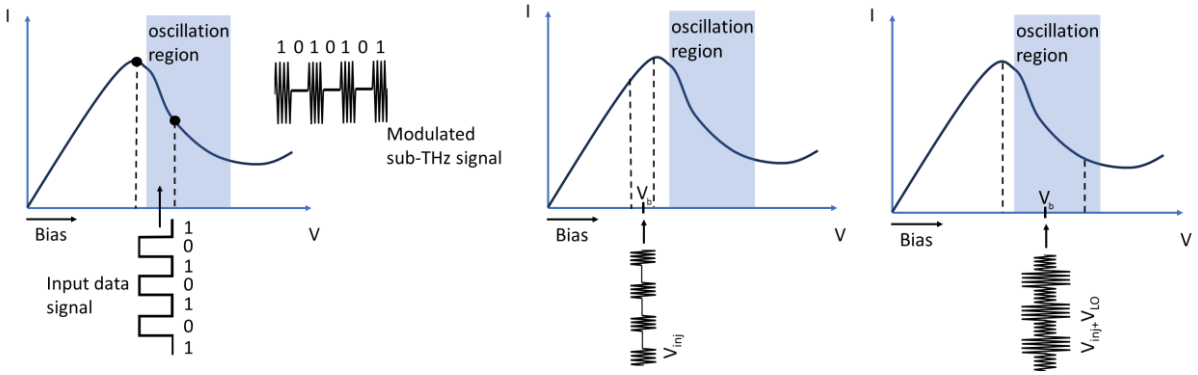
Table 7: Relation between array elements, output power and antenna gain.

	Output power	Antenna array gain
2x2 (4-elem)	6 dBm	16 dBi
4x4 (16-elem)	12 dBm	22 dBi
8x8 (64-elem)	18 dBm	28 dBi
16x16 (256-elem)	24 dBm	34 dBi
32x32 (1024-elem)	30 dBm	40 dBi

2.4.4 Modulation aspects

In terms of modulation, one of the most common choices is to use on-off modulations such as on-off keying (OOK) or amplitude shift keying (ASK). This is due to the RTDs nonlinear I-V curve which takes advantage of such on-off modulation schemes, by appropriately adjusting the bias voltage on the transmitter and the receiver (an illustration is given in Figure 2-40). One might think that the drawback of using these lower order modulation schemes is the lower spectral efficiency, however this can be compensated by the larger operation bandwidth (i.e, higher symbol rate) of the RTDs. A 3-dB bandwidth of more than 100GHz has been reported in [KAW20], allowing to achieve at least 50-Gb/s of data transmission with simple OOK. Moreover, using these simpler modulation schemes considerably reduces the effort in terms of digital post-processing as opposed to using higher spectral efficiency schemes.

For higher order modulation schemes (QPSK/QAM) and implicitly higher data rate transmission, an external vector-modulator can be used in an amplifier and mixer free approach (Figure 2-41), leading to a much-simplified system compared to conventional quadrature modulator-based transceiver architectures with driver/PA.



(a) (b) (c)

Figure 2-40: a) Tx-operation: intensity modulation, by superimposing a modulation signal over the RTD oscillation signal-Tx operation; b) Rx operation - direct detection; c) Rx-operation – coherent detection.

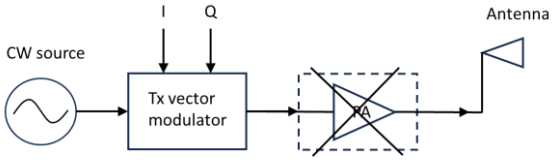


Figure 2-41: RTD-based source modulation using a vector modulator.

Performance of RTD-based devices versus other solid-state technologies in terms of data rate and link distance is reflected in Figure 2-42, where some of the most relevant scientific results up to 100 Gb/s are included. As can be seen the link distance is in most cases limited below 10m, with larger distances being enabled through the use of high-gain antennas, coherent receivers or power amplifier stages. On the other hand, RTD-based devices can achieve comparable performance with respect to other competing technologies, while making use of simpler modulation schemes. As such, RTD-based technology can pave the way towards a simple, low-cost solution suitable for future short-range high-capacity wireless communication links (e.g., ultra-broadband short wireless links between interactive smart devices, instantaneous data transfer in wireless kiosks).

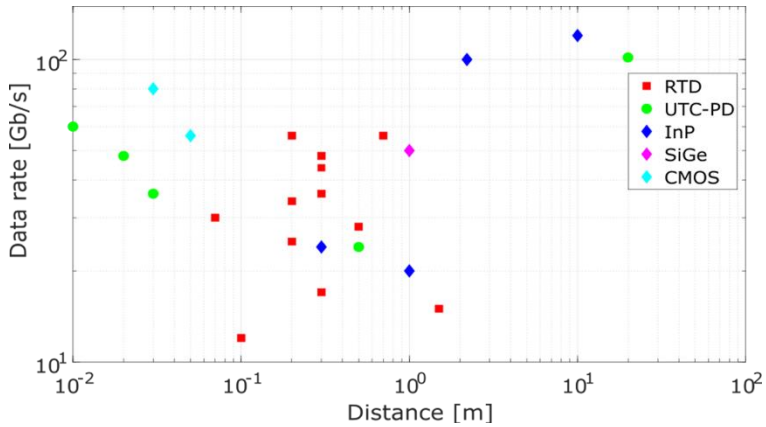


Figure 2-42: Data rate vs link distance of several solid-state technologies.

Based on the features and their performance, RTD-based devices could pave the way towards future 6G communication systems, in use cases which target mainly wireless short-distance links with extreme high-data rates (i.e., data kiosk applications or even small range hot-spots delivering high-speed data to demanding applications such as augmented or virtual reality).

2.5 Switched beam antenna lenses

2.5.1 Phased arrays

Phased arrays are the most common beamforming architecture employed at mmWave. A typical implementation is of 2 RF chains, one per polarization, connected to an array of dual-polarized antennas, where each antenna polarization is equipped with a phase shifter and a power amplifier (PA). It is important to note that in this configuration, the number of PAs is twice the number of antennas in the array, i.e. it scales linearly with the number of antennas. Benefits of phased arrays are well known and, among others, include a compact implementation amenable to integration in devices, simple and cost-effective array design and the convenience of electrical beam steering by means of electrically tunable RF phase shifters.

With operating frequency moving into sub-THz range, negative aspects of phased arrays become more prominent. These involve beam squint due to typically large relative bandwidths at sub-THz and increased losses of RF switches needed for implementing phase shifters [RPB+23]. A very serious problem is also incurred by comparatively large numbers of antennas needed to support link budgets similar to those in lower mmWave. A large number of antennas will result in a large power consumption closely connected with

integration problems (many RF connections in a limited space) and heating issues. The aforementioned negative aspects of phased arrays will severely limit their applicability in sub-THz [RPB+2023].

2.5.2 Switched-beam antenna lenses

A promising alternative to phased arrays is found in lenses that collimate the electromagnetic waves, thus providing a beamforming gain. The idea of using lenses to collimate RF signals stems back to very first days of radio research [LH89]. The directivity/gain of lenses depends on their physical aperture, which limits the use of lenses at lower mmWave frequencies to niche applications (e.g., military radar, space research) due to large physical size needed for providing a satisfying gain. At sub-THz frequencies, lenses with gains of 15-30 dB, which may be considered satisfactory for some applications, will have a diameter of a few centimeters, which enables their integration in radios with a small form factor such as mobile devices and small access points.

Lenses are commonly optimized for and used in a fixed-beam configuration, i.e., where the beam stays static over time. An attractive property of lenses is that they also enable dynamic beam switching by means of dynamic antenna port selection. Figure 2-43 (taken from [HJY+17]) illustrates the operation principle of switched-beam lenses. As shown in the leftmost panel for the hemispherical lens, N radiating elements are placed in the focal plane of the lens. Choosing the desired beam is equivalent to activating the corresponding radiating element and corresponding RF transceiver circuitry, which can jointly be referred to as antenna port. The principle of beamforming by antenna port selection extends to other lens types, such as Luneburg and half Maxwell fish-eye lenses with a gradient dielectric profile, where the focal points are located on the surface of the lens. Gradient profile lenses and all-metal lenses such as geodesic lenses [QLC+22] enable a scanning range that is typically wider than hemispherical lenses.

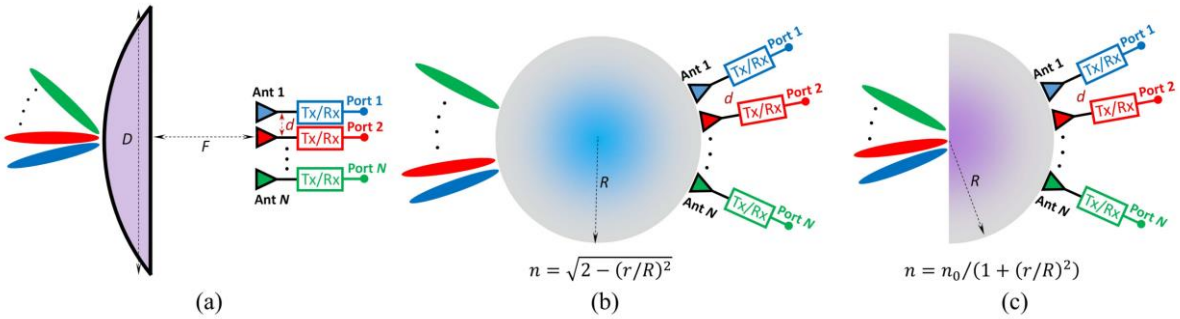


Figure 2-43: Switched-beam antenna lenses, a) hemispherical lens with homogeneous dielectric, b) Luneburg lens with gradient-index dielectric, c) half Maxwell fish-eye lens with gradient-index dielectric. Taken from [HJY+17], copyright 2017 IEEE.

Lenses have significant advantages over phased arrays when operating in sub-THz. One is the lack of beam squint due to the absence of phase shifters. Another key advantage is that only one transceiver chain per polarization is active at any given time. In other words, it is necessary to activate only one PA per direction and polarization. This is in direct contrast with M (where M is large, typically on the order of tens or hundreds) PAs active in a phased array with M antennas. Consequently, one can expect a significant reduction of transceiver power consumption when using switched-beam antenna lenses compared to using phased arrays. The interplay between array and lens size, power consumption and EIRP for the two competing beamforming systems is analyzed in the following section.

2.5.3 Power consumption of phased arrays versus lens-based architectures

The architectures under comparison are illustrated in Figure 2-44, assuming single polarization for simplicity. As explained previously, beam switching in the lens-based architecture is done by selecting an appropriate antenna port from the set of N ports by activating the corresponding transceiver circuitry (works equivalently

for transmit or receive modes) and connecting the said circuitry to the digital baseband. Transceiver circuitry at unused antenna ports is switched off for the purpose of saving power. In contrast, beam switching in the phased-array architecture is done by choosing an appropriate codeword of phase shifts $[\phi_0 \ \ \phi_1 \ldots \ \phi_{M-1}]^T$ for the M antennas of the phased array. For the latter configuration, it is assumed that all M PAs are active. For the sake of clarity, it should be noted that the number of antenna ports N in the lens-based architecture is generally not equal and is unrelated to the number of antennas M of the phased array architecture.

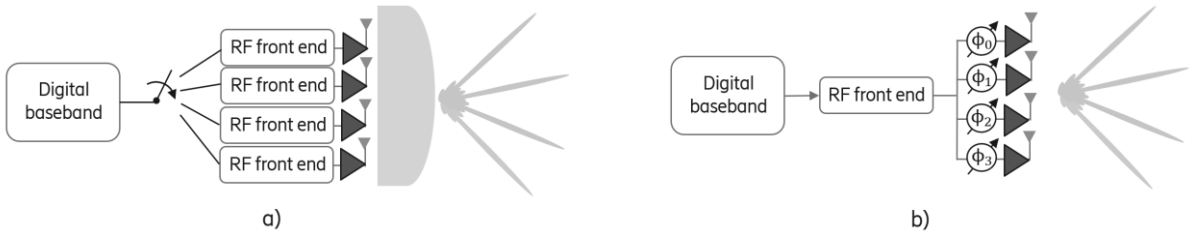


Figure 2-44: Illustration of beamforming architectures being compared, a) switched-beam lens-based architecture, b) phased-array-based architecture.

For the lens-based architecture, a hemispherical lens with diameter d_{lens} is assumed. Directivity of such a lens as function of physical aperture is [RCB+21]

$$D_{lens} = 10\log_{10}\left(\frac{4\pi\eta A}{\lambda^2}\right), \tag{2-22}$$

where η is the aperture efficiency, λ wavelength and A physical aperture of the hemispherical lens, calculated as

$$A = \frac{\pi d_{lens}^2}{4}. \tag{2-23}$$

Assuming an antenna element with gain G_{ant} and ignoring the dielectric losses in the lens and directivity losses at the antenna-lens interface, the effective isotropic radiated power (EIRP) at the lens boresight related to peak PA output power is calculated as

$$EIRP_{peak,lens} = P_{PA,out,peak} + G_{ant} + D_{lens}, \tag{2-24}$$

where $P_{PA,out,peak}$ is peak (saturation) output power of the power amplifier.

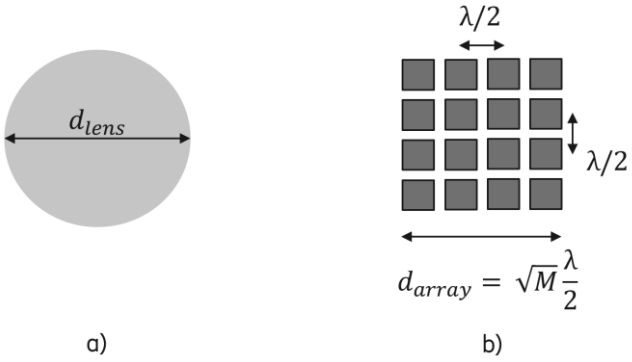


Figure 2-45: relevant geometry parameters for a) lens, b) phased array, front view.

For the phased array-based architecture, a square array with M array elements is assumed, with inter-element distance of λ/2. Array geometry is as shown in panel b) of Figure 2-45Figure . The EIRP of the phased array architecture can be calculated as

$$EIRP_{peak,array} = P_{PA,out,peak} + G_{ant} + 20\log_{10} M, \quad (2\text{-}25)$$

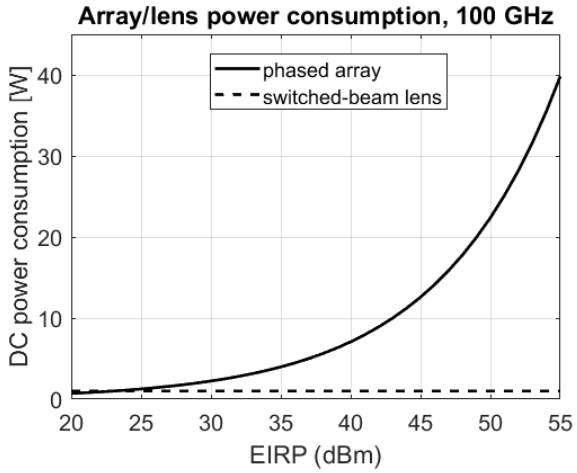
in which $10\log_{10} M$ is the contribution of M active PAs and another $10\log_{10} M$ is the (approximate) array directivity.

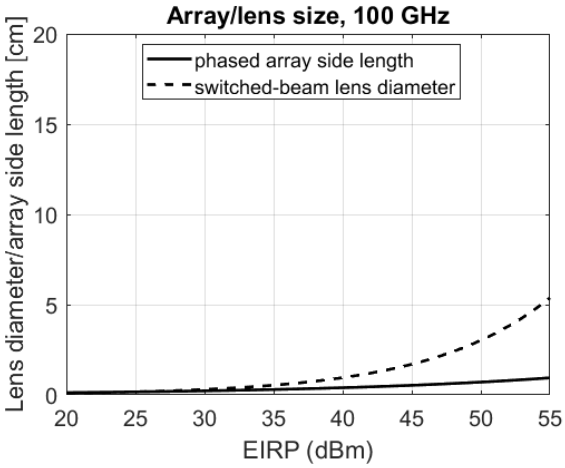
Expressions (2-24) and (2-25) allow for comparing lens and array geometries for a given EIRP, where target directivity for the lens is determined by the size of the lens, and the directivity of the array by the number of antennas M. Another interesting aspect of comparison is the power consumption of the beamforming architectures, represented by the DC power consumption of the PAs. A set of typical values for $P_{PA,out,peak}$ and PA power efficiency $PE_{PA,peak} = P_{PA,out,peak}/P_{DC}$, resulting in DC power consumption P_{DC}, is assumed based on the typical numbers collected in [WHM+21], for several relevant bands in the 100 – 300 GHz range and listed in Table 8.

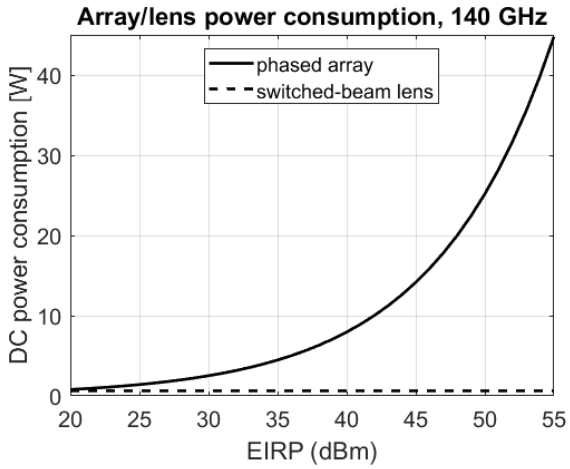
Table 8: values of typical PA key performance indicators in 100 - 300 GHz frequency range.

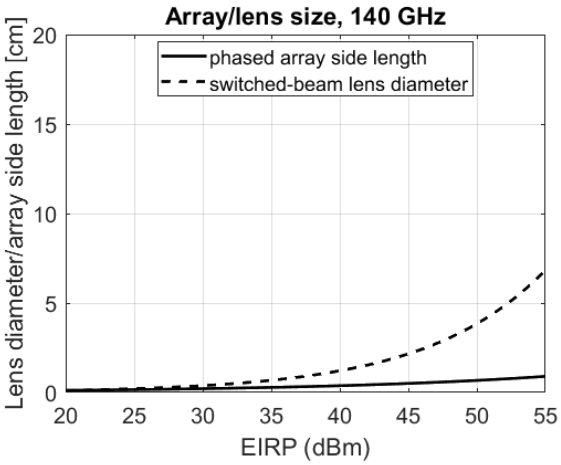
Band	$P_{PA,out,peak}$ [dBm]	$PE_{PA,peak}$ [%]	P_{DC} [W]
100 GHz	20	10	1
140 GHz	15	5	0.63
220 GHz	5	2	0.16
275 GHz	0	1	0.1

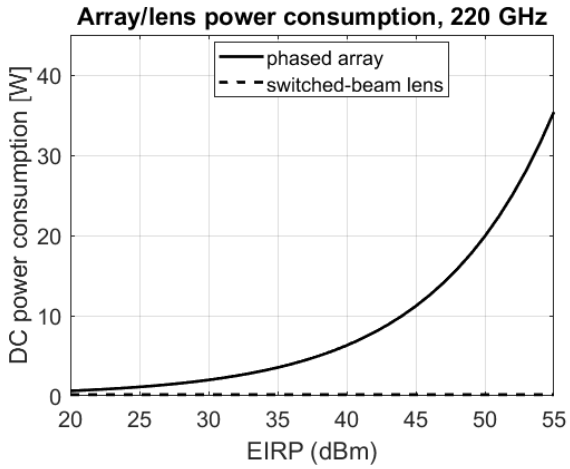
Power consumption of lens-based transmitter will simply be P_{DC} (as there is only one active PA), whereas the power consumption of the phased array will be MP_{DC}.

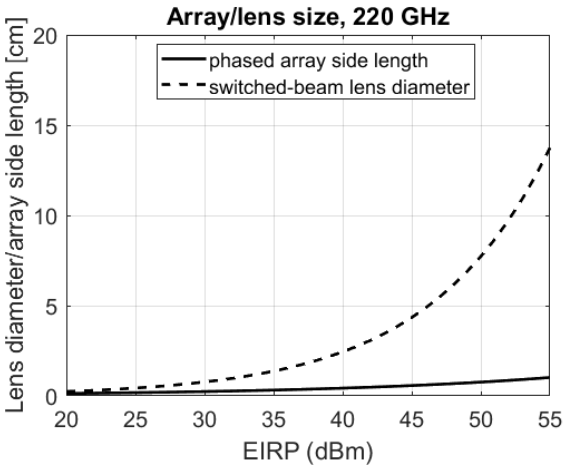

Array/lens power consumption, 100 GHz
phased array
switched-beam lens
DC power consumption [W]
EIRP (dBm)


Array/lens size, 100 GHz
phased array side length
switched-beam lens diameter
Lens diameter/array side length [cm]
EIRP (dBm)


Array/lens power consumption, 140 GHz
phased array
switched-beam lens
DC power consumption [W]
EIRP (dBm)


Array/lens size, 140 GHz
phased array side length
switched-beam lens diameter
Lens diameter/array side length [cm]
EIRP (dBm)


Array/lens power consumption, 220 GHz
phased array
switched-beam lens
DC power consumption [W]
EIRP (dBm)


Array/lens size, 220 GHz
phased array side length
switched-beam lens diameter
Lens diameter/array side length [cm]
EIRP (dBm)

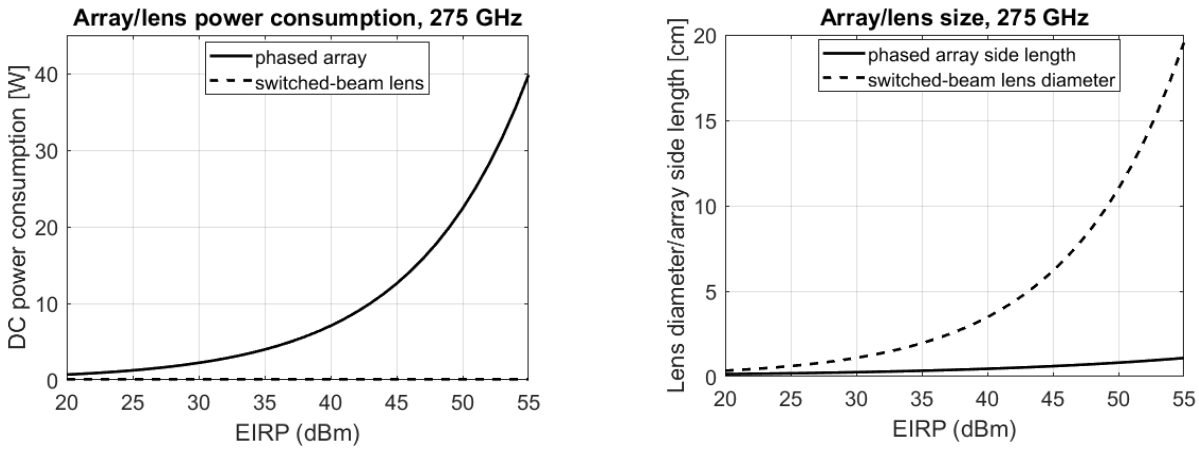


Figure 2-46: scaling of power consumption (left-hand side) and size (right-hand side) of lens-based and array-based with target EIRP.

Combining the data from Table 8 and employing (2-24) and (2-25) gives an opportunity to see how lens size, array size, and power consumptions of lens-based and array-based architectures scale with target EIRP. The results, obtained assuming $G_{ant} = 3\ dB$ and lens aperture efficiency $\eta = 0.5$, are given in Figure 2-46.

Several important conclusions can be drawn from the results:

- Power consumption of the lens-based architecture is always smaller than the power consumption of the phased array-based architecture, due to only 1 PA being active at any given time. For the assumed PA power numbers, the power consumption of the PA section is $\leq 1\ W$ for all frequency bands of interest, which should be an acceptable power consumption for mobile devices;
- Phased array power consumption is $> 20\ W$ for EIRP $> 50\ dBm$ and array sizes of few cm^2 – it is safe to assume that heat dissipation will be such a problem for phased arrays that it will render them infeasible at sub-THz for large EIRP, even when implemented at network radios (base stations/access points);
- In devices, where area is scarce, lens-based architectures will likely not be able to support EIRP > 40 dBm due to excessive lens sizes required beyond that operating point. Designs operating at 100 -140 GHz have an advantage over devices operating at > 200 GHz since they require comparatively smaller lenses for the same EIRP (due to PAs having higher peak output power).

It can be concluded that substituting phased arrays with lenses should be considered for both devices and network radios at sub-THz, primarily due to significantly lower power consumption incurred. Physical size of lenses will limit the EIRP achievable at devices and consequently UL link budget; maximum ranges of few tens of meters are to be expected, which is still acceptable for e.g., indoor applications. Research efforts in sub-THz radio design should focus on the problem of efficiently integrating optimized lens designs with optimized RF designs – research efforts so far have mainly focused on one or the other.

3 Reflective intelligent surfaces design

Reflective Intelligent Surfaces (RIS) have recently emerged as a potential solution to enhance the coverage of LOS-dominated propagation at high frequencies. By adaptively refocussing the incoming signal to a specific direction, they can enhance the system reliability by improved propagation diversity and offer a solution to link blockage problems.

In this section, a simulation environment is used to model the RIS impact on the system performance. The simulation can compute the received power in the target direction as well as non-wanted directions. Simulation results are validated by comparison to a measured prototype. Both active and passive RIS configurations are considered.

A second prototype based on varactors is investigated. It enables extra flexibility in the RIS control. S parameters and radiation pattern are simulated.

Additionally, RIS integration challenges are considered. Control aspects are essential for the interface definition and performance. The controlling can be performed from infrastructure, UEs, or other devices. The key properties of a RIS solution and related control protocols are described and quantified.

3.1 RIS models at mmWave

A RIS is a new component for a radio communication system to modify the radio wave propagation between the transmitter and the receiver. Without a RIS the signal power at the receiver antenna is the sum of line-of-sight (LOS) and non-LOS signals. With a RIS mounted in the area covered by the transmitter antenna, the radio energy transmitted toward the RIS can be directed to the receiver. This generates a new non-LOS signal path between the transmitter and the receiver. This can be used to decrease the signal loss, increase the diversity factor, or allow to bypass blocking objects.

A way to evaluate the behaviour of a RIS in an environment without going through complex and costly on-site measurements is by simulation. We create a simulation model of a RIS and validate the simulation model with measurements on our RIS hardware that were performed under laboratory conditions. With this we validate the RIS simulation model so that it can be used for system level link simulation, e.g., by using a ray-tracing field simulator. Results from these simulations can also be valuable as input for RIS hardware design and optimization or to improve overall system performance.

The hardware simulation model of the RIS describes the relaying properties of one RIS HW realisation that was built and tested. The definition of the coordinate system which defines the orientation of the RIS and the directions of incoming and outgoing signal are shown in Figure 3-1.

Depending on the direction of the incoming signal and the RIS parameter settings, the signal will be reflected by the RIS. The goal of the simulator is to determine how the RIS forwards the incoming signal to the desired outgoing direction and also how the RIS distributes the incoming power into directions other than the desired outgoing direction. This is done by calculating the radiation pattern of the RIS for an excitement with an incoming signal. To make handling of the results easier for ray tracing simulators, the radiation pattern is also analysed and information about signal strength, outgoing direction and polarisation on significant lobes is determined and provided. As a result, the data needed to represent the radiation behaviour is greatly reduced in comparison to a full radiation pattern.

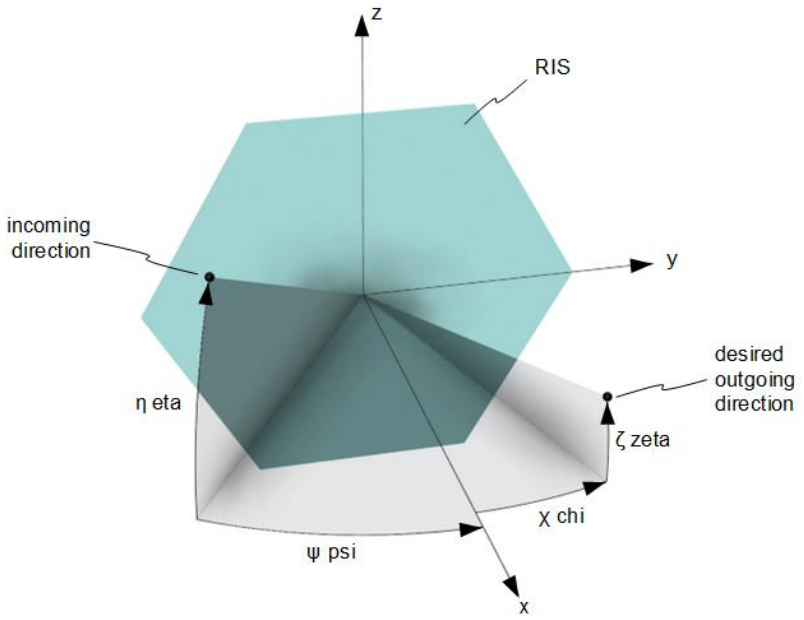


Figure 3-1: RIS coordinate system.

The RIS consists of many unit cell elements that each allow a controlled backscattering of incoming signals. The elements are arranged on the yz-plane (x=0). The positive x-axis is perpendicular to the RIS and points into the half-sphere, where signal interaction takes place.

The incoming signal travels towards the RIS and is defined with the azimuth angle ψ (psi) and the elevation angle η (eta). The desired outgoing signal direction is at an azimuth angle χ (chi) and an elevation angle ζ (zeta). In the simulator, the element configuration settings of the RIS are determined so that the elements are controlled in a way that the maximum possible part of the incoming signal is radiated into the desired outgoing direction. Also, for this element configuration the radiation pattern of the RIS is calculated on a grid of azimuth angle φ (phi) and elevation angle ϑ (theta). The angular increment (resolution) of the radiation pattern can be set in the simulator.

The radiation pattern of the RIS also depends on the radiation pattern of the unit cells and the number and position of the unit cells. This depends on the HW realisation of the RIS and is—for now—not changeable. In the current simulation model, the effects of the RIS element properties are based on calculations and measurements of the known RIS realization. The results therefore will match the effects of this realisation closely.

For creating the RIS simulation model MATLAB has been used. The simulator contains data of a realized RIS. The properties are hardcoded into the simulator and contain properties like element layout and a model of the radiation pattern of the element itself. The actual control circuit of each element can however be configured so that different control strategies can be explored. For the validation of the simulation model, measurements in an anechoic chamber have been performed with a RIS that was realized in the mmWave frequency band. The RIS consists of 127 elements that can be individually controlled and are arranged in a hexagonal pattern. This RIS realization has two modes of operation:

- Passive mode, where the RIS elements reflect incoming signals with two possible phase settings. For this mode, vertical polarization is used.
- Active mode, where each RIS element employs and amplifier to increase the signal level before reradiation. To improve stability in the active mode, the RIS receives signals in horizontal polarization and reradiates the amplified signals in vertical polarization. For control of the elements the amplifier can either be switched off or on.

3.1.1 Simulator input parameters

Simulation results can be calculated depending on the input parameters shown in Table 9.

Table 9: RIS simulator input parameters.

Parameter	Description	Example
ψ (psi)	incoming azimuth angle in rad.	-10/180 π
η (eta)	incoming elevation angle rad.	15/180 π
χ (chi)	outgoing azimuth angle rad.	10/180 π
ζ (zeta)	outgoing elevation angle rad.	10/180 π
f	center frequency in Hz	$24 \cdot 10^9$
r	radius of sphere on which the results will be computed in m	10
φ (phi)	azimuth values of the grid on which the results will be computed	$[-\frac{\pi}{2}:\frac{\pi}{90}:\frac{\pi}{2}]$
ϑ (theta)	elevation values of the grid on which the results will be computed	$[-\frac{\pi}{2}:\frac{\pi}{90}:\frac{\pi}{2}]$
PhaseStates	reflection phase states for passive mode in rad.	[0 π]
PowerStates	reflection gain states for active mode	[0 2]

3.1.2 Output parameters

After finishing the calculation the radiation pattern of the relay gain of the RIS is returned.

The relay gain of the RIS (G) is represented as an array. The first dimension relates to the angles specified for ϑ (theta), the second dimension relates to the azimuth angles φ (phi).

With RIS gain G, the formula for calculating the received power at the direction specified by φ and ϑ is as follows:

$$P_{\mathrm{Rx}} = P_{\mathrm{Tx}} \cdot G_{\mathrm{Tx}} \cdot \left(\frac{\lambda}{4\pi d_{\mathrm{Tx}\rightarrow\mathrm{RIS}}}\right)^2 \cdot G \cdot \left(\frac{\lambda}{4\pi d_{\mathrm{RIS}\rightarrow\mathrm{Rx}}}\right)^2 \cdot G_{\mathrm{Rx}}$$

The RIS relay gain assumes polarization matching between both the incoming and outgoing signals with the RIS.

You can imagine the relay gain as a fictional gain inserted between a 0 dBi isotropic receive antenna and a 0 dBi isotropic transmit antenna that are both located at the centre position of the RIS. The gain summarises the antenna characteristics of the RIS for reception and transmission and also the signal attenuation or amplification of the RIS. The gain is calculated so that it produces the same field strength as the RIS in the direction specified by φ and ϑ.

Lobe information of the radiation pattern of the RIS is calculated and described in Table 10. This is a $N \times 6$ array where each line contains information of a lobe in the pattern. The first line always shows the information of the lobe which points into the desired outgoing direction. The other lines hold information of sidelobes and are sorted according to their peak gain in descending order. This data is extracted from the radiation pattern to directly use the results in ray-tracing simulations.

Table 10: Lobe data simulation output.

column index	**1**	**2**	**3**	**4**	**5**	**6**
quantity	relay gain	lobe energy	elevation angle of peak	azimuth angle of peak	elevation -3dB beamwidth	azimuth -3dB beamwidth
unit	dBi	relative to total radiated power	radians	radians	radians	radians

The beamwidths denote the angular region in elevation and azimuth where the gain decreases by 3 dB compared to the peak gain of that lobe.

Lobe analysis is stopped as soon as 99 % of the radiated power is accounted for. The number of lines in the lobes matrix therefore depends on the radiation pattern.

3.1.3 Verification

The RIS design, prototyping and simulation model developed in [RIS21-23] can be operated in active and passive mode. The simulation model is enhanced so that it can be used for evaluating the benefits of RIS for a communication system.

3.1.3.1 RIS realization

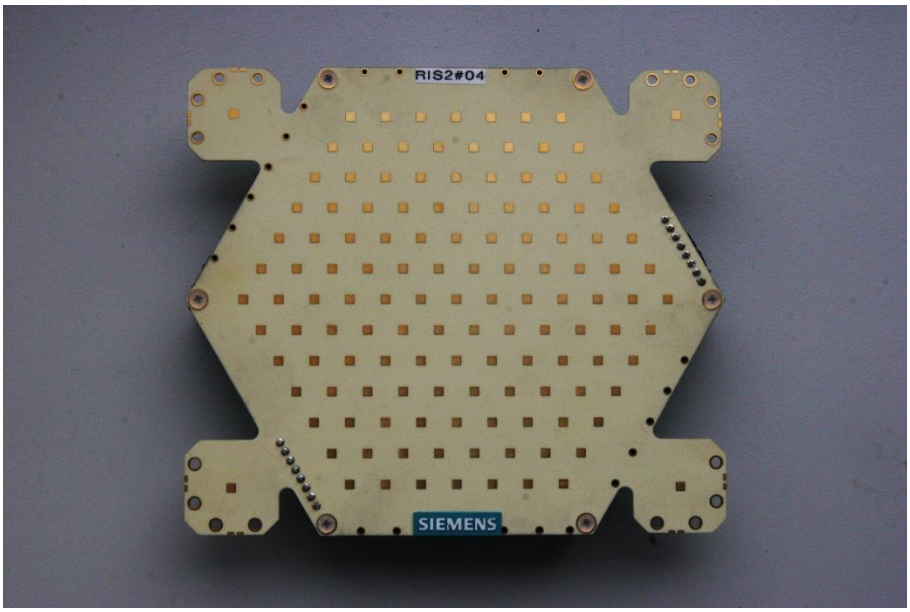


Figure 3-2: RIS HW realisation with 127 unit cells.

3.1.3.2 Active mode

In the following is an example of measurement and simulation for the active mode of the RIS (see Figure 3-2). The simulation is done for an effective element gain in the RIS elements of 3 dB. This was the achieved value at 23 GHz. At 24 GHz and above the amplifier gain drops slightly.

The radiation pattern matches the simulation (see Figure 3-3) very well and the relay gain for the desired lobe also matches closely.

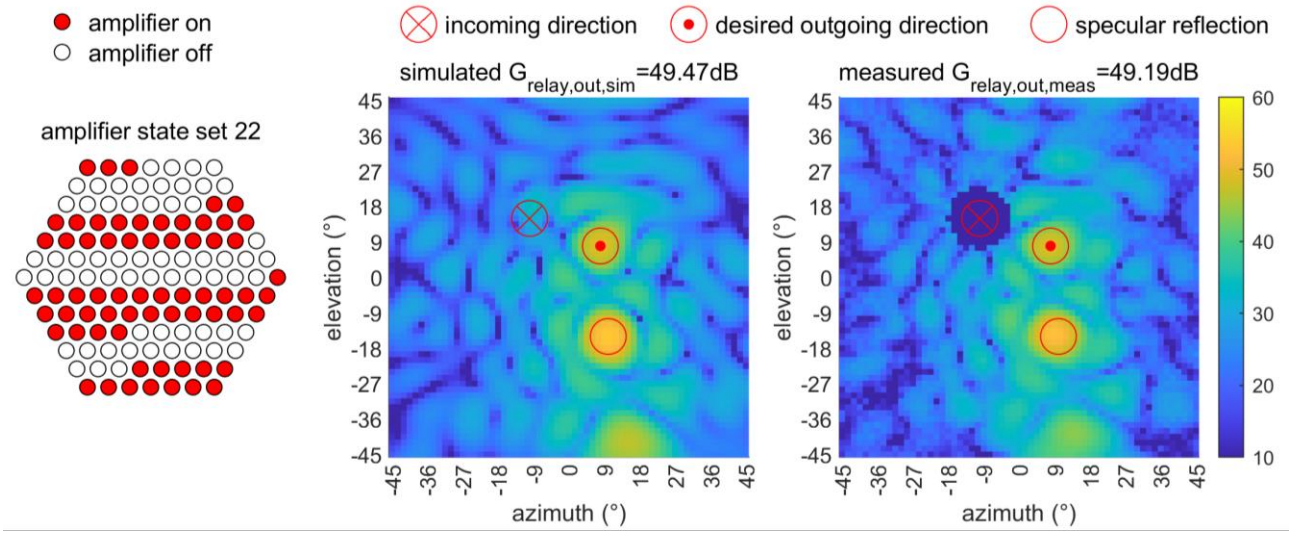


Figure 3-3: Comparison of simulation and measurement of the active RIS for one scenario. To the left, the configuration of the elements which was determined for the incoming and outgoing directions is shown.

3.1.3.3 Passive mode

In the passive mode, the simulation (see Figure 3-4) is done with lossless reflection at the antenna elements. Only the phase shift of the two discrete states can be set. Since the RIS was optimized for the active mode, reflection losses are higher in the physical elements and there is a discrepancy in the peak relay gain. However, the shape of the radiation pattern matches well between measurement and simulation. In future work the frequency characteristic of the reflection behaviour can be included in the simulator. No effort has been put into that as the RIS was not expected to perform very well in passive mode.

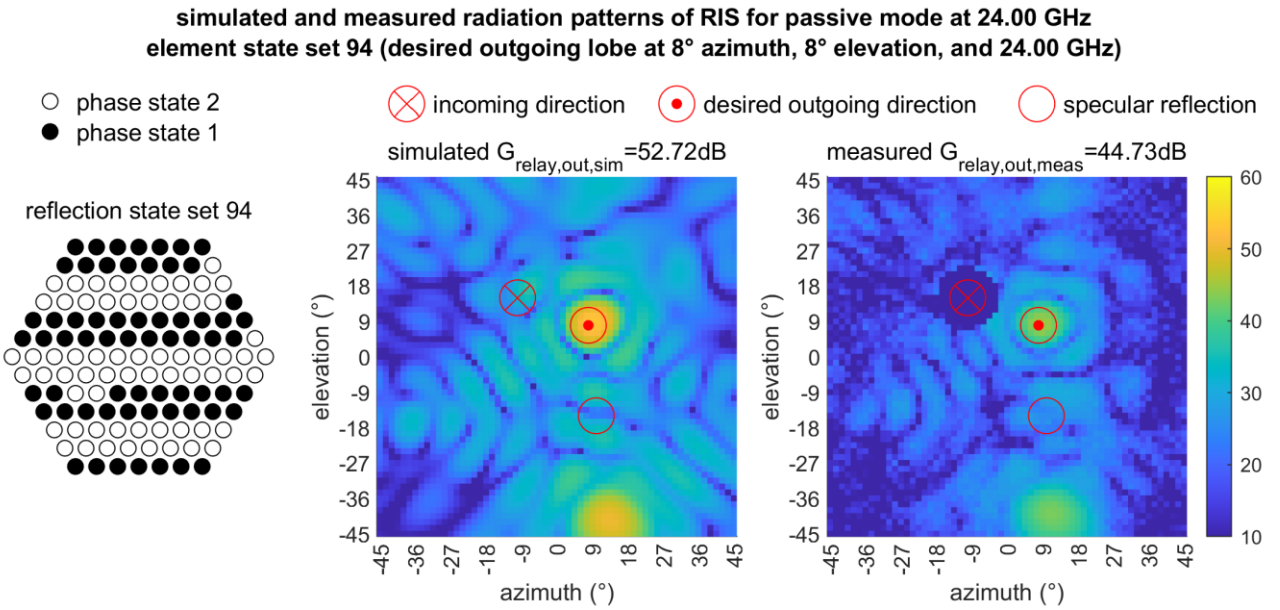


Figure 3-4: Comparison of simulation and measurement of the passive RIS for one scenario. To the left, the configuration of the elements which was determined for the incoming and outgoing directions is shown.

3.1.4 Future work

For simulations on the behaviour of RIS, it can be of interest how the RIS reflects and scatters signals that do not come from the incoming direction. In future work, these results can be provided by calculating the element configurations from the incoming and desired outgoing direction, but calculating a radiation pattern for which the RIS is excited by signal coming from a different direction than the incoming direction.

In passive mode the RIS is fully reciprocal. However, in active mode, the RIS can only receive in one fixed polarization and transmit in another polarization. This can be solved by employing two RIS next to each other which have orthogonal polarization alignment. Another option could be a time-synchronized RIS which is capable of switching the direction of amplification and is synchronized with the up- or downlink packet timing.

3.1.5 Varactor-based RIS

For the design of the unit cells, a continuous control of the reflection phase coefficient instead of a quantized one (2 bits) has been chosen to allow in addition to the typical beamforming to be able to shape the beam, add phase offset in the reflected field, that can bring advantages for Reflectarray/RIS applications.

The reconfigurable surface design based on a controllable High Impedance Surface (HIS) [SZB+99], [SSJ+03]. At the resonance frequency, the surface acts as an artificial magnetic conductor (phase = 0°). The control of the reflection phase coefficient is done by adjusting variable capacitances placed on each unit cell of the surface. Varactor diodes are used to control a variable capacitance by tunning the inverse voltage applied. These capacitances modify the resonance frequency of the unit cell and then its reflection phase coefficient at the working frequency. Several prototypes has been developed with a surface control principle and tested for a single polarization reflectarray working in 5.0-6.0 GHz frequency band [RBF+13] (Figure 3-5) and a dual polarization one in K/Ka band [BRB+22] (Figure 3-6). A dual polarization surface has been designed working in the 5G mmWave frequency band [RIS23-D34] [Rat23], the geometrical configuration and performance will be presented below.

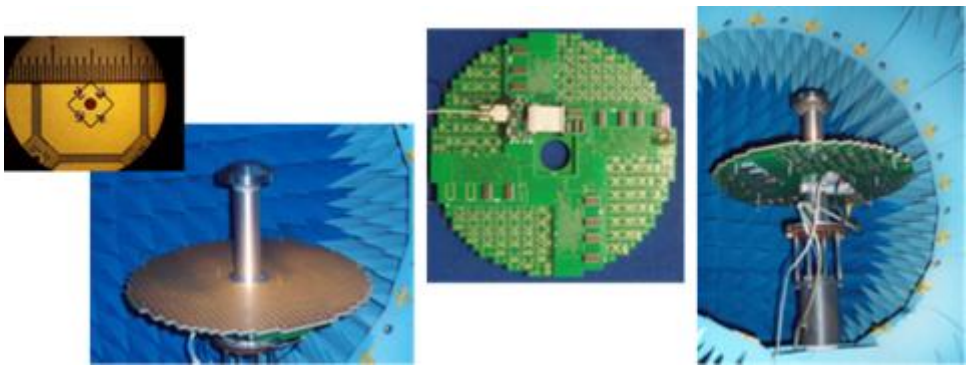
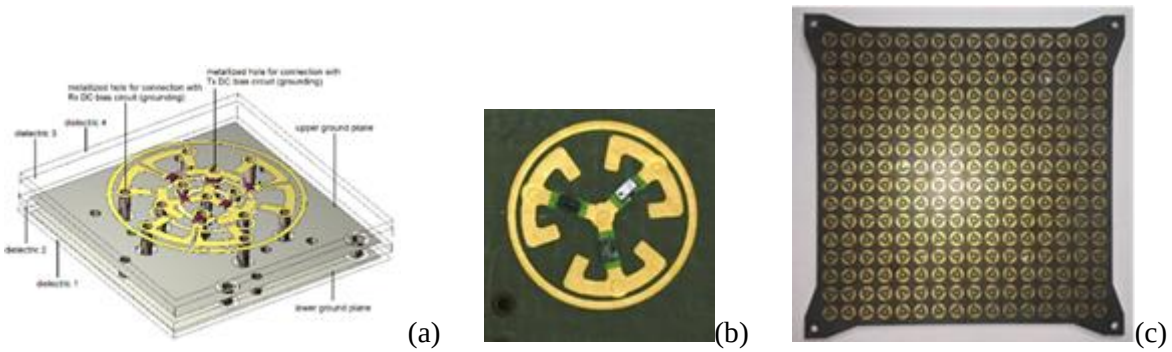


Figure 3-5: Reconfigurable Reflectarray Prototype working between 5.0 to 6.0 GHz.



(a) (b) (c)

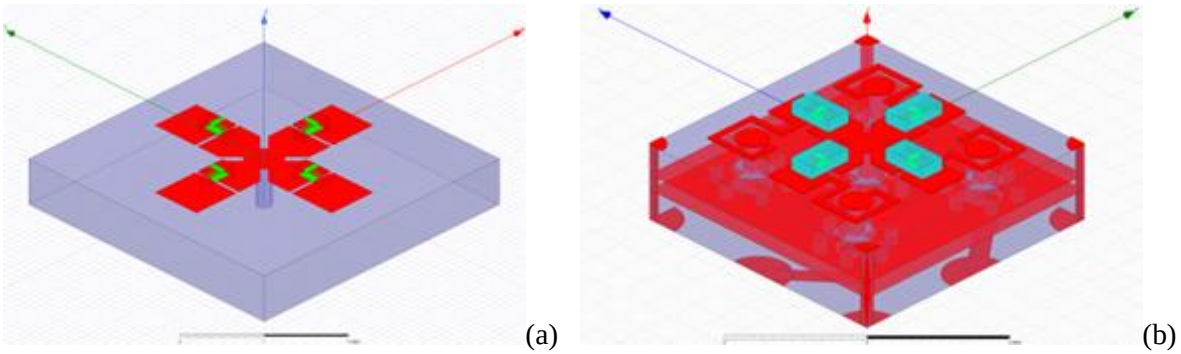
Figure 3-6: Dual frequency band (K/Ka) unit cell (a), K band unit cell (b), K band reconfigurable prototype (c)

3.1.5.1 Unit Cell Configuration

The requirements for the unit cell are the following:

- Continuous reflected phase control with at least 300°
- Full n258 working frequency band (24.25-27.5 GHz)
- Low level of reflected field depolarization
- Unit-cell size below the half-wavelength to avoid parasitic effect.

To achieve these requirements, the unit cell is composed by two orthogonal metallic strips on a dielectric substrate with a metallic central post connected to the ground plane (Figure 3-7.a). A varactor diode is placed on each arm of the crosse above a slot and connects the two parts of each arm. To be able to control the voltage applied on the 4 varactor diodes, a biasing circuit is added with the goal to minimize the discrepancies on the RF performance of the unit cell (Figure 3-7.b). the biasing circuit is composed of RF short circuits (shielded metallic vias) and RF blocks (spiral inductances).



(a) (b)

Figure 3-7: 5G mmWave pre-design unit cell (a), final design unit cell including biasing circuit (b).

This unit cell can control independently two orthogonal polarizations. The design has been done to keep the same RF unit cell part (crosse+diodes) and only the voltage panel control behind changes to addresses single or dual polarization control of the unit cell [Rat23].

3.1.5.2 Simulated results.

3.1.5.2.1 S11 (20-30 GHz)

Figure 3-8 and Figure 3-9 present the ability of the unit cell to control independently two orthogonal polarizations. For the normal incidence, we can observe that the variation of the capacitance (2 diodes) along x or y axis controls only the respective parallel polarized field for the phase (Figure 3-8) and amplitude (Figure 3-9). The coupling between the 2 polarizations (Figure 3-9) stays below -20 dB.

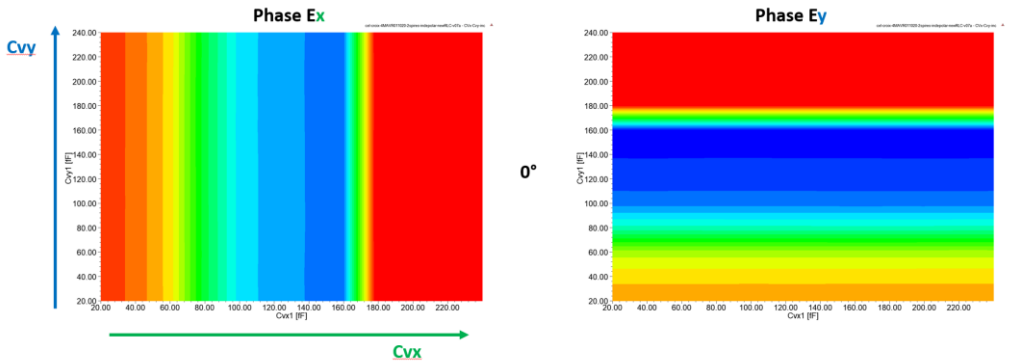


Figure 3-8: Phase controls at 26 GHz and 0° of incidence.

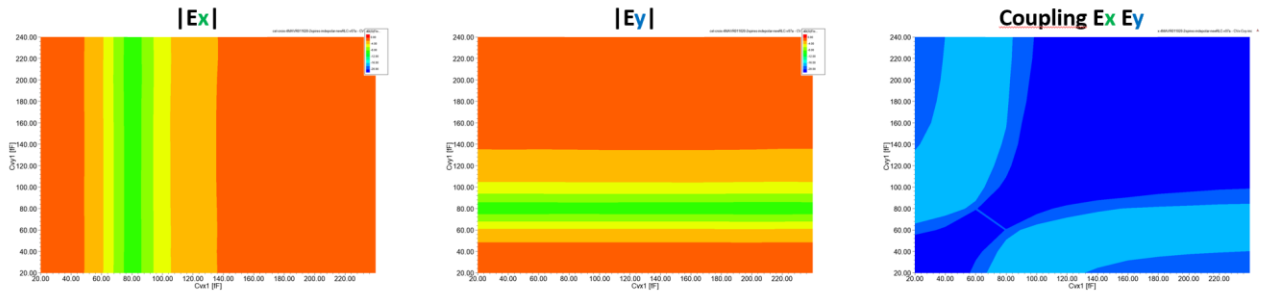


Figure 3-9: Amplitude of reflection coefficient and polarization coupling at 0° of incidence at 26.0 GHz.

We can observe the reflection phase coefficients depending on different configurations:

- Figure 3-10 presents the reflection phase coefficient of the unit cell in normal incidence for the varactor capacitance varying between 20.0 fF to 240.0 fF at 24.25 GHz (green line), 26.0 GHz (red line) and 27.5 GHz (blue line),
- Figure 3-11 presents the reflection phase coefficient between 20.0 GHz and 30.0 GHz in normal incidence for several value of capacitances and for a 100.0 fF capacitance value for several incident angles of the EM field,
- Figure 3-12 presents the reflection phase coefficient between 20.0 GHz and 30.0 GHz for a 100.0 fF capacitance value for several incident angles of the EM field.

In the working frequency band, a 300° phase excursion is observed on Figure 3-10 and Figure 3-11 and a stable reflection phase coefficient depending on the incident angle on Figure 3-12.

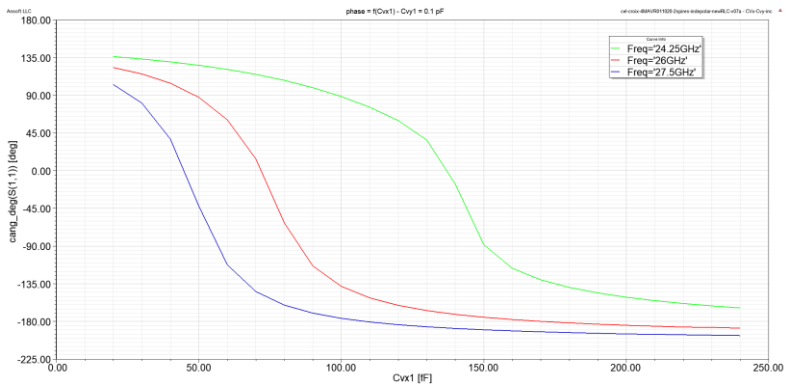


Figure 3-10: Reflection phase coefficient depending on capacitance excursion at 24.25, 26.0 and 27.5 GHz.

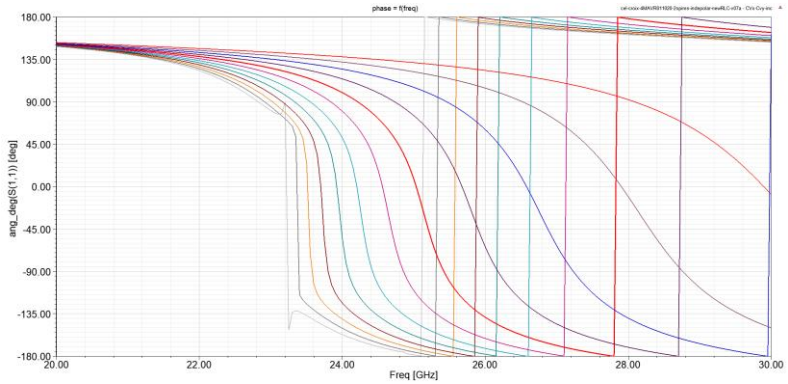


Figure 3-11: Reflection phase coefficients depending on frequency for several capacitance values at 26.0 GHz.

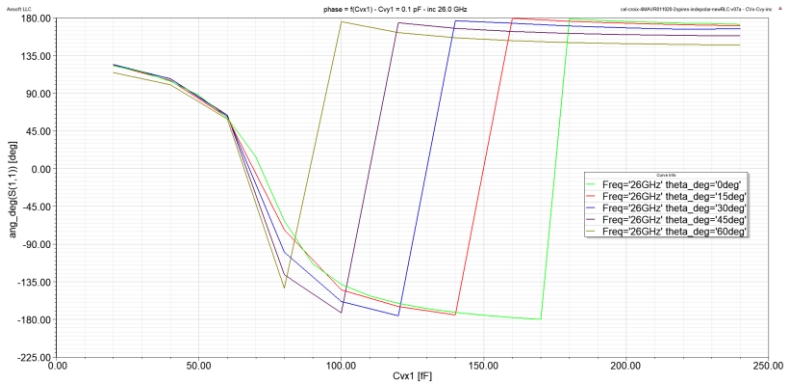


Figure 3-12: Reflection phase coefficients depending on frequency for several angle of incidence of the EM field at 26.0 GHz.

3.1.5.2.2 Bandwidth of influence (BoI)

The BoI [APK+23] characterizes the frequency range where the reconfigurable surface (RIS) modifies the reflected field compared to a "dumb wall". This frequency range is larger than the required working frequency band because of EM properties of the unit cell that are not negligeable outside the working frequency band.

To evaluate the BoI of the mmWave reconfigurable unit cell, an extended frequency band has been simulated. The S11 modules and phases for several capacitances values are presented respectively Figure 3-13 and Figure 3-14.

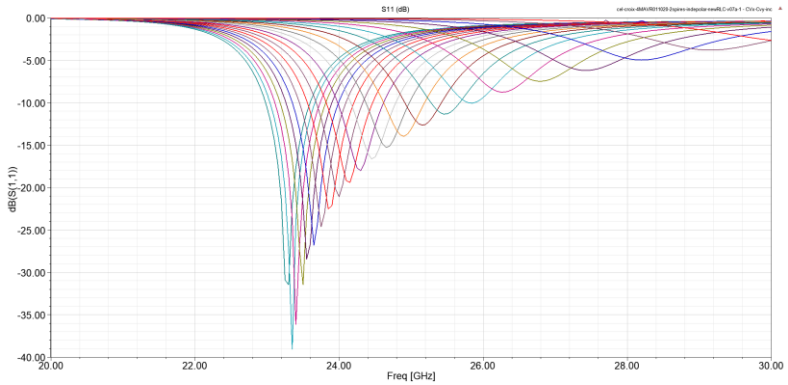


Figure 3-13: Reflection module coefficients depending on frequency for several capacitance values (20-40 GHz).

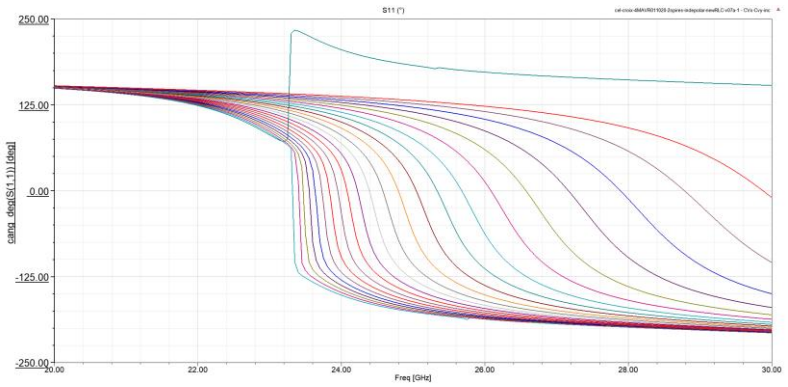


Figure 3-14: Reflection phase coefficients depending on frequency for several capacitance values (20-40 GHz).

For this unit cell, the BoI, as defined in [APK+23], starts at 22.3 GHz and stops at 38.5 GHz is presented in Figure 3-15. We can observe that the BoI as expected is larger than the working frequency band of the phase control, the RF behaviour of the unit cell impacts the reflected field above 28 GHz because of its size below the half wavelength in the working frequency band.

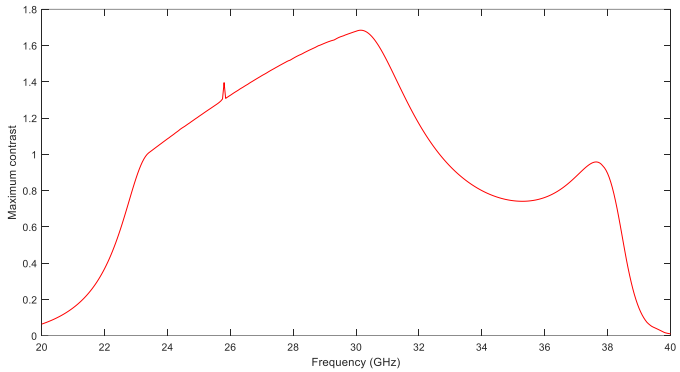


Figure 3-15: Bandwidth of influence of the reconfigurable unit cell.

3.1.5.2.3 3D radiation patterns

Figure 3-16 presents the radiation patterns for several capacitance values (from 0.02 to 0.24 pF every 0.02 pF) at 26.0 GHz. The pattern shape changes with the capacitance because of the change of EM behaviour of the unit cell, a lower level is observed around 0.08 pF because of the resonance of the unit cell at this frequency and this capacitance to act as a magnetic conductor.

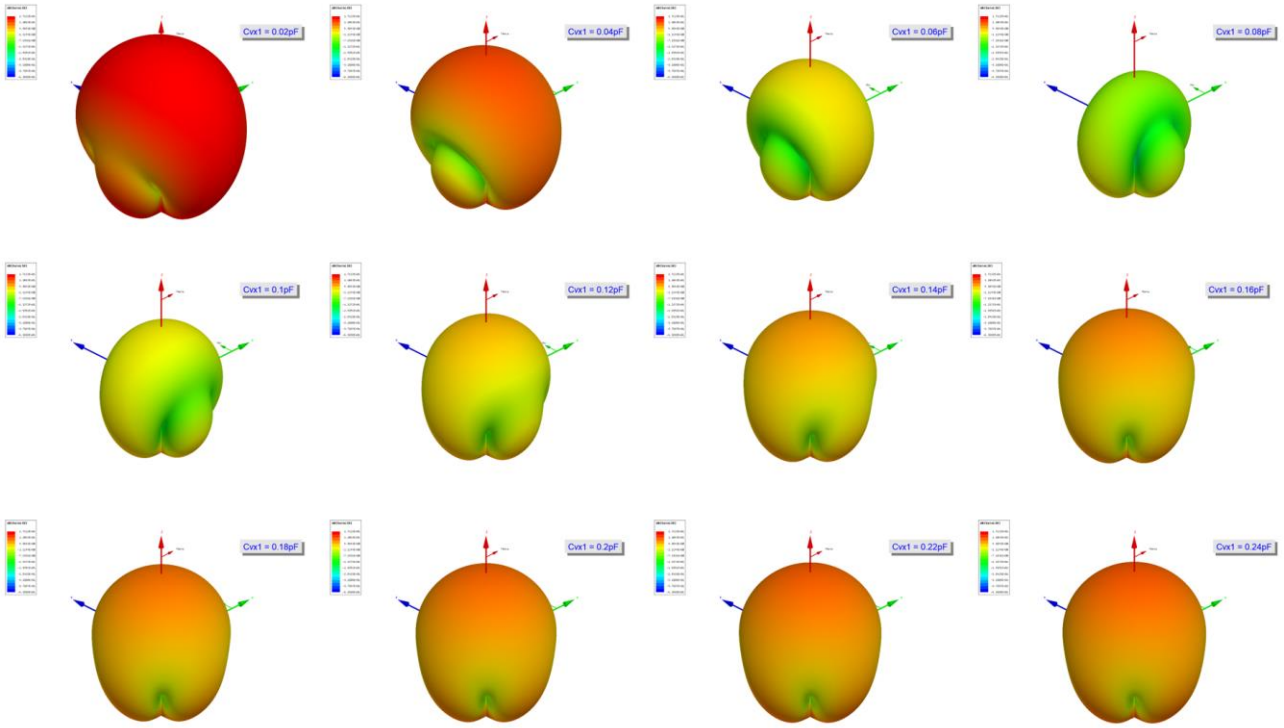
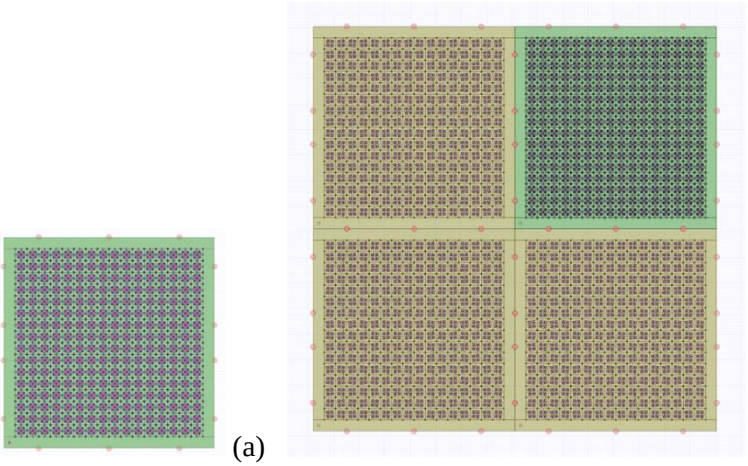


Figure 3-16: 3D radiation patterns for the unit cell for several capacitance values at 26.0 GHz.

3.1.5.3 Prototype and manufacturing

A prototype is under manufacturing to validate the presented design. A modular tile has been designed to be able to adjust the size of the RIS by assembling the adequate number of tiles. 16x16 unit cells compose the tile with an external contour to fix them together on a support as can be seen in Figure 3-17.



(a) (b)

Figure 3-17: 16x16 unit cells tile (a) and assembling of 2x2 tiles (b).

3.2 RIS system integration

For the integration of RIS in a radio communication system the following aspects are relevant for integration and control. Figure 3-18 shows an overview of the RIS control architecture. The RIS properties describe the capabilities and limitations of the RIS. They determine in which manner the RIS can be functionally integrated into the radio communication system. The RIS-local-controller-functions offer means to control the RIS on the physical level (e.g. element control) based on information that is available for the communication links (e.g. CSI, RSSI, position, movement, …). Control commands for the RIS-local-controller are defined in the RIS-control-interface and follow the RIS-control-interface-protocol. The RIS-central-controller prepares the setup data for the RIS based on information about the radio links. This information can be obtained by the

radio communication network (A), the user equipment (B) or others (C, e.g. a factory automation system). The RIS-central-controller-functions are provided as a generic description to control the RIS. The goal is that the same RIS-control-interface can be used for all three control types. In Figure 3-18, the three different architecture types are shown.

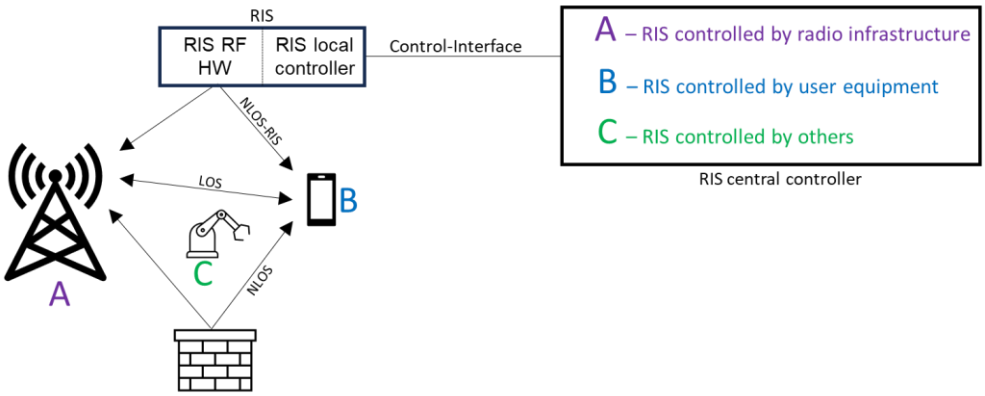


Figure 3-18: RIS control architecture types.

3.2.1 RIS properties

A RIS realization can be described with the following properties. The typical value range shown in Table 11 gives examples for RIS that might be employed in the near future, but it is no general limitation. The properties listed are independent of the used technology.

Table 11: RIS properties.

Property	Base unit	Typical value range	Comment
frequency range	GHz	for FR1, FR2, FR3	specifies for which frequency range the RIS is functioning
typical relay gain	dB	>40 dB	Typical relay gain that can be expected from the RIS during operation. It is defined as the gain of a fictional amplifier between a 0 dBi isotropic receive antenna and a 0 dBi isotropic transmit antenna located at the RIS centre position that produces the same field strength as the RIS in the desired target direction.
accepted incoming polarisation	-	H&V, H, LHCP, V, -45°, ±45°, …	Polarizations that the RIS can accept
produced outgoing polarization	-	H&V, H, LHCP, V, -45°, ±45°, …	Polarizations that the RIS can produce

polarization reciprocity	-	Yes Yes, with up/downlink awareness No	Polarization reciprocity defines whether signals can go in both directions without a change in polarization. Most passive RIS will be reciprocal. An active RIS might change the polarization and need direction switching dependent on up/downlink slots.
boresight alignment	dimensional axis	x	Boresight (center) direction of beam steering range with respect to the RIS construction / mounting instructions
polarization alignment	dimensional axis	e.g. vertical polarization aligned with z-axis	Alignment of polarization mode(s) with respect to the RIS construction / mounting instructions
typical beam width	degree (°)	from few degrees to some 10 degrees	Typical beam-with of the main beam when trained on the receiver.
beam control range	degree (°)	from few 10 degrees up to 150 degrees	Angular range in which the main beam can be steered.
switching time	µs	from µs to seconds	Transient time it takes to switch from one RIS configuration to the next (beam switching)
timing accuracy	µs	> a few µs	Accuracy of the time instant at which a pre-scheduled change of a RIS configuration (e.g., beam switching) takes place.
power consumption	W	> a few W	Power consumption of the RIS in operation.
maximum incoming signal power density	W/m²	from few W/m² to several 10 W/m²	Intended maximum power level at the RIS where it conforms to radio regulations regarding linearity.
dimensions	x, y, z	around 15 cm or larger	RIS size depends on operation frequency and typical relay gain
weight	kg	from below 1 kg to several 10 kg	

3.2.2 RIS-local-controller functions

The RIS-local-controller is part of the RIS and has the task to execute commands and configurations coming from the RIS-control-interface input/output protocol. For this it must control the RF hardware of the RIS (for instance its unit cells) appropriately. The RIS-local-controller should offer at minimum the following functions:

Table 12: RIS-local-controller function.

Function	Description	Comment
Set relay parameters - simple	- incoming direction - desired outgoing direction - polarization (if applicable) - start time info - stop time info The RIS will perform as a relay specified by the directional parameters and during the times specified by start and stop values.	Basic operation of the RIS.
Set relay parameters - advanced	List of parameter sets, each containing: - incoming direction - desired outgoing direction - polarization (if applicable) - relative lobe power and timing info: - start time info - stop time info The RIS will perform as an advanced relay specified by the directional parameters and during the times specified by start and stop values.	Advanced operation mode of the RIS where the RIS is optimized to relay signals into multiple directions and create nulls for specific input and output directions.
Power management	Put the RIS in sleep mode or operating mode.	
Time synchronization	Network time	Synchronize the internal clock of the RIS with the network time so that beam control can be synchronized and triggered with signal transmissions.
Hardware ID FW version	report on used HW and FW version.	
FW update	Support download of new FW versions.	

3.2.3 RIS-control-interface, protocol, and functions

The connection of the RIS-local-controller with the RIS-central-controller will use the RIS-control-interface and the RIS-control-interface-protocol. Whereby the interface can be realized wired or wireless and can use standard or customised physical and Medium Access Control (MAC) layer realisations.

The RIS-control-interface-protocol shall include minimum the following commands:

Table 13: RIS-control-interface-protocol.

Function	Description	Comment
set	transmit parameter values to the RIS local controller	
time-sync	command to synchronize the internal clock	Initiates or executes a time synchronization procedure
get	read parameter values from the RIS local controller	
download	download datablocks	e.g. for FW updates

The RIS-central-controller shall include at least the following functions:

Table 14: RIS-central-controller functions.

Function	Description	Comment
calculate RIS relay parameters	Calculate relay parameters (e.g. to achieve a desired beam) with timing info to be sent to the RIS.	Depending on the known or expected x, y, z positions, orientation and movement of UEs, BSTs and RIS the relay parameters are calculated for specific time ranges.
control RIS	Send control commands to the RIS.	

Depending on the architecture type the central-controller-functions can be executed in different devices, e.g. BST, UE, or any other device.

The RIS-central-controller calculates the desired parameters for the RIS (incoming direction, outgoing direction, polarization, timing, ...) based on information about the

- Position and orientation of the BST and the RIS
- Position, orientation, and movement of the UE

If this information is not available an estimation or assumption of the relay parameters will be necessary. For this, training sequencies and the use of AI could be helpful.

4 6G System-on-Chip architecture

The development of SoC architectures for 6G devices is a challenging task, influenced by diversity of performance and cost requirements, scalable and efficient signal processing, security aspects, power management and the integration of AI functions. This section addresses specific areas of research, including the applicability of RISC-V for 6G signal processing and the design and integration of AI components in SoCs. Another aspect is the development of secure and scalable SoC architectures, where the primary challenge is to create an integrated trustworthiness that includes hardware, runtime operations and the operating system (OS). It also emphasizes the importance of energy harvesting for the continuous operation of IoT devices, addressing the conversion, regulation, storage, and management of electrical energy from ambient sources. The research work aims to develop a power management integrated circuit (PMIC) and an energy combiner (EC) for efficient utilization of multiple energy sources, as well as to investigate low-power machine learning (ML) algorithms for predictive energy-aware management.

4.1 DSP and AI SoC components

The Hexa-X project has outlined six distinct use case families along with their corresponding requirements, as detailed in Hexa-X-II deliverable D1.2 [HEX223-D12]. Recognizing the inherent diversity encapsulated in these varied use cases, the anticipation is that numerous device classes will emerge in the age of 6G [HEX223-D52]. Each of these classes will be tailored to meet diverse operational, deployment, performance, and cost requirements. In alignment with this strategic 6G vision, a plethora of cutting-edge computing, communication, networking, and sensing technologies will be introduced to enable a broad spectrum of novel smart applications within the 6G ecosystem. Consequently, novel SoC architectures must be developed addressing a wide range of technical considerations, including:

- ultra-low power consumption to support the proliferation of IoT devices,
- high computing power and efficient processing capacity to handle the complexity of signal processing, control, management and optimization algorithms in 6G communication technology,
- configurability and adaptability to diverse use cases, including ultra-reliable low-latency communications, massive machine-type communications, and enhanced mobile broadband,
- efficient AI processing capability by integrating specialized hardware for AI acceleration and efficient execution of ML algorithms,
- Security and privacy by incorporating robust security features to protect against potential cyber threats and ensure the privacy of user data,
- support for advanced wireless technologies, such as sub-THz communication, massive MIMO, and RIS.

The focus of this part of work is on the investigation, development and integration of signal processing and AI components in 6G SoCs.

The 6G SoC architecture must meet the performance demands of a wide range of applications, including Internet of Everything (IoE), virtual reality (VR), 3D applications, artificial intelligence, and massive machine-type communications (mMTC). The 6G SoC architecture shall represent a significant advancement in wireless communication technology, offering a comprehensive suite of signal processing capabilities to address the evolving requirements of 6G wireless systems. The architecture's computation capability scales from tiny IoT devices to high-performance devices, meeting the demands of ultra-high data rates, ultra-reliable low-latency communications, and global connectivity. Additionally, it addresses the need for energy-efficient and intelligent networking capabilities. These capabilities are instrumental in supporting the peak user data rates of 100 Gbps and ultra-low latency 0.1ms and ultra-reliability 10^{-9} frame error rate.

The general concept of SoC architecture is illustrated in Figure 4-1. The SoC HW comprise CPUs, DSPs, AI accelerators, dedicated HW accelerators and high-performance I/Os for RF-frontend, memory, and system integration. The HW architecture is complemented by SW layers dedicated to system configuration and adaptation, signal processing applications, HW-resource allocation and task scheduling and HW configuration and adaptation. Note, that SPF and system management functions can be implemented and/or complemented by AI.

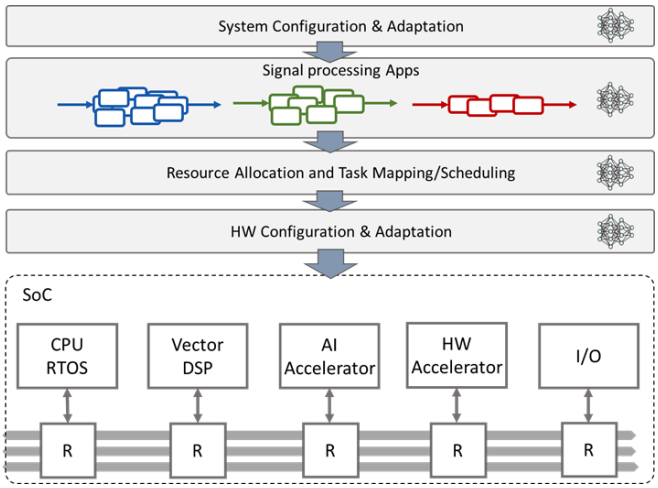


Figure 4-1: Principal SoC architecture employing general purpose processor, Vector DSP, AI-accelerator, HW-accelerator components connected by chip interconnect.

4.1.1 AI based signal processing

The evolving landscape of 6G wireless systems demands a deep exploration of wave propagation phenomena and the development of sophisticated signal processing techniques to address the diverse requirements of modern communication scenarios. The multidimensional nature of the wireless channel and signal models is illustrated in Figure 4-2. The wave propagation characteristics of 6G systems vary across different applications and scenarios, necessitating a comprehensive understanding of the channel behaviour and the associated signal models. This complexity, combined with the need to meet specific performance and cost requirements, presents a substantial challenge for the development of efficient signal processing algorithms and HW/SW architectures.

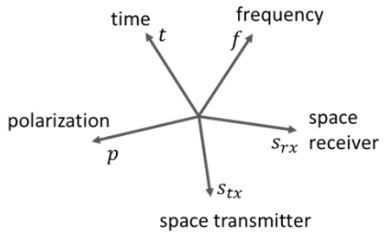


Figure 4-2: Illustration of multidimensionality of wireless channel and signal models.

The development of signal generation and reception capabilities in such complex environments necessitates the utilization of advanced physical layer techniques. These techniques encompass a spectrum of functions including waveform processing, linearization, beamforming/precoding/combining, modulation/demodulation, channel estimation, channel equalization, denoising, detection and coding/decoding. Additionally, higher protocol layer functions play a crucial role in enabling effective signal transmission and reception. The complexity and type of signal processing algorithms employed are directly linked to the specific communication performance requirements and are closely aligned with the diverse dimensions of the signal/channel environment.

The integration of AI in wireless communications is poised to revolutionize network management, optimizing performance, and enhancing operational efficiency. By leveraging AI, wireless networks can achieve increased capacity and capability, as well as improved resource utilization and cost efficiency. The following key areas demonstrated the benefits of AI on wireless network management and optimization:

- Link Level Performance: AI enables adaptation to network and channel conditions, facilitating interference prevention and mitigation, channel optimization, waveform optimization, and efficient

spectrum and spatial resource usage. This adaptation enhances link performance and optimizes resource allocation, contributing to increased network capacity and capability.

- Optimization of Radio and Processing Resources: AI-driven spectrum monitoring, sensing, localization, spectrum usage, beam optimization, and massive MIMO optimization optimize radio and processing resource usage. This leads to improved resource utilization, cost efficiency, and enhanced network performance.
- Network Management Automation: The integration of AI in wireless network management enhances operational efficiency by automating manual processes, reducing human error, and optimizing spectrum usage. This automation streamlines network operations, leading to improved performance and cost-effectiveness.

Even though AI is primarily used on the network infrastructure side, many applications on devices have recently demonstrated the benefits of AI over conventional signal processing algorithms. AI techniques has been employed to enhance signal processing in devices for tasks such as noise and interference reduction, modulation, demodulation, signal detection and decoding, leading to improved communication performance. In addition, AI can be utilized for spectrum monitoring, enabling devices to efficiently utilize available frequency bands. By leveraging AI, devices can optimize spectrum usage, adapt to dynamic channel conditions, and enhance the reliability and quality of wireless communication. Moreover, AI-driven techniques are employed to optimize algorithms and processing resource usage in devices. This optimization leads to improved resource utilization and cost efficiency. Finally, AI can facilitate localization and sensing capabilities in devices, enabling improved spatial awareness and location-based services. Devices can enhance their ability to sense and localize signals, leading to more efficient communication and interaction with the surrounding environment.

In general, AI techniques can be categorized according to their relation to the signal processing function (SPF) (Figure 4-3):

- stand-alone SPF – employs AI models for SPF
- AI-enhanced SPF – improves capability of non-AI SPF by supplementing it with AI
- parameter/model optimization and adaptation of SPF

In Figure 4-3, $f(x)$ and $f(x,\theta)$, $g(x,\theta)$ represent some function and AI model, respectively. Moreover, θ are trained model parameters and $\oplus$ some function combination operator. The benefits of improving communication performance in wireless systems using AI-based signal processing have been widely reported in various areas. However, in many cases, the computational requirements increase disproportionately, rendering many techniques impractical for application.

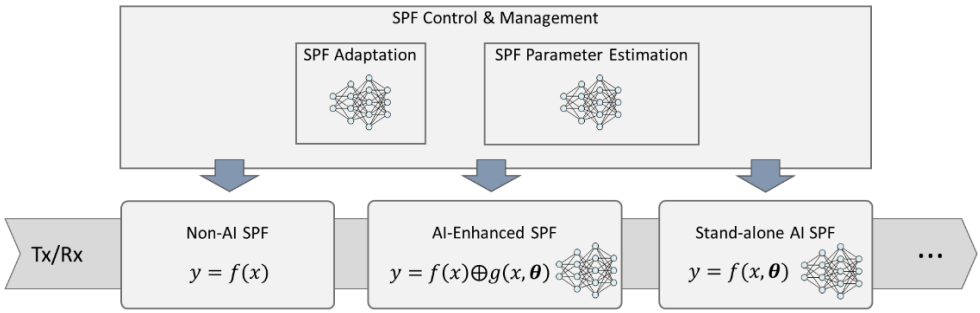


Figure 4-3: Illustration of typical scenarios of employing AI capability in 6G devices for modem signal processing: i) Stand-alone AI, ii) AI-enhanced, iii) Control & Management.

We can discern two primary trends in the integration of AI into PHY signal processing (Figure 4-4). In several physical layer signal processing problems, optimal, cost-effective non-AI solutions (f_b) are available for simplified channel and operational scenarios. However, these models often fail to adequately mirror real channel/operational conditions, leading to suboptimal performance in practical settings. In such cases, data-driven AI approaches can significantly enhance communication performance by accommodating real-world

characteristics. However, the computing costs increase significantly in this scenario compared to the non-AI approach (red line). Conversely, there exist non-AI solutions that offer exceptional communication performance (f_a), accurately representing real channel and operational models. However, even if this is beneficial, the computation complexity of these solutions can exceed what is feasible for certain applications. Here, an AI-driven approach becomes valuable by mitigating solution complexity while almost preserving communication performance (blue line). Both cases are exemplified in the Figure 4-4, where the red and blue trajectories delineate the trade-off between communication performance and complexity costs for the respective AI solutions. The design goal typically revolves around maximizing communication performance, aiming to approach the $f(x,\theta)$ AI-solution denoted by the elliptical region.

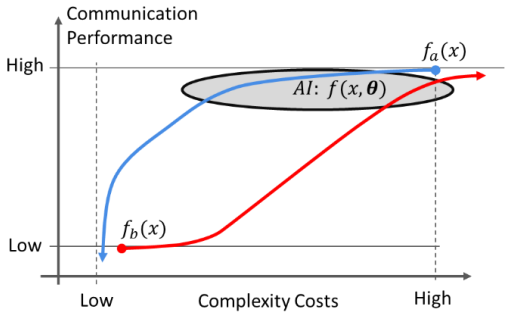


Figure 4-4: Two typical examples of replacing non-AI optimal solutions with an AI approach.

The need for proper selection and investigation of relevant candidate AI-based SPFs is paramount in the context of Hexa-X-II. These insights collectively underscore the critical role of advanced computational resources in enabling the potential of AI-based signal processing in 6G wireless communication.

When optimizing AI acceleration at the physical layer (we assume inference problems in this context and don't consider model training), it's critical to discern the specific application—whether it's tailored for classification or regression tasks. Classification networks often excel with lower-resolution parameters, whereas regression algorithms, aimed at prediction tasks, typically demand high precision for accurate modelling and forecasting of signals. To accommodate both applications within a single SoC, two specialized accelerators optimized for each task or a single versatile architecture supporting both regression and classification algorithms are possible. In this study, we investigate flexible AI accelerator architectures that facilitate dynamic scaling of parameter precision. The primary focus lies on the bit-serial approach, i.e. all calculations and memory accesses are carried out bit-serially by bit-streaming of the operands and results [Fri23]. This enables scaling of the calculation accuracy at bit level. In this approach, the precision of network parameters and computations is initially determined during the design phase based on the worst-case scenario. However, during runtime, the resolution of operands is dynamically configured to meet the specific requirements of tensor operations. This dynamic re-configuration enables highly adaptive computations that are optimized layer-by-layer for deep network architectures.

In addition to variable precision computations, typical AI algorithms exhibit a high degree of inherent parallelism. This characteristic can be leveraged by parallel AI accelerator architectures to effectively handle demanding computation and latency requirements. For instance, Figure 4-5 illustrates the seven algorithmic dimensions of a typical neural network convolutional layer [Fri23]. Conversely, the kernel dimensions 1 and 2 collapse to one with fully connected layers. Hence, to accommodate the flexibility required across various network layers and tensor operations—encompassing convolutional, fully-connected, dilated and transposed - we propose the 3D array HW accelerator architecture (Figure 4-6), primarily linking array dimensions to input channels, output channels, and feature maps of algorithm. However, it's important to note that various other strategies, including algorithm partitioning, folding, fusing, and mapping, are also viable within this framework. Note that this architecture is flexible regarding the configuration of the dimensions, so that 2D and 1D arrangements are also possible in addition to the 3D configuration by statically reducing certain dimensions to 1 at design time or dynamically at runtime.

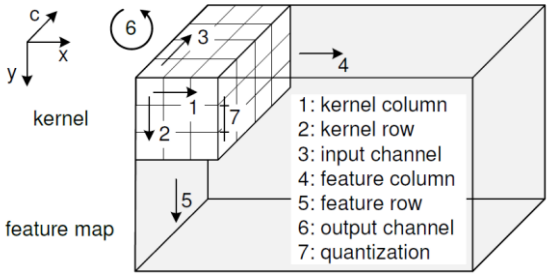


Figure 4-5: Typical algorithmic dimensions of a convolutional layer.

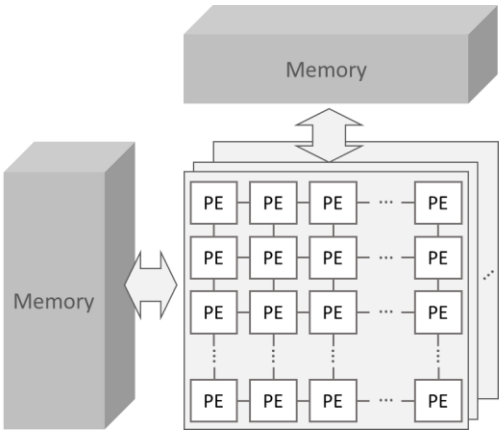


Figure 4-6: Concept of 3D array AI-accelerator architecture. PE represents processing elements that implements AI specific operations e.g. MAD, scaling, bias and activation.

4.1.2 AI accelerator

AI accelerators, known by various aliases such as Neural Processing Unit (NPU), Tensor Processing Unit (TPU), Neural Engine, etc. are components implementing control and processing logic for efficient execution of ML algorithms. For brevity, we will use the abbreviation NPU in following text.

In essence, the majority of ML algorithms inherently leverage a high degree of parallelism and regularity, facilitating effective processing through parallel architectures. Recent advancements have seen the proposal of diverse parallel architectural approaches, each tailored to specific application domains and the demands of training and/or inference capabilities [CXS+20] [RMJ+20] [RMJ+21].

As such, an NPU typically integrates a multitude of parallel processing elements, closely coupled with high-bandwidth shared memory and a robust high-bandwidth interconnect. Architectural distinctions arise in several key aspects:

a) Distribution of computing capacity: This ranges from straightforward memoryless multiple-and-accumulate processing elements to intricate Very Long Instruction Word (VLIW) processors equipped with local cache/memory.
b) Memory Hierarchy: Architectures differ in their approach to memory organization and hierarchy to store input/output features and model data.
c) Control and Data Flow Management: Variances exist in how control and data flows are managed within the architecture.
d) Support of ML models: Depending on application domain, diverse computational and structural AI/ML models are supported

These nuanced differentiators contribute to the overall efficiency and effectiveness of the NPU in executing AI tasks, aligning its design intricacies with the specific requirements of the given use case or application domain.

In this work we introduce initial design of NPU for purpose of on-chip inference. The proposed NPU (Figure 4-7) consists of a lightweight RISC processor core with dedicated program memory, a tightly coupled separate memory, DMAs and the PE array. The integrated processor is an open-source implementation of the Rocket Chip RISC-V core. The memory scheme, in combination with a regular instruction set architecture, enables the utilization of the regularity of fully linked and convolutional layers and results in a low instruction footprint.

For AI acceleration, in accordance with concept in Figure 4-6, NPU implements a 2D array of PEs with vector Multiply-Add (MAD) operation supporting elementary computations of state-of-the-art DNN, CNN and fully connected networks including layer fusion. In addition, piecewise linear activation approximation is supported, enabling programming of a wide range of activation functions. The PEs, computation pipeline and memory architecture implement above mentioned bit-serial processing that enables efficient support for mixed-precision operand computations of inner-products, additions, convolutions, pooling. Furthermore, these advantages also apply to the acceleration of the more general dilated and transposed convolution operations with efficient zero-skipping. The scalable and transferable architecture of the NPU architecture allows integration in platforms with different performance requirements. Thus, a small sized accelerator can be used in small IoT devices while a large one for high performance systems. Moreover, NPU can be configured for low-precision classification and high-precision regression applications, respectively.

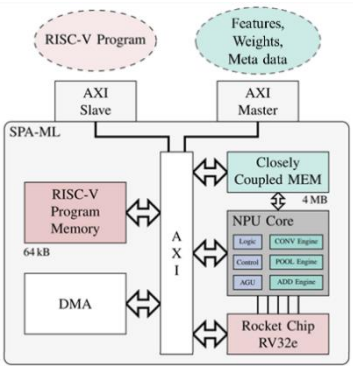


Figure 4-7: Block diagram of proposed AI accelerator.

4.1.3 RISC-V based signal processing

To increase computing efficiency, modern processor architectures leverage data level (DLP) and instruction level (ILP) parallelism. Meanwhile, there is renewed interest in vector processors, which serve as scalable, adaptable and energy-efficient computing platforms. Usually, standard CPUs integrate extensions for vector calculations within their instruction set architectures (ISAs), such as ARM's scalable vector extension (SVE) or RISC-V's "V" vector extension (RVV).

RISC-V has gained a lot of traction due to its open ISA approach. While RISC-V and its RVV primarily target general-purpose and data-intensive applications, recent studies have highlighted their potential as effective solutions for signal processing tasks. Therefore, this research aims to investigate the capability of RISC-V in processing 6G baseband signals. For our initial investigation, we selected a prevalent baseband benchmark algorithm commonly used in HW performance evaluation [SMB+12]. This algorithm represents MIMO SC-FDMA receiver functionalities, encompassing MIMO channel estimation, MIMO combining, demodulation, and decoding. The list of signal processing kernels forming the baseband benchmark is in Table 15 and task graph of benchmark is in Figure 4-10a. In the analysis, the benchmark was set up with 50 resource blocks (RB) with 12 sub-carriers/RB and 1ms subframe comprising 14 symbols, employing a 4x4 MIMO spatial multiplexing alongside QAM modulation. In addition, the chosen HW platform for experimentation was the open-source RISC-V RVV processor ARA [CSZ+20, PCW+22], configured with a 4-lane vector extension and a vector length of 4096 bits. The ARA scalar and vector cores operate at clock frequencies of 1.25 GHz and 1.0 GHz, respectively. Each ARA-lane supports two 64-bit input operands, leading to two 32x32-bit and four 16x16-bit multiplications per cycle, respectively. The aggregated computation performance of four lanes per cycle thus results in eight 32x32-bit and sixteen 16x16-bit multiplications. It's important to note that most computations in benchmark kernels were implemented using 16-bit precision, resulting in a 2x16-bit

configuration for a 32-bit precision in a single complex value. The baseband signal processing kernels were initially implemented for scalar computation and later vectorized to efficiently leverage ARA RVVs.

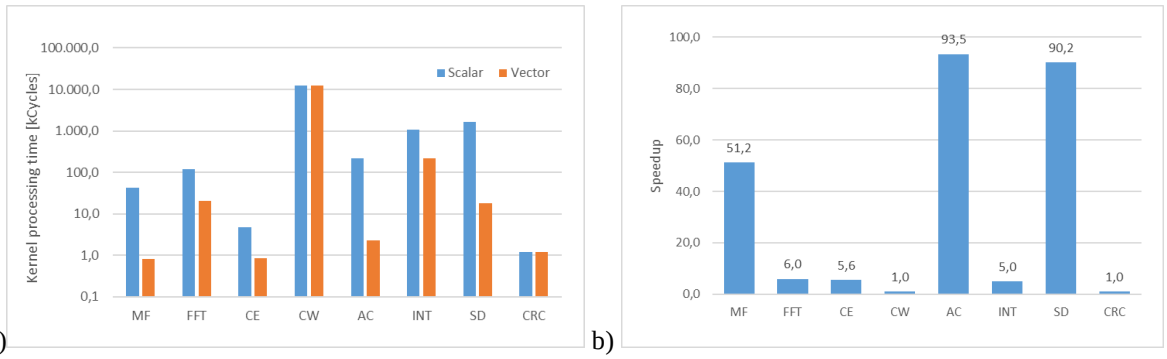
As first, the performance impact of signal processing kernel vectorization and vector processing was examined. Figure 4-8a depicts processing time in cycles for both the scalar and vectorized kernels. Additionally, the Figure 4-8b illustrates the vectorization speedup achieved by individual kernels. It's apparent that most kernels benefit from vectorization. Note, that combiner weight computation kernel (CW) employing matrix inversion using Cholesky decomposition was implemented in scalar version only and hence doesn't demonstrate any improvements from vectorization. CW optimization will be subject of future work. Additionally, the estimated speedup of the vectorized 16-bit 1k-FFT is six, based on scaling factor of the results in [Viz23].

Secondly, the overall baseband algorithm benchmark in Figure 4-10a was analysed in Figure 4-9, which represents the receiver signal processing of a single subframe (SF). In addition, the execution count of signal processing kernels within benchmark (kernel frequency/SF) is depicted in Figure 4-10. Figure 4-11 and Figure 4-12 breaks down the overall benchmark into execution time cycles, milliseconds and percentage for individual kernels. Note that even in this scenario the performance of scalar CW version doesn't benefits from RVV. It's evident that CW dominates the overall performance (this can be expected even after CW vectorization). To disregard the impact of CW on overall performance, Figure 4-13 illustrates the breakdown of the benchmark without CW. It is clearly visible that the vectorized benchmark approaches an execution time of 2 ms, which using 2 RISC-V cores underlines the real-time capability ~1ms/SF of this use case.

This initial analysis serves as an indication of the RISC-V capability in terms of baseband signal processing. In addition, the results can be projected to other use cases and used as such to define minimum SoC requirements in terms of computing power. Furthermore, the critical functions for which special accelerators should be developed can be identified.

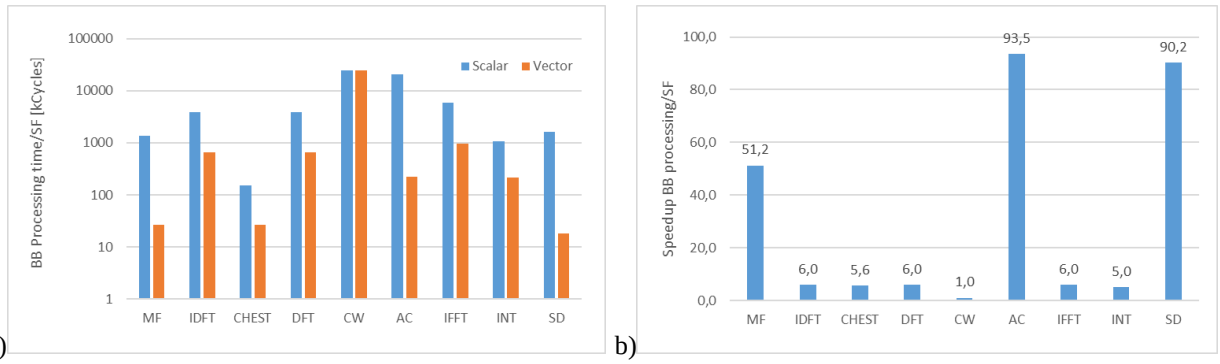
Table 15: List of baseband signal processing kernels forming the core of the baseband benchmark [SMB+12].

Baseband Kernel	Abbreviation
Matched Filter	MF
Discrete Fourier Transform	IDFT/DFT
Channel Estimation	CHEST
Combiner Weight Estimation	CW
Antenna Combiner	AC
Inverse Fast Fourier Transform	IFFT
Interleaver	INT
Soft Demap	SD
Channel Decoder	TC
CRC	CRC



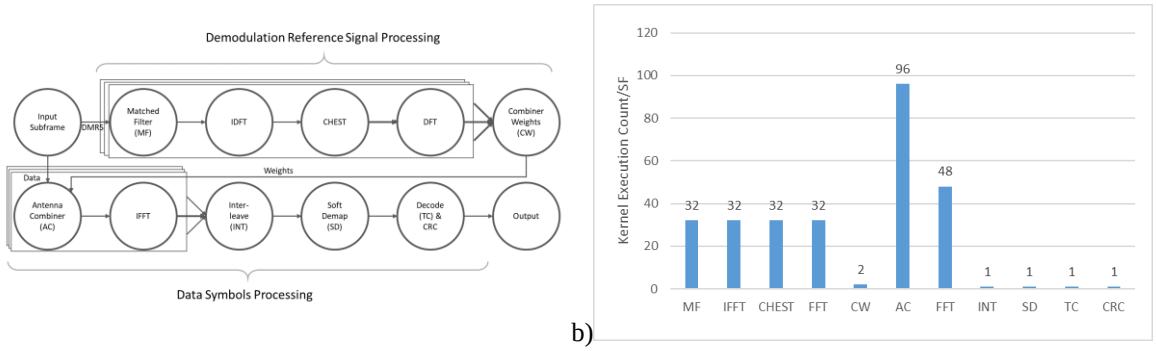
a) b)

Figure 4-8: Processing time of baseband kernels [SMB+12] on RISC-V processor w/o and w/ vector extensions a) and speedup of vectorized kernels b). Note that channel decoder is not subject of vectorization on RISC-V processor.



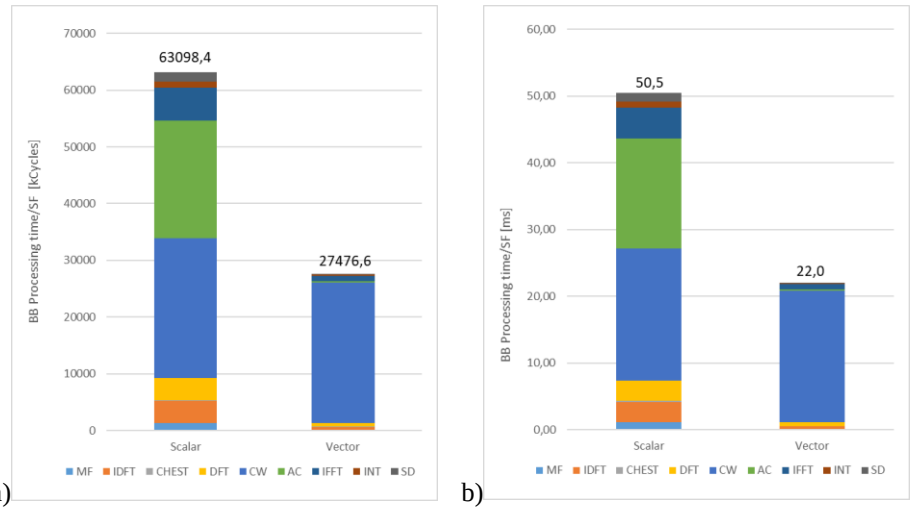
a) b)

Figure 4-9: Processing time of baseband algorithms within one subframe [SMB+12] on RISC-V processor w/o and w/ vector extensions a) and speedup of vectorized algorithms b).



a) b)

Figure 4-10: Task graph of receiver signal processing benchmark and baseband kernel execution count within one subframe of benchmark with 14 symbols and 50 RBs employing 4x4 MIMO spatial multiplexing and QAM.



a) b)

Figure 4-11: Processing time in cycles a) and ms b) of baseband algorithms within one subframe [SMB+12] on RISC-V processor w/o and w/ vector extensions.

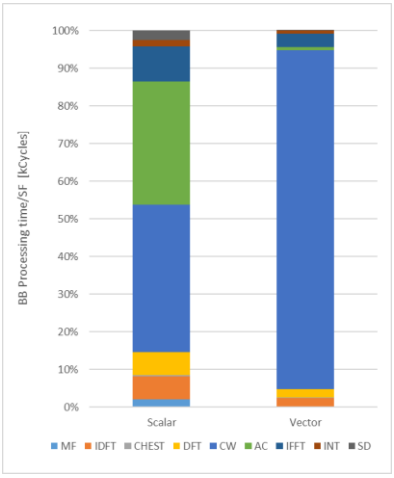
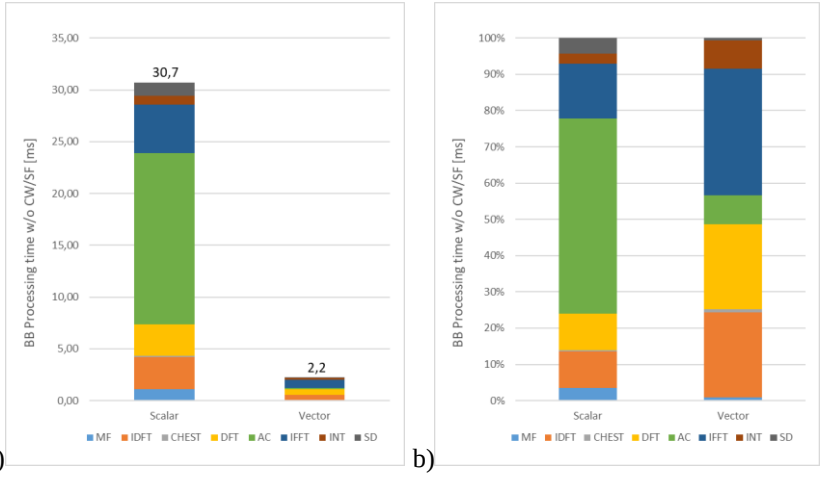


Figure 4-12: Breakdown of the processing time distribution in cycles of baseband algorithms within one subframe [SMB+12] on RISC-V processor w/o and w/ vector extensions.



a) b)

Figure 4-13: Processing time in cycles a) and ms b) of baseband algorithms w/o combiner weight computation within one subframe [SMB+12] on RISC-V processor w/o and w/ vector extensions.

4.2 Secure and scalable 6G SoC design

New 6G mobile networks open up unique features and opportunities for smart systems and communication devices. Generational technology upgrades provide unprecedented data rates and processing power. At the same time, these new platforms must address the growing security and privacy requirements. This poses two main challenges concerning the digital processing hardware:

1. We need to provide integrated trustworthiness covering hardware, runtime, and the operating system (OS). Whereas integrated means that the hardware must be the basis to support secure runtime and operating system needs.
2. 6G applications cover a wide range of requirements in terms of performance and energy consumption. This requires a scalable hardware solution to cover differing needs in terms of processing resource requirements.

4.2.1 Security architecture

In Hexa-X-II, we are working on a secure and scalable SoC architecture tailored to a microkernel-based OS, that is designed to deal with the mentioned security and privacy challenges. The approach is based on a tiled architecture with an example depicted in Figure 4-14. Physically separated tiles are connected by a network-on-chip (NoC) and contain computational logic (processing cores, accelerators), memory (SRAM), or interfaces to external resources (DRAM, Ethernet). Furthermore, each tile includes a hardware component called *trusted communication unit* (TCU), which isolates all tiles from each other so that no communication is possible by default. The OS microkernel runs on a dedicated tile and configures the TCUs. The OS manages communication channels between the tiles while the TCUs enforce them in hardware. The platform including the OS is called M³ (Microkernel-based System for Heterogeneous Manycores).

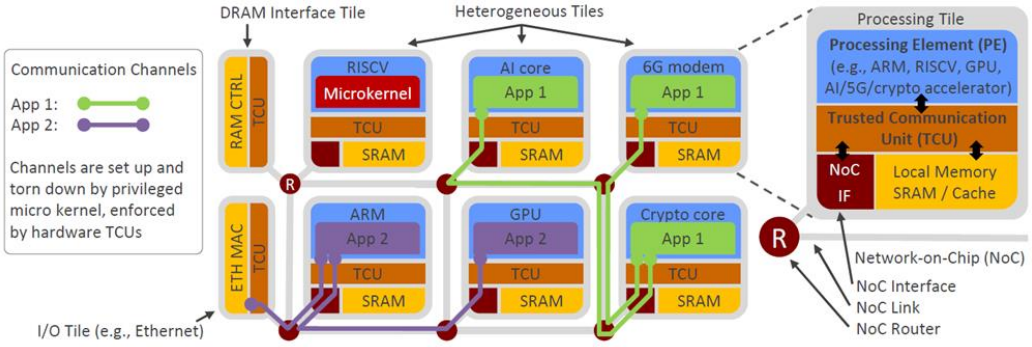


Figure 4-14: General architectural concept of a secure and scalable SoC architecture [PHH23].

Furthermore, the TCU implements a set of commonly used communication primitives such as message passing and remote direct memory access (RDMA). Source and destination address for such primitives are set by the microkernel and strictly enforced by the TCU. If permitted by the microkernel, an application can span over multiple different processors, accelerators, and interfaces without putting unrelated applications running on other tiles at risk. Applications form isolated islands of different processing resources. The TCU interfaces with the local NoC interface (NoCIF), the memory, and the PEs of the tile. The NoCIF provides access to the NoC which allows communication with other system components on the chip. The microkernel sets up and releases communication channels with other components via the NoCIF by configuring the TCU's internal registers.

As an example, Figure 4-14 shows two isolated applications (App 1, App 2). App 1 uses the AI core that sends results via a channel to the crypto core, where data is encrypted, finally encrypted results are channelled to the 6G modem tile and sent to a receiver. App 2 uses an ARM core for application logic, it has been given access to the Ethernet interface and to the GPU tile. Each application only needs to rely and trust those components, that it actually uses. This means, a malicious component that is not part of the application "island", cannot interfere or corrupt our application, because there is simply no way to access it.

4.2.2 Integrating a general-purpose core

One tile must include a general-purpose processor that runs the OS microkernel which configures the TCUs and hence manages communication between tiles. For that purpose, we implemented a tile with a general-purpose processor [HA22]. We used a RISC-V Rocket core [Asa16] with caches, which is available as open-source and can be configured (e.g., instruction set, cache sizes, debug features).

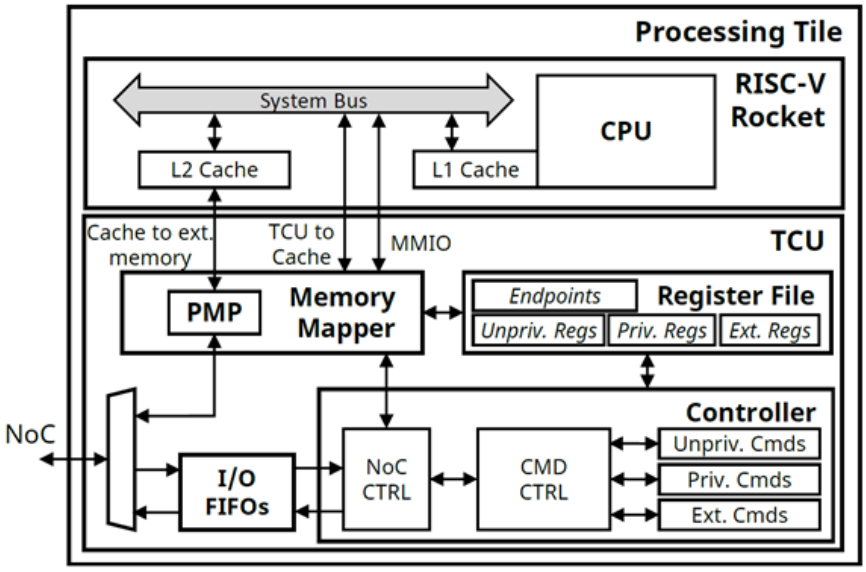


Figure 4-15: Processing tile with TCU and RISC-V Rocket core [HA22].

Figure 4-15 shows a tile that integrates the TCU and the Rocket core. The core communicates with the TCU via memory-mapped I/O (MMIO). The TCU has a tightly coupled memory interface to access the internal caches of the core via its bus system. Other cores without caches or simple processing units with only scratch-pad memory can be connected in the same way. Interrupt signals of the core (not shown in Figure 4-15) are also connected to the TCU. Hence, the TCU can trigger the core at certain events such as message reception. Since the TCU interconnects the NoC and the logic within the tile, it provides a uniform interface to other tiles that simplifies the management and collaboration of heterogeneous tiles.

The TCU requires an interface to the NoC as well as an interface to local resources of the tile. As depicted in Figure 4-15, our hardware implementation of the TCU contains three main blocks which cover the main functionalities: controller, register file, and memory mapper. The TCU controller implements finite state machines to executes commands as requested by the core or from the NoC. The register file includes registers such as endpoint registers (briefly called endpoints) to store access permissions. The memory mapper can multiplex accesses from the core to the local memory and to the register file (MMIO). Additionally, the physical memory protection (PMP) block forwards memory requests of the core to the NoC. These requests are validated according to access permissions stored in dedicated endpoints.

To enable usage and configuration of the endpoints, the TCU commands and registers can be accessed by three interfaces: unprivileged, privileged, and external interface. The unprivileged interface enables commands for DMA transfers (read, write) and message passing (send, receive, reply). During command execution, the TCU checks the related endpoints and, if allowed, performs the data transfer to the target tile. A command is finished when a read response or a message acknowledgment was received. The privileged interface is only accessible by privileged software running on the core and enables support for virtual memory and context-switching in general-purpose cores. The privileged software uses the MMU of the core to ensure that unprivileged software can only access MMIO addresses of the unprivileged interface. To support virtual addressing, the TCU holds a software-loaded translation lookaside buffer (TLB) to store recent address translations. The external interface

is only used by external tiles (e.g., the tile running the microkernel) to configure endpoints and to enable features of the privileged interface.

4.2.2.1 Implementation results

As preparatory work for T5.3, we already implemented the presented SoC architecture on a Xilinx Virtex UltraScale+ FPGA (VCU118 board) [AHW+22]. On the FPGA, there are eight processing tiles with a single RISC-V Rocket core each. For further evaluation possibilities we can reconfigure the platform and integrate RISC-V BOOM cores [ZKG+20] which are the out-of-order variant of Rocket. Furthermore, there are two memory tiles with interfaces to external DDR4 DRAM, and a dedicated Ethernet tile with a hardware UDP/IP stack implementation for configuration and debugging purposes. All tiles are connected by a 2x2 star-mesh NoC topology with four routers. Even though the scalable NoC architecture would allow to build a larger design with more NoC routers and tiles, we are limited by the given resources of the FPGA.

Table 16 shows the consumed FPGA resources of the major components in our hardware platform. The presented TCU configuration includes all extensions and corresponds to the implementation in the processing tile including the RISC-V core. In comparison to the BOOM and Rocket core, the TCU only requires 10.6% and 32.6% of the FPGA logic (Look-up tables, LUTs), respectively. Since the TCU contains no memory or caches, the number of required BRAMs is negligible compared to the cores. This is an advantage especially when considering a real chip implementation where memory consumes a substantial part of the area.

Table 16: FPGA area consumption: Logic and LUT-RAM (LUTs), registers (Flip-flops, FFs), block RAM (BRAM) with 36 kbit per block [AHW+22].

	LUTs [k]	FFs [k]	BRAMs
RISC-V BOOM core	143.8	71.8	159
RISC-V Rocket core	46.6	22.0	152
NoC router	3.4	2.2	0
TCU	15.2	5.8	0.5
Controller	10.3	3.3	0.5
NoC CTRL	3.2	1.5	0
CMD CTRL	7.1	2.8	0.5
Unprivileged commands	6.2	2.5	0.5
Privileged commands	0.9	0.3	0
Register File	2.0	1.0	0
Memory mapper + PMP	0.6	0.2	0
I/O FIFOs	2.3	0.3	0

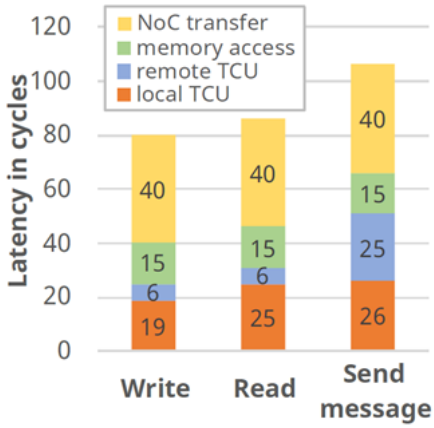


Figure 4-16: Latency of TCU commands [HA22].

We measure the latency of the TCU to evaluate its timing overhead when executing commands initiated by the core. Throughput is limited by the bandwidth of the NoC (16 bytes/cycle) since the current TCU implementation supports the full NoC bandwidth. For these experiments, we use two identical processing tiles each including a TCU and a RISC-V core. The tile which initiates commands contains the so-called *local* TCU, while the receiving tile contains the *remote* TCU. The tiles are connected to a single NoC router to minimize the delay induced by the NoC. Figure 4-16 shows the measurement results of a DMA-write and read as well as for sending and receiving a message (with 8-byte payload data each). For each transfer, the local TCU requires 11 cycles to initiate the command by reading the corresponding endpoint and validating the access permissions. The TCU also performs a TLB lookup which requires at least 4 cycles if the associated TLB entry is found immediately. If more TLB entries have to be scanned, a higher number of cycles is needed. If the privileged interface (virtual memory support) of the TCU is disabled, the TLB access is skipped and the latency of the command is reduced accordingly. The remaining cycles of the local TCU are spent to prepare the access to the RISC-V cache and to finish the command as soon the response/acknowledgment packet has been received. The time of a memory access of the RISC-V is specific to this core and depends on its internal cache access times. For the write and read commands, the remote TCU requires 6 cycles to forward the requests to the memory. For message passing, the remote TCU takes about 25 cycles on average. This includes finding a free slot in the memory-mapped receive buffer, setting up an endpoint for a reply, and informing the software in case the received message belongs to a currently paused application on the core (only if the privileged interface is enabled). The NoC transfer delay is about 20 cycles per packet via the single router, which is mainly caused by the synchronization registers in the NoC interface. Asynchronous transitions enable to set different clock frequencies of the NoC and the tile. Since each command consists of a data and a response/acknowledgment packet, we measured 40 cycles for the total transfer.

In summary, the latency introduced by the TCU functionalities and security features is not significant. Memory read delays to the RISC-V caches and packet transmission times via NoC are in the same range or higher. Furthermore, processor-intern routines of the RISC-V (e.g., interrupt handling) typically show up to two magnitudes higher latencies.

4.2.3 Integrating an accelerator

The above-described architecture and implementation has already been developed. These results are the basis to take the next step and to tailor the platform towards 6G applications. In particular, in task T5.3, we are working on the integration of accelerators to improve performance and energy efficiency of, e.g., cryptographic and ML algorithms. In particular, we will integrate TUD's ML (AI) accelerator as described in Section 4.1 into the secure and scalable SoC platform. In this report D5.3, the initial design will be presented. Ongoing progress and the final design will be presented in D5.4 and D5.5, respectively.

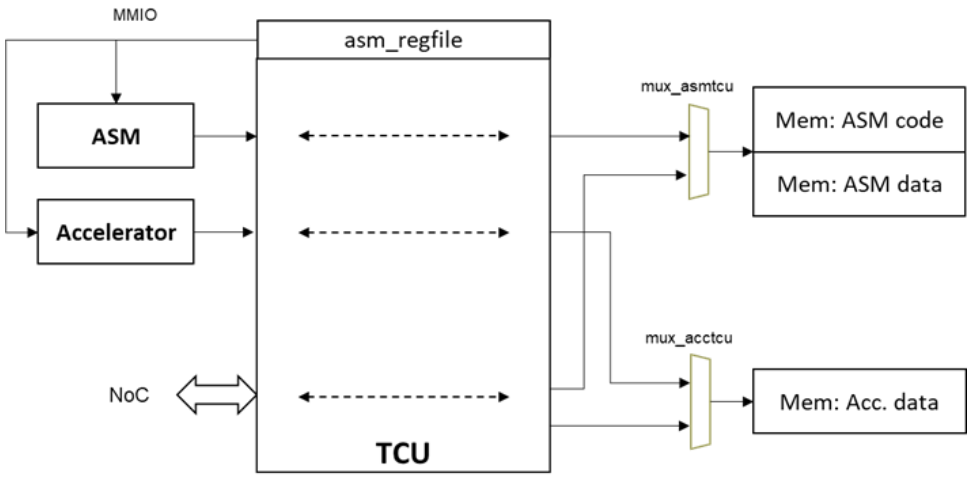


Figure 4-17: Initial design of a processing tile with TCU, hardware accelerator, and accelerator support module (ASM).

In contrast to the integration of general-purpose cores as described above, the integration of accelerators into the tiled architecture adds further challenges.

1. The hardware accelerator is primarily built to support high performances and thus requires high throughputs to memory and data transfers.
2. Usually, a hardware accelerator has no or only limited programmability. Hence, it may not autonomously use endpoints of the TCU to access external memory or communicate with other tiles.

Throughput of the accelerator is limited by the NoC bandwidth of the platform. To fully use the bandwidth, local memory should be available in the processing tile. This allows to apply advanced features to load data from memory (e.g., double buffering, stream processing). As depicted in Figure 4-17, there is a local memory for the data to be processed by the accelerator, which can be accessed through the TCU.

The idea to enable inter-tile communication of the hardware accelerator is to add another component to the tile, called accelerator support module (ASM). As depicted in Figure 4-17, the ASM will be integrated next to accelerator and should allow the following tasks:

- Configuring the accelerator by setting basic properties like algorithmic-specific attributes or address mapping.
- Load data from/to external memory to/from the local accelerator memory.
- Support features of the OS like process communication, file system accesses, and system calls.

The ASM has access to the accelerator via MMIO. The ASM can be a general-purpose core or a custom hardware component. To allow high flexibility, our initial design will feature an ASM as a general-purpose core with access to local memory through the TCU. Later, the general-purpose core can be replaced by a custom hardware implementation to decrease costs (chip area, power consumption).

4.3 Multi-source EH and power management

The core of energy harvesting (EH) lies in extracting energy from ambient sources to enable perpetual operation of IoT devices. Harvesting energy from ambient sources involves primarily two steps, converting ambient energy into electrical energy using a transducer and then regulating, storing, and managing generated electrical energy. Transducer-generated electrical energy often has voltage levels in milli- or microvolt scale which necessitates an intermediate system capable of upscaling the generated voltage. This system, called the Power Management Integrated Circuit (PMIC) serves as the connecting point between the energy source and load, facilitating optimized extraction of energy from the source and storing and releasing harvested energy to the load as demanded. The primary tasks of a PMIC can be identified as:

1. Extract millivolt output generated by transducers and boost it to a voltage range usable for the electronics and sensors.
2. Act according to the electrical characteristics of the energy generator to ensure optimum harvesting.
3. Store the harvested energy in an energy buffer such as super capacitor and release it to the load.
4. Monitor the energy level of the energy buffer and generate warning signals if the energy in the buffer goes below the minimum operating range.
5. Should not add any or minimum energy overhead to the overall system design

The technology behind PMICs is well matured and many off-the-shelf solutions are available in the market. There are several players in the market, both established and startups who are actively producing EH PMICs. Nonetheless, a major drawback of these PMICs is that they are tailored for a specific source of energy and can work only with this specific source. Thus, if a system design requires harnessing energy from multiple ambient sources, it must incorporate a PMIC for each source, resulting in an inflated Bill of Materials (BoM) and form factor of the device. Even then, simultaneous exploitation of the sources is highly improbable as all these PMICs are independent and there is no system that can orchestrate the synchronized operation of the PMICs.

With the proliferation of EH and energy neutral (EN) devices, there is a demand for PMICs that can seamlessly work with multiple sources and harness energy simultaneously from all the sources. Nonetheless, the technical hurdles in building a multisource EH PMIC are many. A major technical challenge is the unique electrical characteristics of each energy converter. For instance, a solar and thermal energy generator (TEG) will generate DC voltage and current whereas an RF or Piezo generator will produce AC output. Moreover, a solar cell has impedance that varies with the incident radiation, whereas a TEG is a low impedance system with relatively constant impedance over temperature ranges. Thus, an MPPT algorithm for solar will not work for TEG or RF harvester. Therefore, building a multisource energy harvester would require PMICs with power front end that can detect the type of source and adapt to electrical characteristics of the source.

Assuming we managed to harvest energy from all the sources, the next question is how we combine all these energies and charge a single energy buffer. In a single source harvester, there is only one source and one buffer. Therefore, the source has exclusive access to the buffer and can be accessed at any time. However, when there is more than one source and each source is harvesting energy, exclusive access to the buffer is not guaranteed anymore. This can create serious issues such as reverse current flow and energy losses when more than one harvester tries to access the source. One solution to this problem is to allow the harvesters to access the buffer in a time multiplexed manner. But this method will turn inefficient as the number of sources increases. A better approach is to combine energy flow paths between the sources and the buffer and create a single energy flow path to the buffer. Combing energy from all the sources requires an intermediate circuit that accepts multiple inputs, combines all the inputs and creates a single output to the buffer. Such a circuit, usually called the Energy Combiner (EC) facilitates efficient merging of energy from all the sources simultaneously.

In recent days, attempts have been made to run low-power ML algorithms on EN devices. Furthermore, energy prediction and energy-aware algorithms are increasingly becoming popular. Thus, EN systems are moving beyond the horizon, and they are not any more confined to only sense and transmit applications. Naturally, there is a demand for advanced PMICs that can provide more insights into EH and consumption, not just harvesting and energy management. For instance, energy-aware schedulers need to know EH rate and load consumptions in real time so that they can allocate tasks based on the availability of energy. EN designs with reconfigurable energy buffers require task wise energy consumption of the load so that they can allocate energy buffers in tandem with the load demands. Moreover, any hardware or software-based approach requires an accurate estimation of the Relative State of Charge (RSoC) of the energy buffer, which is usually a supercapacitor. One might assume that the RSoC estimation of the supercapacitor is relatively easy considering the fact that the energy stored in a capacitor is proportional to its voltage. However, like rechargeable batteries, the validity of Peukert's law has been confirmed for supercapacitors as well. Consequently, the RSoC of supercapacitors is impacted by many factors such as charging and discharging currents, bias voltages, etc. Therefore, RSoC estimation based on voltage level alone would produce incorrect values. Hence, an advanced PMIC would be able to track the RSoC of the buffer, measure EH rate, track load consumption and allow many sources to be connected.

4.3.1 InfiniteEn: A multisource EH architecture for EN devices

InfiniteEn is a power management system proposed for EN devices [PW23]. InfiniteEn boasts many features that are not available with any so far available PMIC in the market. The features include multisource harvesting, EC, harvest rate sensing, online capacity estimation of the supercapacitor and an ultra-low power Load Monitoring Module (LMM). In addition, a reconfigurable storage architecture allows InfiniteEn to configure its storage capacity on the run. Thus, the capacity of the energy buffer can be adjusted based on harvestable energy or the load's demand. A high-level block diagram of InfiniteEn is shown in Figure 4-18a and the circuit fabricated on a PCB is shown in Figure 4-18b.

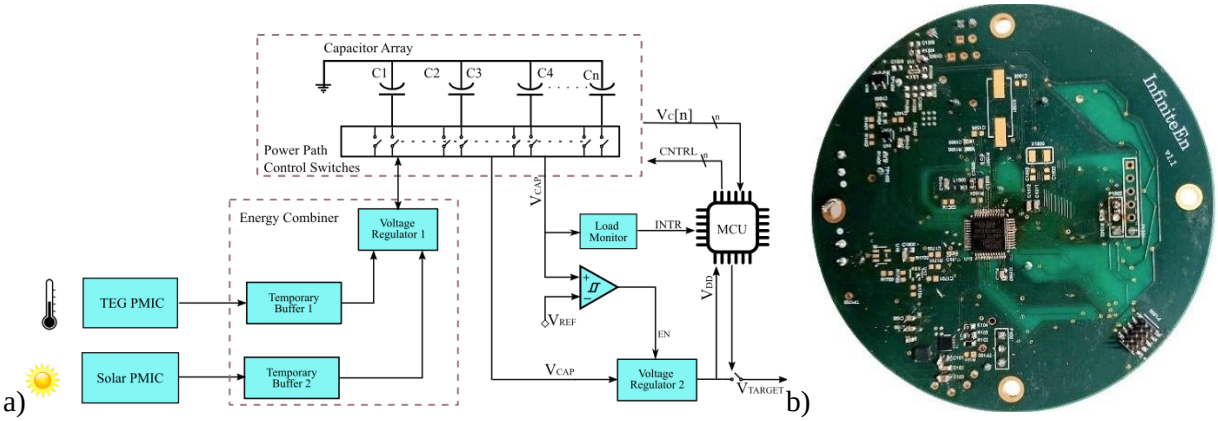
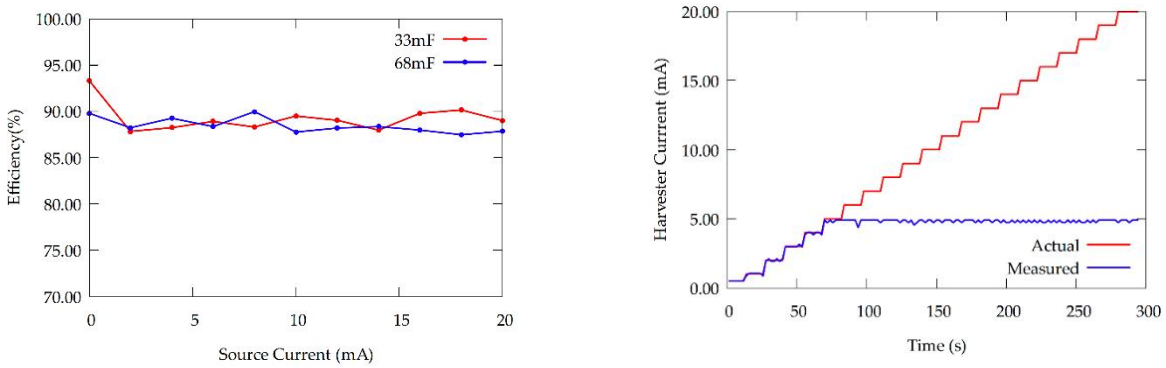


Figure 4-18: InfiniteEn (a) high level architecture of the system (b) complete system fabricated on a 50mm diameter PCB.

An EC that uses the concept of *capacitor-to-capacitor forms* the heart of InfiniteEn. The EC, instead of directly charging the supercapacitor, uses intermediate energy buffer capacitors) of relatively smaller capacity to store the energy before it gets transferred to the supercapacitor. Then, a switching regulator is used to transfer energy from these temporary buffers to the supercapacitors, i.e., by discharging the buffer capacitors. A lower and upper threshold voltages are chosen for each temporary buffer. Every time the upper threshold is reached the temporary buffer is discharged to the main buffer via the switching regulator until the lower threshold is hit. Each source has been assigned a pair of temporary buffers to ensure the source is not left unattended at any time. Using temporary buffers allows EC to effectively shadow all the sources from the supercapacitor and to create only one charging path to the supercapacitor. Moreover, since the buffers are connected through the switching regulator, reverse energy flow never happens. With two energy sources connected, the EC achieves more than 88% efficiency in transferring energy from the harvester to the main buffer. The efficiency of the EC when harvesting energy from two different sources is illustrated in Figure 4-19a.

A remarkable feature of the EC is that it can estimate the harvesting rate and the amount of energy stored in the buffer, making it the best choice for energy aware applications. As the temporary buffers always charge and discharge between two predefined threshold levels, each charging/discharging corresponds to a fixed quantum of energy. Thus, by tracking this energy EC can estimate the total amount of energy transferred to a supercapacitor. Moreover, the number of energy transfers in a fixed slot can be used to estimate the harvesting rate. An energy prediction algorithm or scheduler can take advantage of these features to improve their performance. The harvesting rate sensed by the EC for different harvesting conditions is shown in Figure 4-19b.



a b

Figure 4-19: Evaluation of EC (a) efficiency of energy combing when harvest from multiple inputs (b) harvest rate sensed with input current to the switching regulator limited to 5 mA.

Often energy-aware real time schedulers on EN devices require task-wise energy consumption of the load so that the energy-aware scheduler can optimize the execution of any device task. On the other hand, reconfigurable storage needs this information to optimize the storage buffer. An EN device with reconfigurable storage allows the system to adjust the capacity of its storage buffer in response to different energy harvesting and load conditions. This helps the device to choose an optimum energy buffer based on the energy harvesting and load demands. Conventional designs hardcode task-wise energy consumption values in the code and are used during the run time. Thus, any change in the firmware structure would require energy profiling the code. The LMM, a novel concept incorporated with InfiniteEn allows tracking the energy state of the load on the run. The module measures the discharge rate of the energy buffer and converts the discharge rate into energy. Moreover, by tracking the discharge rare, LMM can determine the start and end time of each task. The LMM adds minimum components and negligible energy overhead to the system. The response generated by LMM for different load currents is show in Figure 4-20 and the corresponding load current measured by the LMM is shown in Figure 4-21.

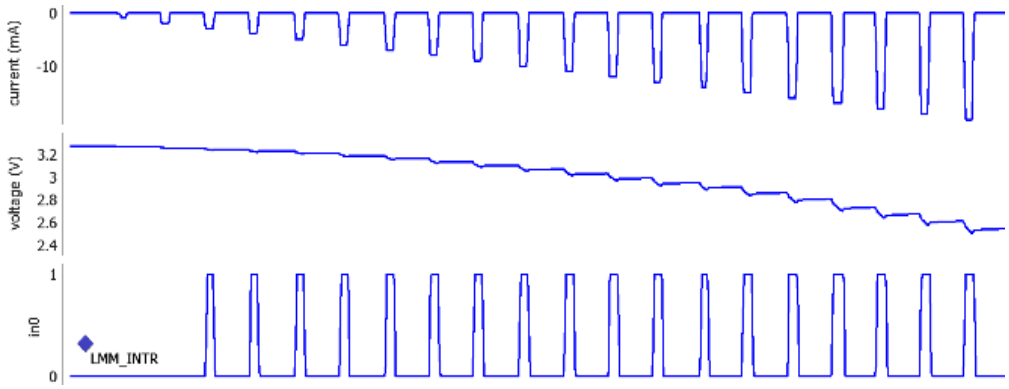


Figure 4-20: Response generated by LMM for different discharge rates.

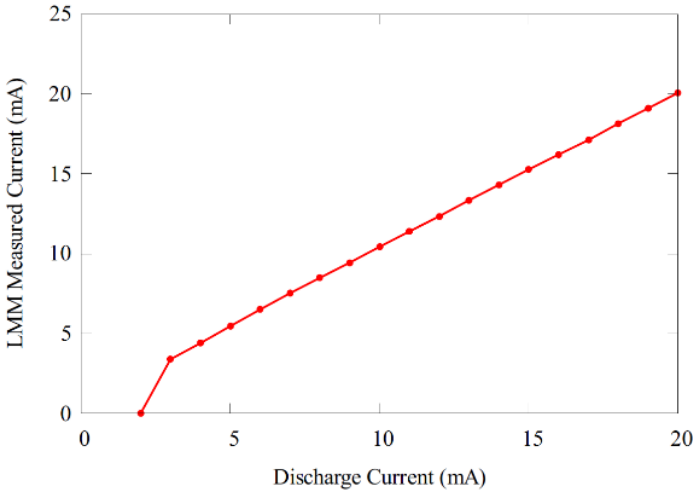


Figure 4-21: Comparison of actual discharge current and discharge current measured by LMM.

InfiniteEn is a first step towards making the power management units for EN devices intelligent and energy aware. Nonetheless, InfiniteEn is still in its early stage of development and requires many more iterations before it can be confidently employed in real-life applications. Currently, the power interface and MPPT are managed by off-the-shelf components, thus requiring a separate interface unit for each source. More research must be invested on making a unified power interface section that can work seamlessly with any source.

5 Ultra-low cost/power devices

Nowadays, numerous terminals are interconnected over the Internet using machine-type communication. In many cases, the overall energy usage of these terminals needs to stay low to support viable use cases. This is not only for keeping a small form factor and the overall cost of the end devices low but also because these devices rely usually solely on a battery, if any, and are required to be of low or no maintenance at all during their lifetime (which can be 10+ years). In the case of rechargeable batteries, keeping the device's overall energy consumption low will reduce the batteries' charge cycles and may ensure its long lifetime. In addition, considering devices with limited or no storage capabilities, or where energy supply is not guaranteed all the time, such as the envisaged energy-neutral (EN) devices using EH, the instantaneous or short-term average power consumption may also be highly constrained. In this section, we first explore energy usage and some KPIs of 6G cellular ultra-low cost/power IoT devices, with a focus on modem contributions. Specifically, we outline principles for reducing power consumption, delve into legacy cellular IoT energy-saving mechanisms, and propose an evaluation methodology. Then, key enabling technologies are discussed in more detail, including EH corresponding to different energy sources, radio frequency (RF) wireless power transfer (WPT), energy-aware and low-overhead/cost protocols, TinyML, and intelligent wake-up. Finally, via a proof of concept (PoC), we demonstrate the feasibility of using ambient backscattering communication (AmBC) in the cellular downlink, specifically utilizing LTE downlink reference signals, to support EN devices in modern and future applications. Through the discussions, we highlight key challenges and offer hints on future research directions.

5.1 Energy, cost, and performance trade-offs

5.1.1 Power consumption analysis for 6G IoT devices

To understand the devices' energy usage and assess energy-related KPIs such as battery lifetime depending on differently configured use cases, and trade-offs with performance KPIs such as throughput and latency, it is crucial to analyse the energy consumption contribution of the connectivity part of the device (i.e., modem) to the total energy consumption as it is commonly the dominant. Herein, we focus on cellular IoT devices for 6G, primarily covering the massive IoT range of applications, as described in Hexa-X-II deliverable D5.2 [HEX223-D52]. Thus, the target is to analyse the power consumption of such prospective modem devices for 6G IoT, to help identify energy saving opportunities for increasing their lifetime, reducing their operational cost, therefore enabling viable ultra-low power devices and respective new IoT use cases. To do this, it is planned to design parameterized yet realistic power consumption models for devices targeting cellular IoT, in order to: i) obtain insight for evolution of power saving mechanisms, and ii) enable battery lifetime KPI assessment of legacy and potential future modem that could be used in 6G use cases involving IoT devices [HEX223-D12].

In the following, we first present some basic principles for low power consumption and saving energy in modem device. Next, we give a brief background of the main standardised power saving modes that drive the energy efficiency in legacy cellular IoT devices currently in market. Then, we describe the current initial methodology directions for designing a flexible power consumption evaluation model. Finally, we report the current progress and planned next steps, and present future power saving enhancing aspects of potential interest to consider in device power model design.

5.1.1.1 Basic principles for reducing device power consumption

Device modem architecture includes several blocks that consume energy, such as antennas, RF front-end and transceiver circuitry for amplification, conversion, and transmitting/receiving of analogue signals, baseband (BB) unit for various processing tasks of digital signals, memories for data storage and access during operation, clock circuitry for internal synchronization of device system operations, processing units for managing modem's operations and control functions, as well as various interfaces transferring data to/from the modem-external world of the device.

Generally, each part from such blocks in the device is connected to a voltage source (also known as power rail), or by a set of rails, and when a device's function includes activity of this part, it draws power. Thus,

power-efficient semiconductor hardware components and architectural design will keep overall device power consumption low even when the respective parts are active. It is also apparent that, within device's complex architecture and functioning, it can be very efficient for different parts to be powered according to usage needs. For example, when the modem is not actively communicating with the network and just processes information or monitors control signals, there is the potential to keep its RF integrated circuit (RFIC) off. Therefore, various sleep modes can be considered depending on which blocks can stay inactive. At the same time, one must consider the trade-off energy effort needed (and duration constraints) to power-off and wake-up specific circuitry and blocks. A dedicated power management processing unit within the device can take the role to adjust operation (voltage and clocking) of parts based on workload, trade-offs, and constraints, and optimize power consumption during different usage scenarios.

5.1.1.2 Device energy saving in legacy cellular IoT

Together with advanced low-power hardware and efficient device system operation, optimization of communication protocols is highly important to ensure that the device is requested to perform tasks only when actually needed. Generally, in legacy low-capability low-power IoT devices, the main sources of device energy consumption include operation for transmitting, receiving, and processing data and control information. For this reason, 3GPP standardized several connection states, power saving modes, and features, to fit different scenarios of operation and allow IoT devices to enter into deep-sleep modes fast and for long periods, as well as to reduce unnecessary downlink monitoring, signalling, or uplink transmissions. The most important specified concepts include [21.914][21.915][21.916][21.917]:

- **Connected discontinuous reception** (cDRX) mechanism targets savings during *connected state* (where device is actively engaged in user data communication or measurements) and involves network-controlled configuration of periodic receiver ON duration (few ms to monitor scheduled messages) and sleep cycle (several ms where the modem can sleep), targeting more scenarios with regular but unpredictable traffic patterns.
- **Idle DRX** (iDRX) mechanism in *idle state* (where device has no immediate data communication expectations but needs to actively monitor signalling), includes paging occasions (of few ms) within periodic paging cycles (of few seconds), i.e., small wake-up time periods with an active receive period where device listens to control channel for potential incoming data notifications or grants from the network, and is fit for periodic and relatively frequent traffic. Idle state can be generally triggered by the network with a configured inactivity timer. Additionally, the **Release Assistance Indication** (RAI) feature allows the device to bypass C-DRX and move to idle state, by indicating no scheduled uplink data and no expecting downlink data.
- **Extended DRX** (eDRX) mode is further introduced to address even lower traffic and "static" (in terms of periodic traffic) applications for devices in idle state, still needing to be reachable by the network, by allowing flexible configuration of small paging windows of device monitoring activity and high sleep cycle durations between those windows (several minutes or hours).
- **Power saving mode** (PSM) mechanism enables the case of minimal signalling with the network, fit for a device with no traffic most of the time. In that case, the device is still registered to the network for reasonably low wake-up and synchronization time yet at a dormant (deep-sleep) state and not reachable apart for a time window after a wake-up and transmission. Transmission occurrences can be mobile originated (i.e., triggered by device's application layer when there is something planned to transmit) or triggered by *area update* procedures, configured by the network for periodic tracking and locating of devices.
- **Power Class** (PC) device options are defined, to allow solutions with maximum transmit power lower than the standard 23 dBm (PC3), such as 20 dBm (PC5) and 14 dBm (PC6), that can have reduced power amplifier drain current and use simpler and more compact battery types.
- Introduction of a densely transmitted **Resynchronization Signal** (RSS) allows device to re-acquire time/frequency faster and spend less time (and power) being active for this purpose.
- **Early data transmission** (EDT) feature allows small data (~100s bytes) transmit and receive as early as in random access procedure, i.e., when device tries to the access network after a period of inactivity, while small data transmission (SDT) procedure allows the device to send small amounts of data while remaining in the inactive state.

- **Wake-up signal** (WUS), a defined compact signal transmitted before (up to 2 seconds) the paging occasion of a UE supposed to be in idle (iDRX or eDRX) or connected mode (cDRX), allows the device to skip paging procedures and to go to a near sleep, very low-power state if WUS is not detected.
- **Paging Early Indication** (PEI) signalling can indicate to the UE whether there is need to decode paging signal within a paging occasion (PO) and can be used to skip unnecessary receptions within the PO.
- **Relaxed neighbouring cell measurements** feature can skip these highly power consuming measurements for cell reselection (up to 24 hours) and can be really useful to stationary UEs which suffer from bad coverage but not inter-cell interference.
- Device is allowed to quickly **release connected state**, by successfully acknowledging receipt of the Radio Resource Control (RRC) Connection Release message from network through hybrid automatic repeat request (HARQ) process, instead of waiting up to 10 seconds.

To provide a more vivid example of some of these concepts employed in legacy cellular IoT, the power saving modes and duty cycle states for User Equipment (UE) in NB-IoT are depicted in Figure 5-1. The communications between a UE and an Evolved Node B (eNB) or next generation Node B (gNB) are managed at Radio Resource Control (RRC) layer, for resource setup, modification, and release. The RRC protocol operates in two primary states: RRC Connected and RRC Idle. A UE can transition to the RRC Idle state when eNB triggers an RRC Release message to UE, which is prompt for transmission upon expiration of an RRC inactivity timer (reset after each data packet transmission), defined by the network operator. In the RRC Idle state, UEs can operate within two power-saving modes, eDRX and PSM, as described above: the eDRX state allows the UE to intermittently listen to the NB-IoT physical downlink control channel (NPDCCH); on the other hand, PSM involves the UE shutting off its radio, becoming momentarily unreachable by the network while remaining registered to it. While PSM conserves more energy than eDRX, the latter allows for improved downlink latency due to periodic NPDCCH monitoring.

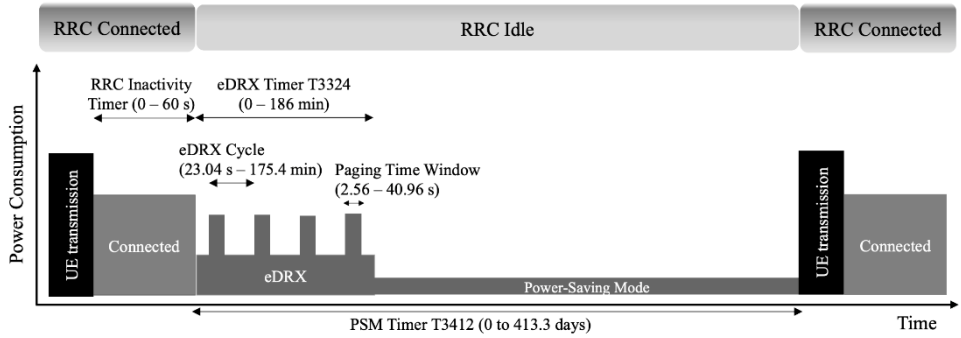


Figure 5-1: Overview of the NB-IoT power saving modes and duty cycle states

5.1.1.3 Evaluation methodology

The following steps of a top-down approach are currently considered to comprise a generic methodology for evaluating the 6G IoT modem power consumption:

1) Understand modem energy usage targets (or equivalently, average power consumption targets) from targeted use cases or usage application. These requirements can be in terms of desired device lifetime, constraints of battery supply, short term energy constraints for device operation, etc. Table 17 gives a visualised such example for quantifying a modem power target.

Table 17: Modem target average power consumption identification (example)

Device target Lifetime / Operation (years / days / min / sec …)	Energy Storage (Ah / mAh)	Device average power consumption target (uA)	Other parts consumption (uA)	**Modem target average power consumption (uA)**
X	Y	$Z= Y_{[Ah]} / X_{[h]} / 10^{-6}$	Z * (100-n) %	**Z * n %**

2) Understand modem operation scenarios and their periodicity from expected traffic scenario of the desired use case. For example, an IoT device may be expected to communicate frequently a small amount of data (e.g., compressed sensor measurements) and less frequently a larger bulk a data (e.g., raw sensor data for calibration or software updates). Furthermore, various different modes for sleep, wake-up, synchronization, inter/intra-cell measurement, etc., may be implemented by device and occur with different durations and periodicities. All these operation scenarios will collectively contribute to the total power consumption (see example in Figure 5-2 and Table 18). It should be noted here that network conditions (e.g., good, or bad signal conditions) as well as the expected network configurations will also play important role on the possible durations and power value ranges experienced from each operation scenario.

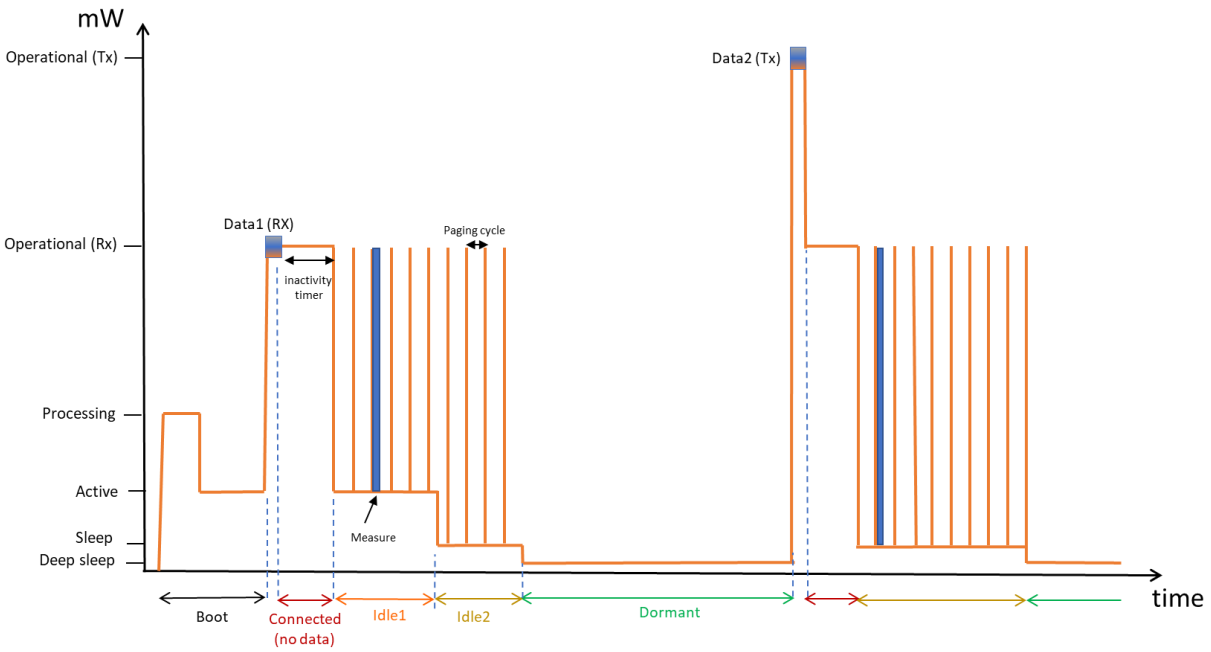


Figure 5-2: Cellular IoT modem power consumption graph (example)

Table 18: Breakdown of modem total consumption to operation scenarios consumption (example)

Operation Scenario	Boot	Data1	Connected	Idle1	Measure	Idle2	Dormant	Data2	…	**Total**
Average power (uA)	A	P_1B	P_2C	P_3D	P_4E	P_5F	P_6G	P_7H	…	A+ P_1B+…

3) The next step is to identify the comprising device states of each operation scenario. For example, considering the legacy cellular IoT, the ‘Idle2’ scenario above could refer to the eDRX cycle where the device commonly needs to monitor paging opportunities, separated by a configured DRX cycle period, within a configured paging time window (PTW), and then do this all over again after a

configured eDRX cycle period (until an expiration timer kicks off). In that case, the eDRX scenario will include device states such as light or deep sleep (between POs and outside PTW), wake-up, synchronize to serving cell, monitor for paging from network, process and then go back to sleep, as illustrated in Figure 5-3 example. In that case, the eDRX scenario will include device states such as light or deep sleep (between POs and outside PTW), wake-up, synchronize to serving cell, monitor for paging from network, process and then go back to sleep, as illustrated in Figure 5-3 example. Of course, the exact device behaviour (states included, order of their occurrence, durations and power consumption) will be always up to device designers' implementation which can also be fitted to address variable system configuration possibilities and signal conditions. The planned power model will take this into account and target to allow flexibility for adjustment to implementation while keeping this simplified approach of discrete scenarios and states to also allow a user-friendly environment for design and analysis.

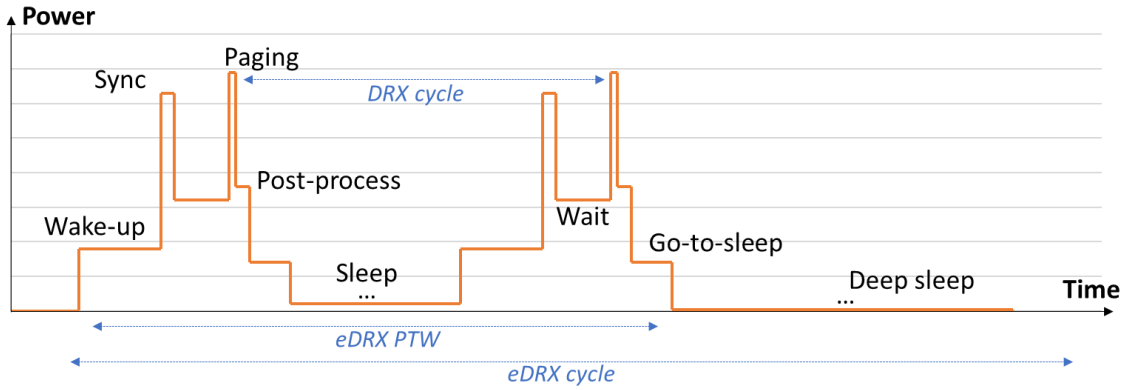


Figure 5-3: Device states and consumption in eDRX cycle scenario (example)

4) Finally, the remaining step will be to model the power consumption of each identified device state. The starting point of the power model design is the legacy cellular IoT devices and associated power saving mechanisms (measurements and field experience) as well as relevant prior studies in literature, such as the recent 3GPP study on UE power saving [38.840].

5.1.1.4 Progress, next steps and future energy saving directions

Currently, the progress in this work has included the setup of a realistic power model for legacy (NB-IoT and LTE-M) modems energy usage in various states and scenarios, according to respective 3GPP technology configuration and lab measurements. An ongoing effort for parameterisation of this model has also started in order to create a generalised model for flexible analysis of several promising features for energy consumption reduction in future EmMTC devices or Zero Energy Devices (ZEDs).

Connectivity operation of future 6G IoT device must be designed to minimize further its power and energy consumption and guarantee its availability to the needs of the respective use cases. For improved, ultra-low power consumption, the various potential 6G IoT devices will need to consider enhanced optimization of hardware components and/or reducing power requirements for the numerous needs of operation. Efficient energy management mechanisms must be implemented at different levels, including device circuitry as well as network communication protocols.

Our future model design and analysis will try to shed light to the effects of a selection of the concepts considered so far in legacy cellular IoT (described above) as well as on a selection of topics of potential interest for energy efficiency that have been identified within D5.2 [HEX223-D52]:

- adaptive low-power sleep modes and duty-cycling at both Tx and Rx (to enter into a deep sleep mode as fast as possible and as long as possible)
- adaptive energy-aware configuration for uplink and downlink transmission (for minimised consumption during network access for data communication)

- energy-aware channel access, synchronization, signalling, and scheduling
- wake-up radio implementation
- advanced battery technologies, e.g., rechargeable battery storage of low self-discharge
- EH forecasting/management mechanisms (ZED without continuous energy supply)
- lightweight security mechanisms
- low complexity sensing methods to minimize consumption from sensor operation (data collection, processing) for detection and respond to input from the physical environment.,
- tiny ML for ZED targeting higher-end applications

5.1.2 Channel coding trade-offs

One example of energy, cost and performance trade-offs in the design of ZEDs is channel coding. It can provide significant performance gain over uncoded transmissions but requires additional complexity. Ultra-low power devices have specific use cases and constraints that should be considered when evaluating the use of channel coding techniques.

[22.840] lists some use cases considered for ultra-low power devices. In most of this use cases data is collected from Ultra-low Power Devices to the Network or Uplink. The payload sizes vary from a few bytes to a thousand bits with most use cases around one hundred bits. In most use cases, Downlink is used for control and the payload size are expected to be smaller than uplink.

Ultra-low power devices have strict complexity constraint. In terms of channel coding, they apply to the encoding for the uplink and decoding for the downlink. Table 19 shows the performance complexity trade-offs for different candidate codes for uplink. Tail Biting Convolutional Codes allow for significant gain over uncoded transmission with a complexity level that is comparable to a Cyclic Redundancy Check (CRC). However, it is possible to achieve even higher performance at the cost of higher complexity using more advanced coding schemes.

Table 19: Complexity/Performance trade-off for uplink channel coding.

Code	Relative encoding complexity[3]	Performance gain over uncoded[4]
Tail Biting Convolutional Code[1]	1	5.5 dB
Turbo Code[1]	2	6.5 dB
Systematic Polar Code[2]	>10	7.2 dB
Notes: 1 [36.212] 2 [VHV16] 3 Extrapolated from [DF03] and [VHV16], Polar Code complexity depend on block length. 4 Eb/N0 gain with respect to uncoded transmission, for BPSK over AWGN channel, 112 bits payloads, 1/3 code rate		

As can be seen in [DF03], decoding complexity is significantly higher than encoding complexity for state-of-the-art channel coding. Furthermore, performance of codes usually decreases for smaller payload. For these reasons, implementing efficient channel coding for ultra-low power devices downlink is challenging. Error detection capability like CRC should be prioritized over channel coding to allow for reliable control payload delivery.

5.2 Enabling technologies

5.2.1 EH technologies

EH from ambient sources is a key driver for the development of ZEDs. The basic architecture of an EH-enabled system consists of (see Section 4.3) i) transducers, to convert ambient into electrical energy; ii) PMIC, to maximize the harvested energy by dynamically matching the output impedance of the transducer to the rest of

the circuit; and iii) energy storage components, e.g., supercapacitors, to compensate for the random fluctuations in the harvested energy.

The EH technology suitable for a particular use case depends on the availability, intensity, and type of ambient energy and the devices' form factor constraints and energy demands [NBM+23]. For example, light-based EH only operates during the day or office hours (depending on whether the deployment of the device is indoor or outdoor) and is sensitive to blockage. Heat-based EH depends on how the temperature changes in the surroundings, i.e., in the form of spatial gradients and temporal temperature variations. Also, such transducers are constantly exposed to mechanical stress and may require a large form factor compared to that of the device to guarantee being exposed to significant temperature variations. Microbial fuel cells leverage the energy released in nature as the result of organic matter processing by colonies of bacteria. Current implementations of microbial fuel cells are expensive, and their lifetime and performance is dictated by the activity of the microorganisms and their living conditions. Vibration-based EH leverages the physical properties of certain material to harness the energy form vibration, impact, deformation, and friction. Vibration-based EH are, in general, compact and reliable devices but are complex to manufacture and require resilience to mechanical fatigue. Flow-based EH turns the kinetic energy of naturally or artificially originated fluid flows, e.g., water streams and wind currents, to generate electricity. Although some implementations incorporate vibration-based EH, traditional implementations such as wind turbines are too bulky to fit small form factor devices. Finally, RF-EH fits well in ultra-compact devices as it only requires rectifying circuits to convert the impinging electromagnetic energy into electricity. The biggest drawback of this technology is its output power, which only suffices to power ultra-low power devices at least when relying on non-dedicated RF transmitters. Table 20 summarizes the advantages, limitations, and some attractive use cases of these EH technologies. The reader can refer to [LMR+23] for a more detailed discussion on development directions, challenges, and applications of the aforementioned EH implementations.

Key performance indicators relevant to EH-enabled systems are [LMR+23]: i) power density, which reveals insights on the achievable harvested energy for given transducer dimensions; ii) conversion efficiency, which is the percentage of the incident ambient energy converted into electricity; and iii) dynamic range, which provides a range of input energy levels for which the transducer conversion efficiency is above a certain value. Some development directions to boost the harvested energy are [LRR+23]: i) using array of transducers; ii) widening the frequency response of the transducer, e.g., multi-junction solar cells, multi-band/low-frequency vibration-based EH, and broadband/multi-band RF-EH; ii) capturing energy from multiple directions, e.g., omnidirectional RF-EH, multi-dimensional vibration-based EH, and concentrator photovoltaics; and iii) resorting to hybrid EH. Specifically, hybrid EH combines the output of multiple transducers to allow EH from different sources, which increases the average harvested energy and potentially boosts the reliability of the energy supply since the likelihood of at least one source available is higher than in the single-source EH case (see the reference architecture discussed in Section 4.3). Alternative implementations of hybrid EH may consider harvesting the wasted energy from the main EH system, such as when combining solar cells and thermoelectric transducers. In such a case, thermoelectric transducers can transform the heat of the solar cells into electricity, and consequently act as coolers which boost the performance of the solar cells.

Table 20: Comparison among EH technologies [LMR+23].

Energy Source	**Power Density**	**Form Factor**	**Limitations**	**Use Cases**
Light	17 mW/cm^2	medium/large	sensitive to blockage	remote applications, indoor/outdoor sensors
Heat	10 mW/cm^2	medium	sensitive to thermal stress	wearables, infrastructure monitoring

Microbial fuel cells	5 mW/cm^3	medium	limited stability/lifespan	smart farming, wastewater treatment
Vibration	20 mW/cm^3	small	sensitive to mechanical fatigue	wearables, infrastructure monitoring
RF	10 nW/cm^2	very small	very low output power	asset tracking

5.2.2 RF-WPT

RF-WPT represents a prominent research avenue in the pursuit of a sustainable energy supply that can support the widespread deployment of low-cost/power devices [LAS+21], [RLA+23]. Unlike, EH technologies described earlier, which are heavily dependent on the availability of ambient energy sources, RF-WPT relies on dedicated RF transmitters, known as power beacons (PBs), which provide a controlled and predictable energy supply for sustaining devices' operation. Specifically, RF-WPT stands out as a technology with the potential to revolutionize the way devices are powered. This is due to its inherent capability of broadcasting energy over long distances and charge multiple devices simultaneously, even in non-line-of-sight conditions, as well as for moving devices. Moreover, the RF-EH circuit form factor and manufacturing costs allow a seamless integration of this technology on existing devices and enable a dual EH from both dedicated and ambient energy sources simultaneously.

Despite the above, it may not seem obvious that WPT deployment will bring additional benefits other than reducing the battery replacement as RF-WPT-enabled deployments require additional hardware with higher upfront costs. To clarify this, the authors in [RLA+23] illustrate how maintenance and maintenance costs scale with both the number of deployed devices and the devices' hardware lifespan. For this purpose, consider i) a conventional grid-powered PBs deployment, where the energy is harvested in a distant location and distributed using the main power grid; ii) a battery-powered PB's deployment, where the energy comes from power banks; and iii) a green-powered PBs deployment, where green energy is locally harvested by each PB. As a baseline scenario, consider a conventional battery-powered devices' deployment. The overall costs of the RF-WPT scenarios are computed considering the costs per kWh of electricity demanded by the PBs' network over the devices' hardware lifespan. As for the devices, we assume that maintenance, i.e., battery replacements, accounts for a fraction of their initial deployment costs. The results evinced that batteries are the most cost-effective solution for powering deployments with few devices. However, as the number of devices increases, deploying grid-powered PBs, and especially green-powered PBs, reduces the overall costs. Furthermore, as illustrated in Figure 5-4, inaccurate hardware profiling, battery imperfections, and/or operating conditions may increase the overall costs far from expected in which case even relying on battery-powered PBs may become more cost-effective.

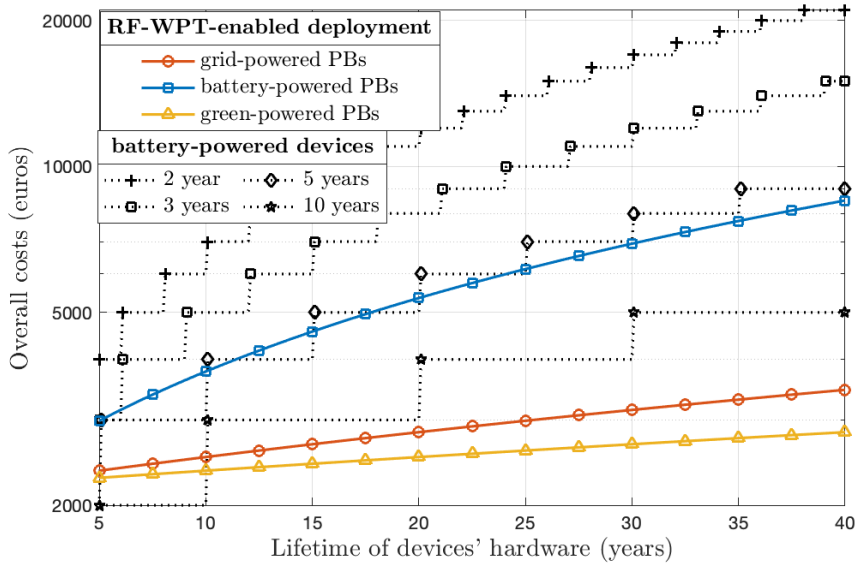


Figure 5-4: Qualitative assessment of the overall costs vs the lifetime of the devices' hardware lifetime for 100 deployed devices and different battery lifetimes [RLA+23].

In RF-WPT systems, low-cost MIMO architectures play a crucial role in charging low-energy devices with a reduced amount of power consumption. This is done by reducing the number of RF chains, which are the most power consuming part of the system. Furthermore, reducing power consumption results in a higher end-to-end efficiency, which is critical when charging a massive number of devices. One of the potential enablers of low-cost wireless charging for large MIMO setups is the novel Dynamic Metasurface Antenna Array (DMAA). DMAA is an emerging architecture, which consists of several waveguides, each one connected to a dedicated RF chain. Furthermore, each waveguide feeds multiple radiating metamaterial elements, thus, the number of RF chains and power consumption can be reduced.

The potential role of DMAA-assisted transmitters in charging multiple low-energy devices with a lower amount of power consumption is investigated in [ALC+23], [ALS+23]. In [ALS+23], energy beamforming is leveraged to focus energy beams toward EH devices to meet their requirements, Therein, it is shown that a DMAA-assisted system outperforms the traditional fully digital architecture in terms of power consumption, especially in the low-frequency regime. Moreover, the authors also investigated the influence of the operation frequency on the system performance, which is an important issue since some devices may operate in high-frequency regime, e.g., mmWave and THz bands, in future wireless systems.

Another promising technology for massive MIMO-based RF-WPT is the radio stripe network [LKS+22], [ALP+23]. Therein, a number of radiating elements are implemented along a cable with one or more CPUs. Interestingly, one CPU is enough for the coordination in radio stripe systems thanks to their compute and forward architecture, which makes their implementation complexity independent of the number of elements. Figure 5-5a demonstrates a potential use-case of the radio stripe systems for real-world RF-WPT applications, Therein, a restaurant is equipped with a radio stripe network in the ceiling, which attempts to charge multiple low-energy devices located in some predefined hotspots. Moreover, the deployment problem of the radio stripes has been addressed in this scenario, aiming to maximize the minimum deliverable RF power to the devices. Figure 5-5b represents the results, where it is observed that the radio stripe-based deployment outperforms the traditional central fully digital architecture in terms of minimum received RF power by the devices. Thus, the radio stripe systems may facilitate the implementation of massive RF-WPT systems, e.g., shopping malls and stadiums, with a much lower cost/complexity [ALP+23].

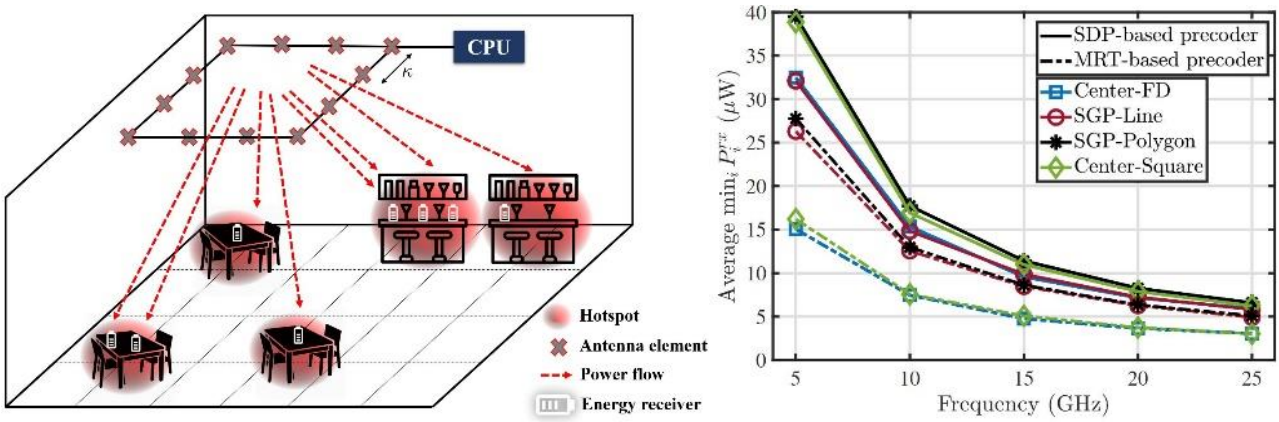


Figure 5-5: (a) Radio stripe system model with a central processing unit, exemplified with a restaurant scenario (left), and (b) average minimum power received by the devices as a function of frequency of operation (right) [ALP+23].

5.2.3 Energy aware protocols

Energy-aware protocols are crucial to meet/improve energy-sustainability KPIs [LRR+23]. As discussed in Section 5.2.1, harvesting resources may not be available all the time especially if they are harvested from ambient or natural resources. Additionally, the amount of power received through WPT may not necessarily be sufficient for instance to perform the subsequent uplink transmission. Furthermore, the amount of energy that can be stored is very limited and the device can easily run out of the back-up stored energy. Consequently, the device might not be able to communicate with the network or other devices during a certain period when the instantaneous harvesting energy is not enough and/or the stored energy level drops below a certain level. Under these conditions, as mentioned above, the device may need to stay in sleep mode or an EH state for a certain time before it can reach an energy level where it can perform transmission or reception. This, in return, can result in the device and the network not being to communicate properly, if the scheduling is performed as is traditionally done in cellular connectivity. To avoid this, it is necessary to select and adapt uplink and downlink transmission and their corresponding scheduling based on availability of energy source in the device.

- ***Adaptation of uplink transmission operation:*** Adaptive time-gaps for uplink scheduling, necessary for a device to perform EH to store energy in its storage device where the time-gaps is adapted according to the device harvesting capabilities and possibilities for receiving power through WPT.
- ***Adaptation of downlink channel monitoring for reception and downlink transmission operation:*** Duty-cycled channel monitoring is typically used to reduce the cost of idle channel monitoring. While the duty-cycle duration is selected to fulfil certain requirements in terms of latency or device availability, the actual duration of the on and sleep time need to be adapted based on the availability of the energy at the device or possibilities for receiving power through WPT. The time for channel monitoring is adapted such that the device can perform EH to store energy in its storage device. On the other word, time-gaps are inserted within the on time where the device can harvest energy.

In both uplink and downlink transmission, base-station and other devices in the network need to adjust their listening intervals and data transmissions to the respective adaptive time-gaps of the device. For this, an assistance information needs to be provided by the device in terms of its harvesting capabilities and its storage device, for instance through device capabilities. During the EH period, where no communication can be performed, the device basically enters a new state, in addition to transmission, monitoring/reception and sleep, where we refer to as power limited or energy harvesting state. It is important that the assistance information is provided early on before the device enters the power limited state. The information related to harvesting capabilities and storage size can be provided to the network, for instance at the registration time through indication of the device type or device category. Further information in terms of traffic conditions, amount of data, energy storage level, and availability of the harvesting resources is provided during the actual data transmission.

In general, communication protocols (but also computation and sensing) for ultra-low power IoT devices must be adaptive and manage energy resources based on availability and demand patterns. Protocols cannot

compromise present and future system states, which inevitably requires energy-awareness. At the MAC, energy-aware grant-free random access protocols are appealing. Indeed, the grant-free feature allows relatively simple protocols, with reduced control signalling. As an example, consider the work in [SLS+23], where the potential gains of energy-aware MAC and computation protocols are investigated when considering a federated learning task in a massive IoT scenario and EH devices. Therein, a Slotted ALOHA policy with multiple channels is used, while the devices have two chances for saving energy: 1) a device decides to engage in an iteration based on a sleep probability; and 2) devices send their local update based on how informative it is for the global model. Herein, it is important to tune the sleep probability for the right functioning of the network. The set goal was to achieve a sufficient energy level at the end of the learning process, i.e., an energy threshold, enabling the execution of future tasks and mitigating energy depletion during the current task. For this, both a full distribution knowledge of the energy income process or just its mean are considered. As depicted in Figure 5-6, the method is compared to the Largest Updates' Norm approach, which only considers the informativeness of updates. From the numerical results, it is possible to see that energy-aware protocols enable a more reliable execution of tasks while increasing energy efficiency.

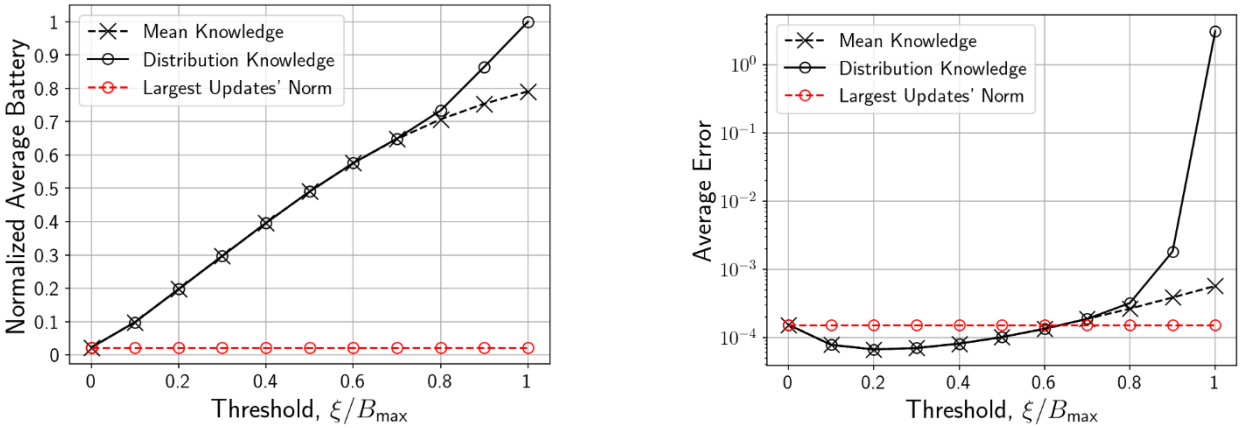


Figure 5-6: Normalized average battery (left) and average error (right) as a function of the normalized energy threshold [SLS+23].

5.2.4 RAN scope dedicated connectionless design

An EN device is a device segment well below the existing cellular IoT devices (NB-IoT, LTE-MTC, RedCap). These devices have usually no guaranteed energy storage as they rely on ambient EH sources, which makes them unreliable. In addition, they have extremely limited processing and data handling capability due to small form factor and low-cost hardware components. Due to such limitations, the device requires simplification at both physical and higher layers compared to legacy cellular networks. Focusing on higher layer design, one simplification would be to shift protocol design from Radio Access Network (RAN) scope connection-oriented mode, e.g., RRC connection-based approach in legacy cellular technologies (e.g., 4G, 5G) to RAN scope dedicate connectionless communication. This means the device will have

- Limited connection with Sixth Generation (6G) RAN which does not involve a dedicated connection, resulting in fewer handshakes between the EN device and RAN node
- Improved lightweight Non-Access Stratum (NAS) security between EN device and 6G core network, and
- Improved lightweight User Plane (UP) security between EN device and 6G RAN node or 6G UP Function (UPF).

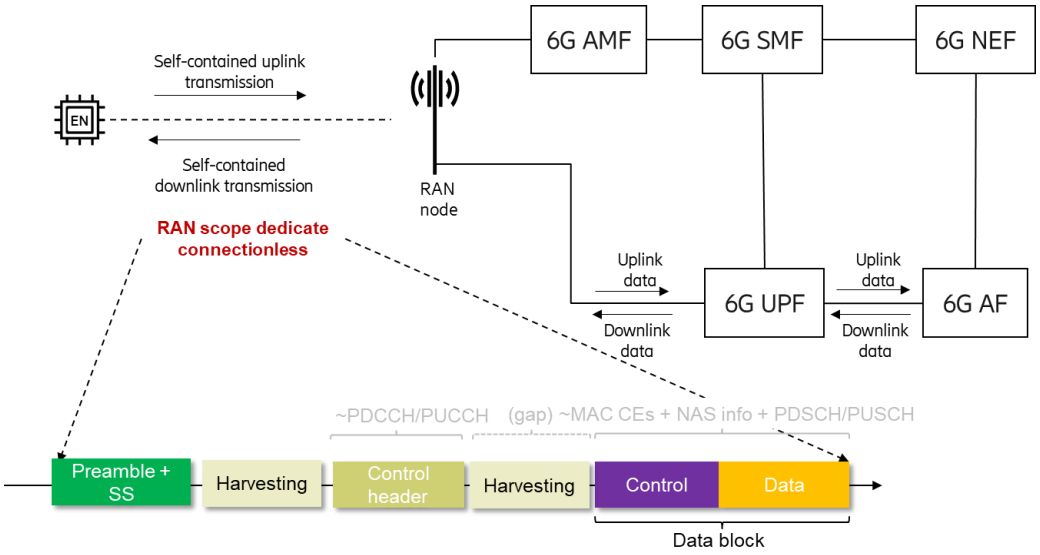


Figure 5-7: RAN scope dedicated connectionless design for EN device handling.

The RAN dedicated scope connectionless protocols are well suited for EN devices as it does not require stringent dedicated handling or connection, stringent Quality-of-Service (QoS) flow, i.e., no dedicated bearer provisioning, and device context in the network will be short lived if applicable. This reduces monitoring load, result in simplified channels, and reduced signaling overhead. Since there is no dedicated handling, the transmissions are made self-contained, which means the user identifier is included in transmissions along with data for its identification at RAN node and Core Network (CN). The identifier can be significantly large, which can be a major component of UL data block or control information. Thus, it is important to define identifier a short function of the CN identifier in order to reduce payload size. These self-contained transmissions in UL can occur using the Random-Access Channel (RACH) or contention resources, and in DL, over paging occasions. Once the initial transmission is performed and user is identified, network can allocate short lived device context and if requires, network can allocate additional resources for subsequent transmissions.

In regard to contention-based UL, the transmission can be prone to collisions, and thus various options can be considered subject to the use case, cell load, EN capability and harvesting infrastructure requirement. In any option, the UL transmission must include an identifier, possibly based on CN identifier to identify the EN device at the RAN node. Furthermore, if there are subsequent transmissions, the RAN node can allocate device context (some signature identifier) for a short period to cater to the identified user's transmissions in UL or DL. Below three options for contention UL and the applicability criteria are listed.

1. Preamble-less self-contained contention UL

The UL user transmits without preamble and utilizes only Synchronization Sequence (SS), header, and data. The preamble is excluded to save transmission cost; however, this may reduce the collision resolution performance. The header is transmitted to indicate RAN node about control information parameter associated with data block.

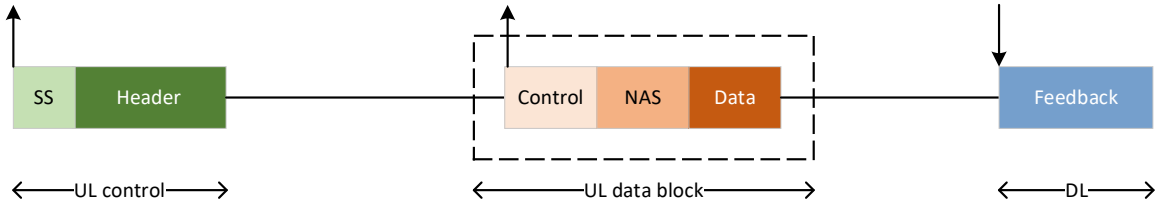


Figure 5-8: Self-contained UL without preamble; the header controlling UL data is in UL as well.

2. Self-contained contention UL

In comparison to above, preamble is transmitted in this variant, thus increasing the collision resolution probability. This is useful in scenarios with high activity or traffic load where contention of resources may result in increased collisions.

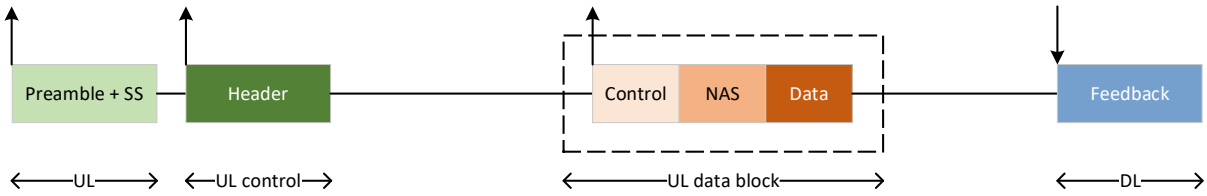


Figure 5-9: Self-contained UL with preamble; the header controlling UL data is in UL as well.

3. Self-contained assisted UL

In this variant, the UL data transmission is subject to intermediate DL response. As opposed to UL header in previous headers, the DL header assist UL transmission and can indicate data transmission control information or contention control commands or correction commands. This is useful in scenarios with high activity loads or EN device with poor hardware (e.g., poor synchronization or drift issues).

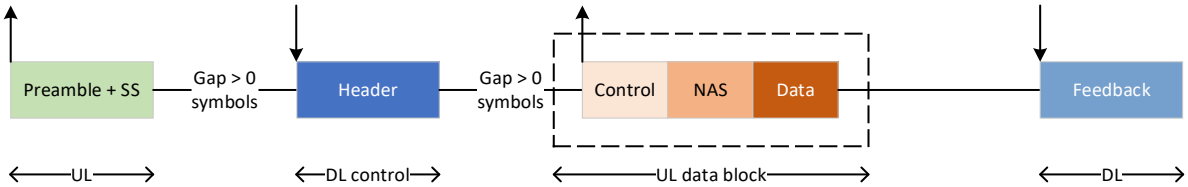


Figure 5-10: Self-contained assisted UL transmission; DL header is provided to control UL transmission.

5.2.5 TinyML

TinyML specializes in crafting computationally efficient and resource-limited ML algorithms tailored for low-power IoT devices. Energy efficient memory access and reduced computational demands resulting from the utilization of techniques, such as quantization, pruning, neural architecture search, etc, makes an ML algorithm tiny. Equipping IoT devices with tinyML empowers them to execute tasks independently, eliminating the need for constant reliance on the cloud. Moreover, having ML algorithm on the cloud requires back-and-forth signalling exchange and also local data needs to be provided to the cloud for improving its ML algorithm. In general, tinyML can be used to perform on-device analysis and do real-time predictions or take real-time decisions, e.g., to forecast future EH values, aiding in formulating proactive EN operational strategies for IoT devices [ASC+23][BTO22].

By enabling on-device data analysis, one can achieve strict privacy and security measures but also minimize communication latency. The benefits of this immediate and local processing of information extend far beyond privacy and security. One of the most significant ones, which is a direct outcome of the communication reduction made possible by the on-device computations, is the energy efficiency of the device. Minimizing the back-and-forth communication with a cloud server automatically reduces its energy consumption, which is especially crucial for ultra-low power or EN IoT devices. Data transmission is also immediately related to cost and scalability. The decentralization of data processing ensures a scalable and cost-efficient ecosystem. Lastly, by not relying on constant connectivity, the reliability of the device is increased, since its operation remains uninterrupted.

Notably, integrating ML models into IoT devices presents a complex challenge. Some difficulties to be considered are the complexity of ML models, deployment strategies and restricted device resources. Furthermore, the trade-off between model accuracy, speed, and power consumption needs to be meticulously balanced. Achieving optimal performance based on the hardware/software restrictions and capabilities is

critical. The end goal is to create an integration where the model operates efficiently without compromising the device's functionality or battery life.

To craft tinyML models, several techniques are at our disposal as illustrated in Figure 5-11[LRR+23]: i) Neural Architecture Search, which is a method aimed at discovering the optimal neural network architecture that can seamlessly fit within the resource constraints of an IoT device; ii) parallel ultra-low power processors, which offer software-level acceleration for tinyML models; iii) model compression, including quantization, pruning, knowledge distillation, and weight sharing, which can significantly reduce the computational demands of the ML model; and iv) efficient memory access, which can be implemented to minimize the average Dynamic Random-Access Memory (DRAM) energy consumption during tinyML inference.

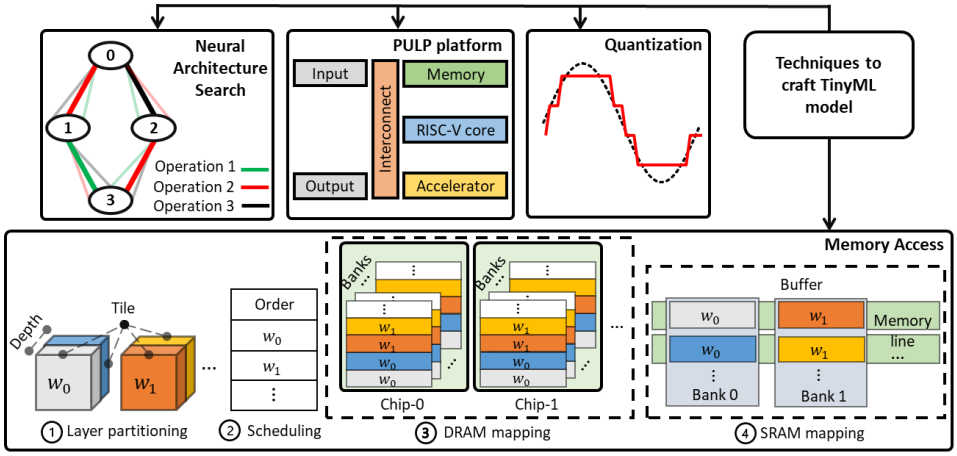


Figure 5-11: Techniques to craft a tinyML model [LRR+23].

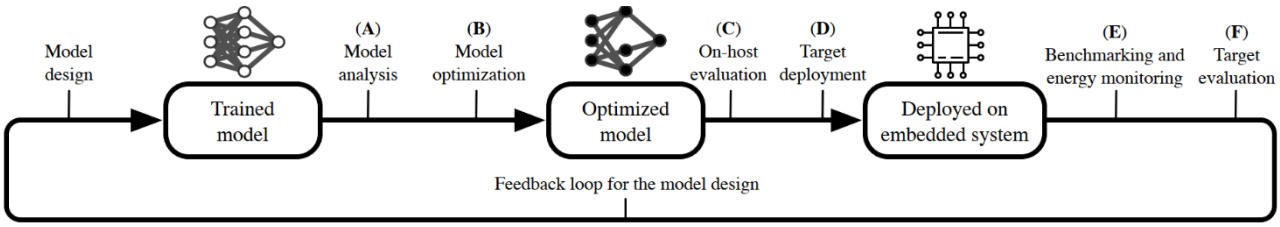


Figure 5-12: Taken from [HBQ+21]

To enable the deployment of tinyML algorithms on battery-less EN devices, we applied model compression (i.e., pruning and quantization) on a pre-trained neural network. Moreover, the device was equipped with an energy-aware model execution and offloading algorithm, to ensure it would only execute the local tinyML model, or offload the captured data SAS+23 to a larger and more accurate edge-cloud model if enough energy is harvested. It can be applied to different applications such as remote monitoring where sensors are hard to reach and require an extremely long lifetime, and fields related to natural sciences where the approach can be used for the detection and counting of different species. As a proof of concept to validate the proposed approach, an application to detecting people on a captured image using low-power cameras was developed. Considering energy awareness, this approach can avoid power failures and maintain forward progress. A Convolutional Neural Network (CNN) was developed, trained, and deployed in an edge cloud server, after which state-of-the-art optimization techniques (i.e., model pruning and quantization) was applied to convert the cloud-based NN models into leaner tinyML models that can be executed on constrained EN devices. By applying different level of compression, multiple tinyML models, with a different energy vs. accuracy trade-off can be created. Based on the predicted harvested energy, the device can then decide which model to execute, or whether to offload data to the cloud for even more accurate inference. The latter requires significantly more energy, as transmission of images is more costly than local computation. This allows the device to always

select the most accurate inference alternative given its energy constraints. The solution is illustrated in Figure 5-13.

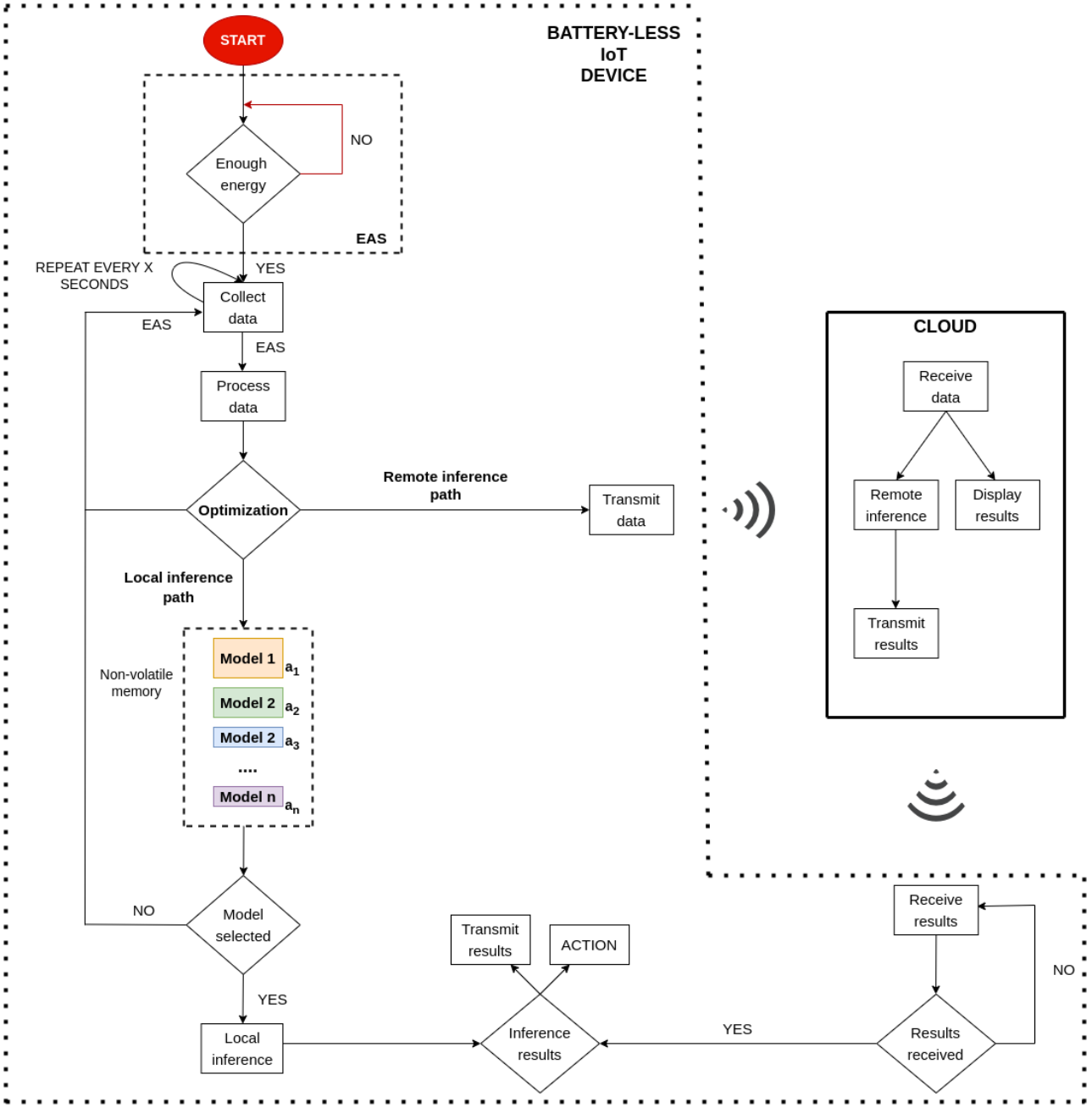


Figure 5-13: The proposed approach for energy-aware management and deployment of multiple tinyML models on the EN device, with cloud offloading capabilities.

The proposed architecture consists of two main parts: (I) the EN device, and (ii) the IoT Gateway/cloud, which are connected using the wireless communication, such as NB-IoT, LoRa, or Bluetooth Low Energy (BLE). In the prototype, BLE was used, due to its lower power consumption compared to long-range technologies and ability to send images even with limited harvested energy. In this way, we enabled both parts of the system to be capable of making decisions by performing inference locally (and additionally send the final result to the cloud) or sending the captured image to the cloud for remote inference. Based on that, the trade-off between different inference strategies is studied, analyzing under which circumstances it is better to make the decision locally or send the data to the cloud where the heavy-weight ML model is deployed, respecting energy, time,

and accuracy constraints. To decide which of these two options is preferable and can satisfy all set of constraints, an energy-aware inference algorithm was defined.

The prototype, as shown in Figure 5-14, was evaluated based on real experiments considering two different environments: (i) a controllable setup with artificial light where the harvesting current and voltage are constant during the full time of the experiment, and (ii) a dynamic harvesting environment based on natural light where the harvesting current and voltage vary over time due to unpredictable sunlight intensity.

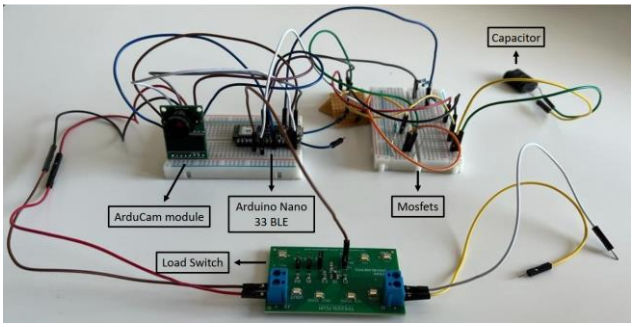


Figure 5-14: The EN tinyML prototype consisting of a low power ArduCam, and Arduino Nano 33 BLE microcontroller, a power manager with load switches, capacitor, and mosfets to measure energy availability, and store and harvest energy from a solar panel

For the first approach, we assumed that the value of the harvesting current was perfectly determined and will not change over time. In reality, this knowledge is not perfect as light (e.g., sunlight) is unpredictable and dynamic, causing the harvesting current to change frequently. For this approach, we considered three strategies: (i) Local Inference (LI), (ii) LI with sending results to the cloud (LIS), and (iii) Remote Inference (RI) that is executed in the cloud. It must be noted that all three considered strategies were deployed and tested on the EN device separately, considering the same configuration and parameters for the fair comparison.

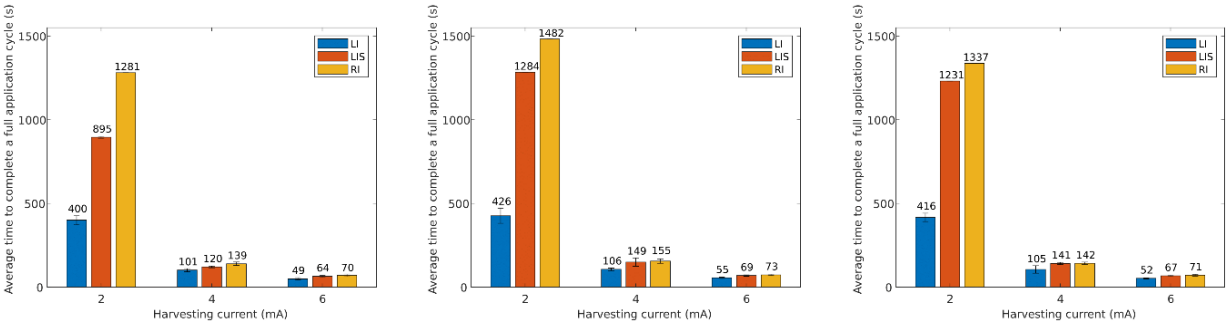


Figure 5-15: Average time needed for execution of the full application cycle considering different inference strategies, capacitor sizes, and harvesting currents (this is the caption), and from left to right these are graphs considering the capacitor size of 0.5, 1, and 1.5F.

The results in Figure 5-15 show that the LI strategy without sending results to the cloud performs best in terms of the average time needed for one application cycle in all considered cases. Considering a harvesting current of 2mA and a capacitor of 1.5F, the LI strategy is able to execute application cycles 3 times more frequently compared to the RI strategy. In the end, this results in more executed application cycles, consuming less energy, but at some cost of accuracy, as the less accurate tinyML model is used for inferencing.

In the second case, a dynamic harvesting environment based on natural light is considered, where the harvesting current changes over time. For this approach, two main inference strategies, the LI strategy without sending results to the cloud, and the RI strategy where the final decision is made in the cloud, are considered.

Considering all set constraints, the energy-aware optimization algorithm was able to decide which of these two inference strategies is more suitable for certain harvesting conditions. As shown in Figure 5-16, the east-side Battery-Less Node (BLN) performed 15.32% more application cycles compared to the west-side BLN due to the higher harvesting currents for nearly the full duration of the experiment. Because of lower harvesting currents, the west-side BLN mostly performed the LI strategy (14.81% more than the east-side BLN).

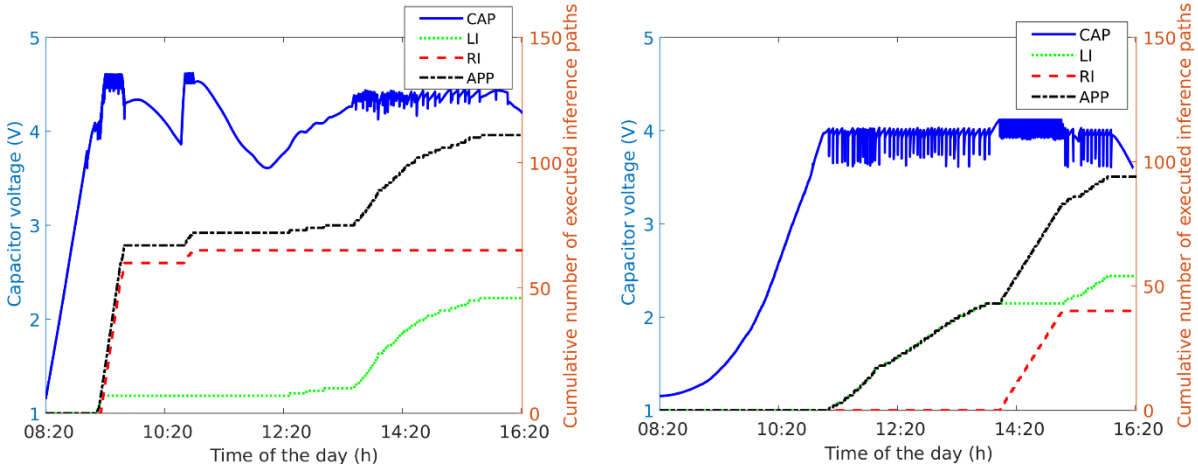


Figure 5-16: Capacitor voltage behavior (1.5F) over time when executing different tasks with a solar panel placed at east and west side windows.

5.2.6 Intelligent wake-up

Wake-Up Radio (WuR) was first proposed in [GZR+01] and allow the activation of a device's main radio only after detection of a Wake-up Signal (WuS), thus conserving energy and extending battery life of IoT devices [LRR+23]. In cellular systems, WuS scheme was first introduced in NB-IoT/eMTC in 3GPP Release 15, where the WuS is a downlink physical signal transmitted prior to paging with the aim to "wake up" a user device from an idle state so that can prepare to receive data. This feature was enhanced and also extended to be applied to NR devices in Rel 16 – Rel 18. With the furthermore introduction of enhanced cross-slot scheduling in 3GPP Release 16, a device can be informed if the time delta between uplink and downlink control information and data slots respectively is sufficient to enter in microsleep mode thus reducing even more its unnecessary RF operations.

Research [FFB+19] shows that a WuR enabled devices' average consumption [ODC+16] is at the magnitude of 1000 times lower than that of the Main Radio (MR). Furthermore, implementing a WuR [MJS+16] is shown to achieve 70 times longer lifetime than Duty Cycling (DC) protocols under a light traffic load. Researchers in a case study [FFB+19] measured a total current consumption of only 390 nA. Using this measurement as a reference, a EN device using a 500 F capacitor bank can continuously decode WuS for over 26 years without entering sleep mode or considering "battery" discharge, or circuit disintegration. Of course, the expected "battery" lifetime is different in a real-world scenario where a receive/transmit and/or sensor acquisition is required, but the numbers can give a helpful insight.

The specifics of the WuR implementation depend on the protocol, circuitry, and purpose. For example, real-life results in [FFB+19] show that a WuR-enabled BLE device targeting an IoT scenario where such battery-powered massive IoT devices do not support direct 3GPP connections, meets the over-10-year lifetime target while satisfying the latency requirements for 5G IoT devices. Similarly, in [MK20] light has been shed on BLE-compatible sensor devices enriched with a WuR, and their results demonstrate energy-efficiency gains in applications with under 2s latencies. Meanwhile, up to 35% energy consumption reduction has been achieved in [RRM+23], by using WuR with an accurate traffic forecasting model. Additionally, WuR has been proposed as super-regenerative solution in [PMV+16] to improve energy efficiency in human-body communication. Therein, WuR operates at a very low data rate, e.g., 1.25 kbps, for higher sensitivity while consuming ~40 µW. Likewise, it has been shown in [PKM+15] the energy efficiency from using WuR capable

of receiving small control commands besides WuS in wireless body area network applications with event-driven traffic.

When listening for WuS, there are basically two types of error that can happen. The WuS is either missed or a non-existing one is erroneously detected. These error events are referred to as miss, and false alarm. A missed detection has no impact on device energy consumption, but the network needs to spend extra resources to re-transmit and an extra delay is added. A false alarm, on the other hand, leads to an extra energy consumption at the device, since it performs unnecessary action to receive non-existing data or perform an unnecessary data transmission. Therefore, when designing a WuS, it is important to select a WUS that it can be detected, with low miss and false-alarm probabilities, by an ultra-low power WuR [LRR+23].

Some key challenges and research directions are:

- WuS may complicate radio resource management and device scheduling in the network due to sleep patterns, reducing potential energy efficiency gains. In this context, employing advanced ML-based scheduling algorithms considering the sleep patterns of devices may be appealing [RRM+23]. ML in Intelligent WuR implies a more intelligent and adaptive approach to scheduling wake-up events, potentially considering historical sleep patterns, predicting future patterns, or dynamically adjusting schedules based on real-time data.
- The traditional design where the WuR utilizes a different frequency band than the main radio increases the complexity and cost of the devices. Therefore, in-band operation and RF integrated circuit (RFIC)-embedded WuR implementation is desired. However, this approach complicates resource management and reduces the available spectral resources for transferring application data.
- Beamformed WuS and mobility management is still an open challenge since beam sweeping for WuS is required to reach a desired device. The network should be able to optimize the number of beams in a single WuS burst utilized for waking up the device.
- Applying WuR brings trade-offs between energy efficiency and other KPIs, depending on the application scenario, like latency, reliability, and robustness [RKL+20]. Therefore, more research should be directed in this direction, especially when dealing with massive low-power IoT scenarios.

Also, intelligent wake-up can be implemented by a tuned TinyML model (refer to Section 5.2.5) that schedules the sleep of the device based on the available energy stored in the capacitor bank (in the case of EN devices). That way, a device can operate “autonomously”, adapting to the ever-changing environment around it while providing measured data only when needed thus removing the need for excessive communication loops when data are not needed or changing. Additionally, utilizing ML at the base station for WuR by gathering diverse data from network devices, aids in training predictive models to optimize wake-up parameters based on features like energy levels and environmental factors. Once deployed, these models intelligently trigger wake-up signals only when necessary, conserving energy by minimizing unnecessary device wake-ups. Continuous model refinement using new data ensures adaptability to changing network conditions, enhancing overall energy efficiency in the network.

5.3 ZE PoC

3GPP has recently a study item on Ambient Internet of Things (AIoT) [38.848] that investigates the feasibility of a new IoT connectivity solution which would provide complexity and power consumption orders-of-magnitude lower than existing 3GPP LPWA technologies such as NB-IoT and LTE-MTC . Backscatter radio is seen as a promising candidate technology for zero-energy AIoT devices. Figure 5-17 illustrates the various deployment topologies for integrating ZE Backscatter Devices (BDs) to the mobile communication system. These topologies presented here resemble but are not exactly aligned on those outlined in [38.848].

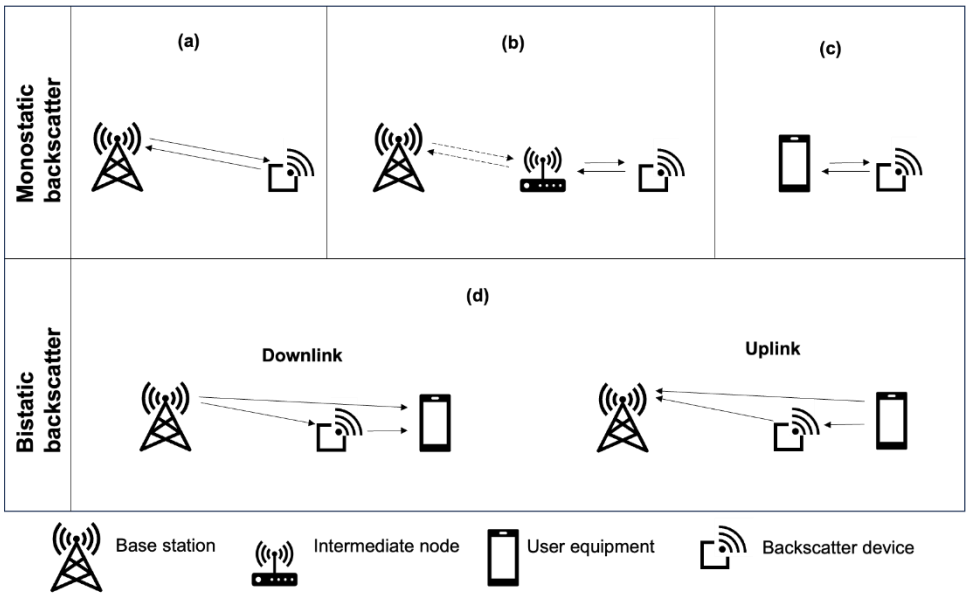


Figure 5-17: Different candidate topologies for BDs.

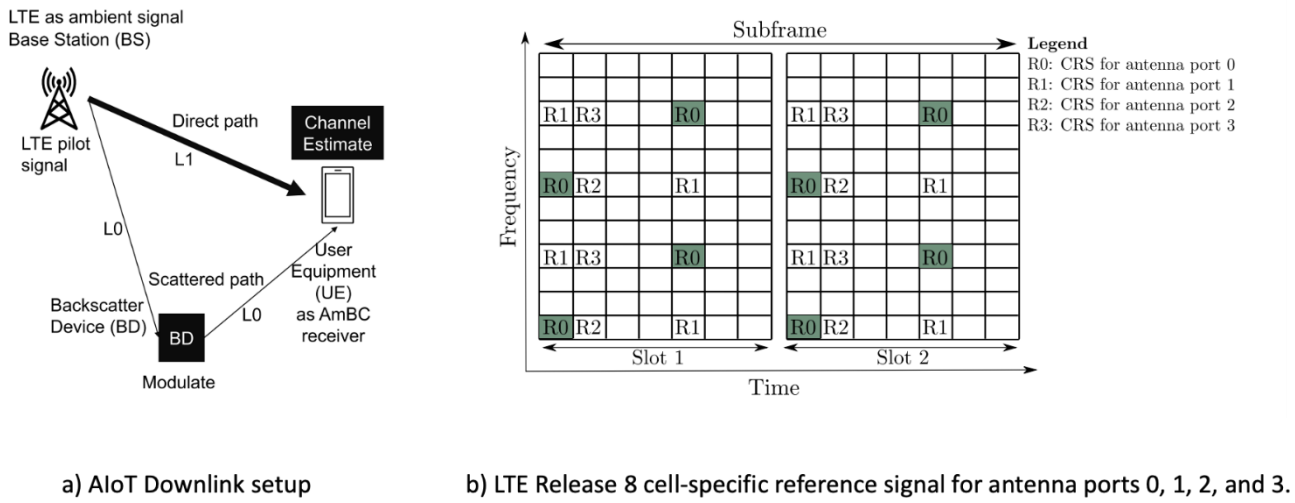
In traditional backscatter communication systems like those used in Radio Frequency Identification (RFID), the reader device, equipped with a full-duplex transceiver, is responsible for generating an unmodulated carrier wave that is transmitted to the backscatter device (BD). The BD then modulates this carrier wave with its own information and backscatters it to the reader for decoding. This architecture places a significant complexity burden on the reader device, as it must be capable of simultaneous transmission and reception. AmBC capitalizes on existing ambient RF signals, such as those from television [LPT+13] or cellular broadcasts [PBR+22], as opposed to requiring a dedicated carrier signal generated by a reader. This approach simplifies the design by reducing both computational and hardware requirements for the reader device as the transmitter and receiver are spatially separated from each other. Additionally, the use of ambient RF signals alleviates the necessity for dedicated spectrum allocation allowing the operators to use their existing spectrum more efficiently.

In this ZE PoC, we demonstrate the feasibility of using AmBC in the cellular downlink. The studied scenario aligns with topology (d) in the above figure. The proposed system design makes use of the periodically transmitted downlink reference signals that the user equipment (UE) uses for channel estimation. BD modulates the downlink signal transmitted by the base station. UE see the presence of BD as an additional multipath component. To enable the UE channel estimator to separate the BD modulated signal path from other paths, the BD performs frequency shift keying (FSK). This causes the BD modulated path to have an artificial Doppler that is higher than the natural Doppler in the channel allowing the UE channel estimator to separate the BD modulated path from the rest of the multi-path components in the frequency domain. The PoC presented here used LTE Cell Specific Reference Signals (CRS) as a signal source and implemented UE channel estimator with software defined radio [36.211]. The proposed receiver could be easily implemented in real UE modem as a middleware update without the need for any new hardware. While the PoC was made using LTE, we note that the same concept can be applied in 5G New Radio and any future mobile communication system that transmits reference signals to enable channel estimation.

5.3.1 System description

The system is composed of three part of hardware, base station (BS), BD and UE. The key idea of the proposed AmBC system is utilizing the channel estimation algorithm embedded in UE. The BD changes propagation environment, which is recognized by UE channel estimation as shown in Figure 5-18.

LTE CRS is known by the UE. CRS in LTE exists in the 0, 4, 7 and 11 OFDM symbol in each OFDM subframe. The period of CRS existence frequency is roughly 4 kHz in nonuniform samples [LWR+23]. CRS exists specific subcarrier as Figure 5-18.



a) AIoT Downlink setup b) LTE Release 8 cell-specific reference signal for antenna ports 0, 1, 2, and 3.

Figure 5-18: AIoT Downlink in LTE and utilized reference signals.

That CRS structure provides channel estimation in each 0.5 ms regularly, 2 kHz. BD changes channel under Nyquist frequency 1 kHz.

Because the structure of the BD generated only two status "on" and "off", the waveform of AmBC is a square wave, toggle in that two status. Under the "on" status, the BD reflects ambient signals, while in the "off" status the BD absorbs the ambient signal. This BD waveform is called square-wave Binary FSK (SBFSK) [LWR+23]. Under constraint of a Nyquist frequency of 1 kHz, the SBFSK is selected at frequency key of f_0 and f_1.

UE channel estimator for the *n*-th subcarrier on *l*-th OFDM symbol is given by $\widehat{H}[n;l] = \frac{S_r[n;l]}{S_t[n;l]}$ where $S_r[n;l]$ and $S_t[n;l]$ denote the received signal and transmitted symbols, respectively. BD assumed to be close to Rx. We assume the AmBC information is only contained in the first tap of channel. In receiver, only the first tap of channel estimation is considered, in other words, mean of channel estimation in time domain: $\hat{h}[0;l]$. The receiver then uses a band pass filter to remove the doppler caused by environment from this tap.

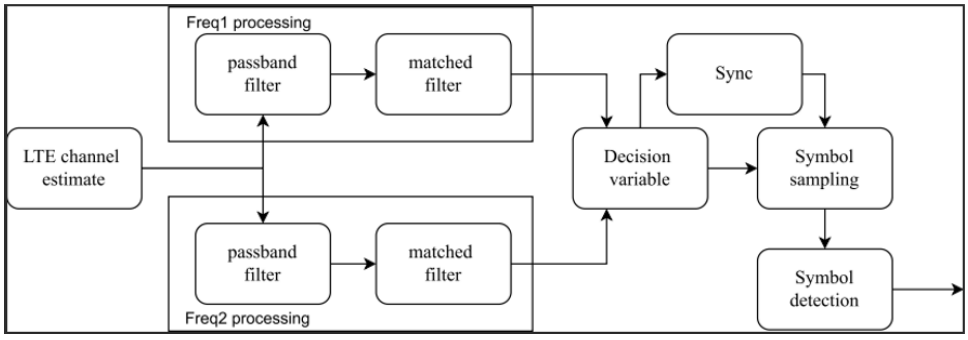


Figure 5-19: Flow chart of the proposed backscatter receiver.

The receiver shown in Figure 5-19 contains two parts, channel estimation and FSK demodulator. The channel estimation algorithm is initially built-in UE, due to LTE protocol [36.211]. The FSK demodulator is a software-defined radio. In principle, the proposed receiver could be implemented in existing LTE modems by firmware update.

The FSK demodulator includes two steps, band pass filter and matched filter. The purpose of band pass filter is to remove Doppler effect around the environment and aliasing caused by channel estimation speed. After filtering the interference, two matched filters make the decision of FSK coherent detector. The coherent detector requires a synchronization with BD and UE. In proposed AmBC system, a synchronization header and a data package compose one backscatter frame. The header is composed of three copies of 7-Barker codes where the last one is inverted.

5.3.2 Zero energy devices

Two zero-energy devices have been tested. The main components of the first zero-energy device are Analogue Devices ADP5091 low power energy harvester and STMicroelectronics SRM32L562 microcontroller. The energy harvester converts and store power from photovoltaic cell to a set of supercapacitors. The microcontroller uses low power timer to generate pulse width modulated (PWM) signal. By keeping the PWM signal duty cycle at 50% and changing the signal frequency, a binary frequency shift keying modulation can be implemented. Another timer is used for symbol timing. The timers used can run independently when the microcontroller is in low power state. Microcontroller can be kept in low power state for the duration of one symbol, and only be woken up to set the timers for new symbol. Thus, for the most of the transmission time the microcontroller is in a low power state.

Running from 3.3V source the microcontroller uses on average 6.6uA (22uW) current during the transmission with 40ms symbol length. 35mm x 21mm solar panel is enough to keep the device running on normal indoors office lightning. Four 0.1F supercapacitors are used for energy storage. The storage is enough to keep the device running for 36 hours with 10 second transmission interval.

The backscatter modulator (tag) implementation employs a RF switch Analogue Devices ADG919 connected to the RF antenna, switching the termination of the antenna between two loads Z_0 and Z_1. The switch consumes 165 nW while modulating. Block diagram and photo of the device is presented in Figure 5-20 below.

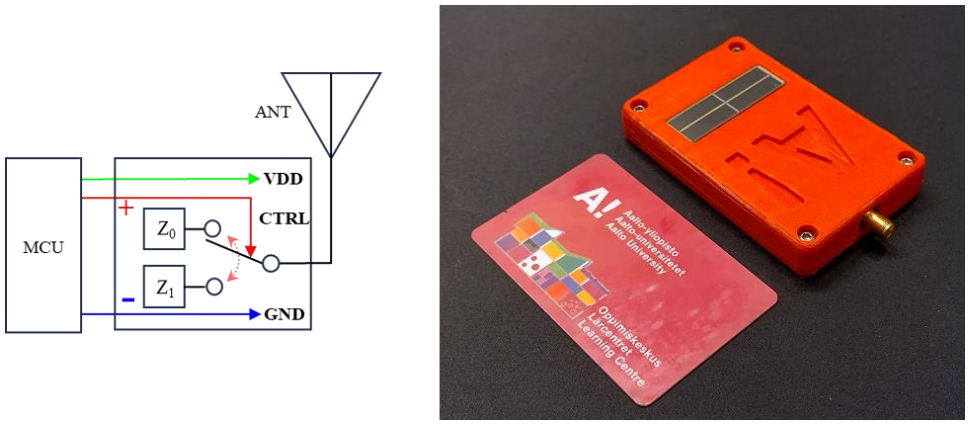


Figure 5-20: The first AIoT ZED device.

The second ZED prototype has been described in detail in [PBR+22]. As illustrated in Figure 5-21 below, a micro-controller (MSP430) controls all components. In particular, it controls an RF switch to set the two branches of dipole antenna into short circuit or open circuit. The MSP430 can be reprogrammed through a JTAG interface. The ZED harvest solar or indoor energy with a solar panel and stores it in a 3V battery coin-cell. The management of energy storage and energy consumption is performed by a dedicated component (BQ25570). In previous studies [HEX23-D73][PBR +22], it has been shown that this prototype of ZED can transmit a message periodically night and day and remain fully energy autonomous ‘infinitely’, under some particular assumptions (sending a message of 96 bits every 10 seconds, using FM0 modulation, 24 h/day and consuming 54 µW, and harvesting 150 µW during 10 hours of bright light per day). In Hexa-X II, the ZED is reprogrammed to support the modulation described in Section 5.3.1.

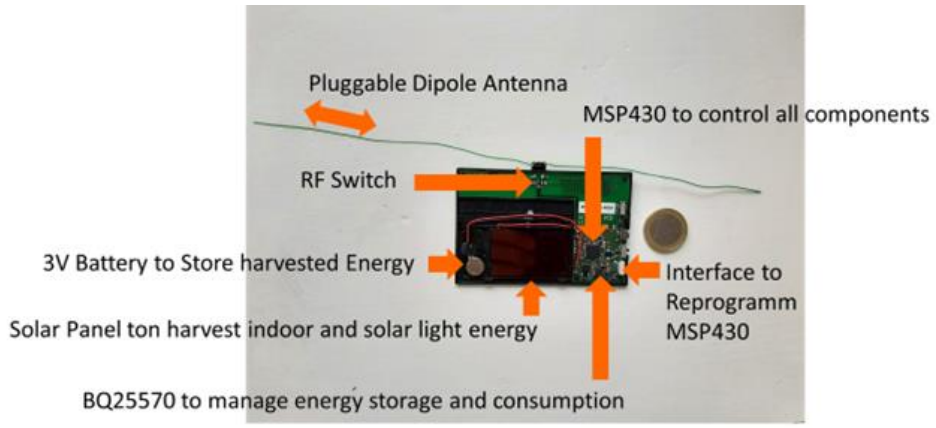


Figure 5-21: The second ZED device.

5.3.3 Laboratory tests and field trial

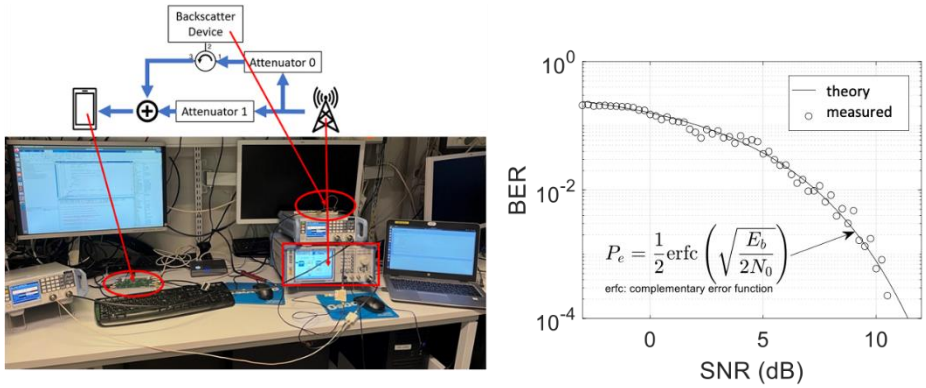


Figure 5-22: Measurement setup and results.

There is a demonstration of proposed AmBC, which proves the the feasibility of using the channel estimator at the UE as a receiver for backscatter signals in LTE downlink [LRJ+23]. The experiment is set up as Figure 5-22. LTE downlink signal is generated by a signal generator, mimicking the BS, and an USRP pretends to be the UE, including channel estimation. Attenuators play role of path loss in real world communication environment.

The experiment received information from AIoT, showing the proposed AmBC feasibility. The performance of the AmBC communication system matches the theoretical FSK bit per error (BER) in Figure 5-22.

After successfully testing the system in the laboratory setting, an experiment was conducted using real base stations. The details of the experiments are summarized in [NPH+23]. The system which is described in Sub-Sections 5.3.1 and which is validated in lab (as described in sub-section 5.3.4) is tested on the field with the ZED prototype presented in 5.3.3 backscattering ambient waves from Orange 4G commercial network to be read by the reader described in sub-section 5.3.1/5.3.4. For this purpose, the ZED is reprogrammed to support the FSK modulation specified by AAU in [LRJ+23]. For these trials, the ZED is programmed to send a frame constituted of a synchronisation sequence (for frame synchronisation) of 63 bits and a data sequence of 57 bits, continuously.

Figure 5-23 illustrates the measurement scenarios. The ZED-to-reader (UE) communication is tested at two different challenging positions: position#1 at 215 meters from a commercial 4G Base Station (BS) with an average 4G Signal-To-Noise Ratio (SNR) lower than 0dB (as roughly measured with a spectrum analyzer), position#2, with an average 4G SNR lower than 4dB, at 110 meters from the same 4G BS. In both cases, the ZED and the reader are very close to each other. In these initial measurements, the ratio of frames for which the synchronisation sequence has been detected is around 66% and 96%, for positions #1 and #2, respectively.

Furthermore, the measured data BER (averaged over detected frames) is 0.1 and 0.04, for positions #1 and #2, respectively.

These very early results will be extended in the future, with larger coverage measurement campaigns, with improved physical layer (synchronisation, channel coding etc.), larger ZED-to-reader distances, etc. Also, the application of this system to indoor localisation of UEs will be assessed.

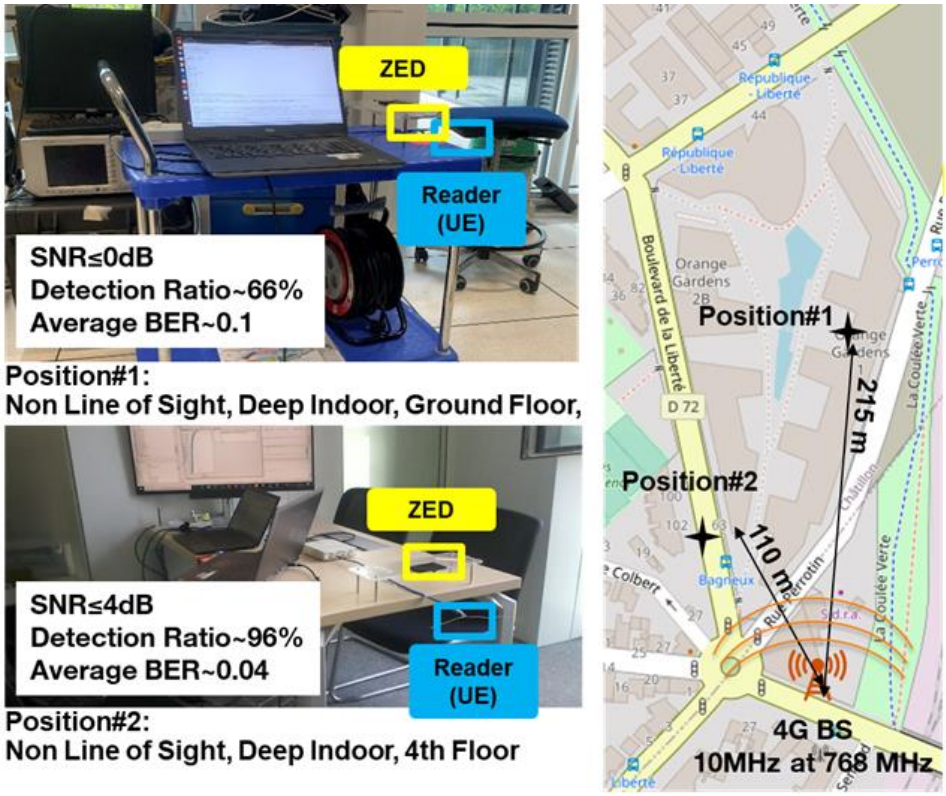


Figure 5-23: Measurement setups and first results.

6 Conclusions

In this deliverable, we have explored many different technologies in order to design efficient and novel 6G devices and infrastructure. Those technologies are clustered in four different directions: sub-THz transceivers, Reflective Intelligent Surfaces, System-on-Chip architectures, and ultra-low cost/power devices.

In Section 2, we have explored how to optimally dimension sub-THz architectures, based on scenario specifications, link budget analysis and a flexible power consumption model. Three types of architectures were considered: full digital, hybrid partially connected, and hybrid fully connected. Each of them can be used in the different hotspot scenarios considered. However, a proper dimensioning in number of antennas, PA output power and spectral efficiency is essential in order to obtain the best energy efficiency. This requires a proper balancing of the power consumed by PA, digital processing and analogue front-end.

The main models for analogue HW non-idealities were revisited, i.e., PA, phase noise, I/Q imbalance, DC offset, Sampling Clock Offset, Carrier Frequency Offset, ADC/DAC, and phase shifter. While some state-of-the-art models can be reused, there are also differences coming from the specificities of sub-THz architectures, i.e., higher carrier frequency, wider bandwidth, higher number of antennas and different architectures. The main extensions to investigate relate to PA memory, updated phase noise, frequency-selective I/Q, refined ADC and phase shifting non-idealities, and random per-antenna differences for PA, phase noise and I/Q.

The role of LO routing on phase noise contribution on wideband arrays was also investigated, showing how throughput can be phase noise limited and not SNR-limited, depending on bandwidth. Asymmetries in LO routing to the different antennas creates specific effects and frequency-dependent phase noise spectra.

Architectures based on Resonant Tunneling Diodes were also explored. Those devices are especially appealing for high frequencies, in the sub-THz to THz range, and they can lead to simple architectures. The main design challenges relate to the output power limitations. They hence require high-efficiency antenna integration and the use of antenna arrays. Simpler modulations are also considered for those architectures.

Finally, switched-beam antenna lenses were investigated as an alternative to phased antenna arrays. By placing multiple antenna ports next to the lens and selecting them adaptively, the lens gain can be combined with angular adaptivity. When the required EIRP and hence array size increases, such architectures promise a reduced power consumption and array size as compared to their phased-array counterparts. The main research challenges relate to the lens integration on top of RF transceivers.

In Section 3, Reflective Intelligent Surfaces are investigated, as they can enhance the connectivity at high frequencies due to the LOS-dominated propagation behaviour. A simulation environment was used to model how the RIS can improve the system performance. Two prototypes were used to validate the results, based on active and passive configurations, and considering classical or varactor-based implementations. A very close match was observed between the simulated and measured radiation patterns. The phase reflection coefficients and related 3D patterns were simulated as function of the frequency and capacity.

RIS integration challenges were also considered. Indeed, control aspects are essential to consider in order to obtain a working solution. Control needs to satisfy a number of properties related to angular accuracy, timing or general constraints such as power consumption. Different control functions are also required and have been listed. The control can be triggered from the infrastructure, from the UEs or from external devices.

In section 4, the aim was the research and development of energy-efficient 6G system-on-chip (SoC) architectures enabling scalable signal and AI processing, HW/SW concepts for trusted SoC to protect against potential attacks and multi-source power harvesting and management.

This work addressed the feasibility of using RISC-V for 6G signal processing, focusing on representative signal processing applications and their basic kernels. The study encompassed the vectorization of these kernels specifically targeting RISC-V vector extensions, followed by an analysis of the performance improvements achieved through the vectorization approach. A comprehensive analysis of the overall signal processing benchmark in conjunction with subframe processing enabled identification of critical gaps for further algorithm-hardware-software optimizations.

Another aspect involved the design and integration of AI components into SoC. A flexible and scalable 3D-array AI-accelerator architecture employing bit-serial processing was proposed enabling dynamically

adaptation to workload precision requirements. The next challenge is in the integration of the AI accelerator into a secure SoC platform and assessment of the impact of hardware/software overhead on performance and cost.

Investigation of security and privacy aspects at the System-on-Chip (SoC) level was another aspect of study. A scalable SoC architecture adopting a tiled structure wherein physically distinct tiles interconnected through a network-on-chip (NoC) was proposed. Each tile integrates a trusted communication unit (TCU) responsible for isolation among tiles. The OS microkernel take charge of managing communication channels and TCUs, ensuring secure and reliable communication throughout the platform. The FPGA prototype was designed, featuring multiple processing tiles equipped with RISC-V cores. Evaluation results showcased that the TCU configuration incurred minimal resource overhead, utilizing only a fraction of FPGA resources. Furthermore, latency assessments indicated that the introduction of TCU functionalities and security features resulted in negligible delays.

Energy harvesting (EH) was another important aspect of 6G IoT devices. In this part of work a power management integrated circuit (PMIC) was developed that could work seamlessly with multiple energy sources simultaneously. In addition, we investigated low-power ML algorithms for EN and EH devices to improve prediction of energy availability and consumption and enable energy-aware schedulers and reconfigurable energy buffers. Proposed approach, such as harvest rate sensing, energy buffer estimation, and an ultra-low-power load monitoring module, achieves high efficiency in transferring energy to the main buffer and estimated EH rates. The module is feasible to track the energy status of the load in real-time and determine the start and end times of tasks with minimal overhead.

We explored next-generation ultra-low power/cost IoT devices in several aspects. Specifically, we discussed their power consumption dynamics (unravelling the intricacies of balancing performance and longevity), the challenges posed by energy constraints, and the imperative to optimize devices for sustainable operation through a comprehensive analysis. A generic methodology for evaluating the 6G IoT modem power consumption was proposed and also several channel codes suitable for low-power/cost devices were overviewed. Key enabling technologies redefining the possibilities of such devices such as EH, RF-WPT, energy-aware and lightweight protocols, TinyML, intelligent wake-up, and AmBC, were revised, related novel results discussed, and relevant research directions highlighted. Key take-aways are: i) the EH technology suitable for a particular use case depends on the availability, intensity, and type of ambient energy and the devices' form factor constraints and energy demands; ii) RF-WPT may support the sustainable deployment of massive ultra-low power IoT deployments with QoS guarantees but end-to-end efficiency issues must be tamed, especially considering novel low-power MIMO techniques; iii) communication (but also computation and sensing) protocols for ultra-low power/cost IoT devices must be lightweight, adaptive, and manage energy resources based on availability and demand patterns; iv) TinyML may be affordable for some IoT devices, and if so, several benefits in terms of privacy and security, and performance gains, including energy-saving capabilities, may be realized; and v) next-generation of wake-up radios and protocols must be more intelligent. Finally, regarding AmBC, we presented a zero-energy PoC, wherein IoT tag devices harness ambient RF cellular signals for backscatter communication/reporting. This demonstrated the feasibility of perpetual, self-sustaining IoT devices in the real world, although much work is still needed before widespread deployment and commercialization.

References

[21.914] 3GPP, TR 21.914 “Release 14 Description; Summary of Rel-14 Work Items (Release 14)” v14.0.0, June 2018

[21.915] 3GPP, TR 21.915 “Release 15 Description; Summary of Rel-15 Work Items (Release 15)” v15.0.0, October 2019

[21.916] 3GPP, TR 21.916 “Release 16 Description; Summary of Rel-16 Work Items (Release 16)” v16.2.0, June 2022

[21.917] 3GPP, TR 21.917 “Release 17 Description; Summary of Rel-17 Work Items (Release 17)” v17.0.1, January 2023

[22.840] 3GPP, TR 22.840, “Study on Ambient power-enabled Internet of Things (Release 19)” V2.0.0, September 2023

[36.212] 3GPP, TS 36.212, “Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding” v18.0.0, September 2023

[36.888] 3GPP TR 36.888, “Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE (Release 12)”, June 2013.

[38.848] 3GPP, TR 38.848 “Study on Ambient IoT (Internet of Things) in RAN (Release 18)” v18.0.0, September 2023

[38.840] 3GPP, TR 38.840 “Study on User Equipment (UE) power saving in NR (Release 16)” v16.0.0, June 2019

[AAK+21] H. Afzal, R. Abedi, R. Kananizadeh, P. Heydari and O. Momeni, "An mm-Wave Scalable PLL-Coupled Array for Phased-Array Applications in 65-nm CMOS," in IEEE Transactions on Microwave Theory and Techniques, vol. 69, no. 2, pp. 1439-1452, Feb. 2021, doi: 10.1109/TMTT.2020.3039517.

[AB14] A. Acharyya and J. P. Banerjee, ”Prospects of IMPATT devices based on wide bandgap semiconductors as potential terahertz sources,” Applied Nanoscience, vol. 4, no. 1, pp. 1-14, 2014, doi: 10.1007/s13204-012-0172.

[AKO+16] K. Alharbi, A. Khalid, A. Ofiare, J. Wang, E Wasige, “Diced and grounded broadband bow-tie antenna with tuning stub for resonant tunnelling diode terahertz oscillators”, IET Microwaves, Antennas & Propagation, 2016

[ALC+23] A. Azarbahram, O. Lopez, B. Clerckx, and M. Latva-Aho, “Waveform and Beamforming Optimization for Wireless Power Transfer with Dynamic Metasurface Antennas”, 2023

[ALJ19] M. U. Aminu, J. Lehtomäki and M. Juntti, "Beamforming and Transceiver Optimization with Phase Noise for mmWave and THz Bands," 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, 2019, pp. 692-696, doi: 10.1109/ISWCS.2019.8877230.

[ALP+23] A. Azarbahram, O. Lopez, P. Popovski, M. and Latva-aho, “On the Radio Stripe Deployment for Indoor RF Wireless Power Transfer”, 2023

[ALS+23] A. Azarbahram, O. Lopez, R. Souza, R. Zhang, R., and M. Latva-Aho, “Energy Beamforming for RF Wireless Power Transfer with Dynamic Metasurface Antennas”, 2023

[APK+23] G. C. Alexandropoulos, D. T. Phan-Huy, K. D. Katsanos, M. Crozzoli, H.Wymeersch, P. Popovski, P. Ratajczak, Philippe Y. Bénédic, M. H. Hamon, S. H. Gonzalez, P. Mursia, M. Rossanese, V. Sciancalepore, J. B. ,Gros, S. Terranova, G. Gradoni, P. Di Lorenzo, M. Rahal, B. Denis, R. D’Errico, A. Clemente, and E. C. Strinati, ”RIS-enabled smart wireless environments: deployment scenarios, network architecture, bandwidth and area of influence,” J Wireless Com Network, 103, 2023. https://doi.org/10.1186/s13638-023-02295-8

[AS16] M. Asada, and S. Suzuki, “Terahertz Emitter Using Resonant-Tunneling Diode and Applications”, MDPI, Sensors 2021.

[AS21] deliM. Asada and S. Suzuki, ”Terahertz Emitter Using Resonant-Tunneling Diode and Applications,” Sensors, vol. 21, no. 4, pp. 1384, 2021, doi: 10.3390/s21041384

[Asa16] K. Asanovic, R. Avizienis, J. Bachrach, S. Beamer, D. Biancolin, C. Celio, H. Cook, D. Dabbelt, J. Hauser, A. Izraelevitz, S. Karandikar, B. Keller, D. Kim, J. Koenig, “The Rocket Chip Generator,” EECS, University of California at Berkeley, Tech. Rep. UCB/EECS-2016-17, 2016

[AHW+22] N. Asmussen, S. Haas, C. Weinhold, T. Miemietz, M. Roitzsch, “Efficient and Scalable Core Multiplexing with M³v”, ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2022

[ASC+23] J. Anuj, A. Sabovic, B. Celikkol, M. Aernouts, P. Reiter, S. Mercelis, P. Hellinckx, and J. Famaey, "An Energy Management Unit for Predictive Solar Energy Harvesting IoT," In Proceedings of the 8th International Conference on Internet of Things, Big Data and Security, 2023 - IoTBDS; ISBN 978-989-758-643-9; ISSN 2184-4976, SciTePress, pages 39-50. DOI: 10.5220/0011839500003482

[ATS+21] A. Arora, C. G. Tsinos, B. Shankar Mysore R, S. Chatzinotas and B. Ottersten, "Analogue Beamforming with Antenna Selection For Large-Scale Antenna Arrays," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 4795-4799, doi: 10.1109/ICASSP39728.2021.9414673.

[AWW18] A. Al-Khalidi, J. Wang and E. Wasige, "Compact J-band Oscillators With lm RF output Power and Over 110 GHz Modulation Bandwidth," 2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), Nagoya, 2018, pp. 1-2.

[BCM01] N. J. Bershad, P. Celka and S. McLaughlin, "Analysis of stochastic gradient identification of Wiener-Hammerstein systems for nonlinearities with Hermite polynomial expansions," in IEEE Transactions on Signal Processing, vol. 49, no. 5, pp. 1060-1072, May 2001.

[BJK+23] J. Bang, S. Jung, J. Kim, S. Park and J. Choi, "A Sub-50-fs RMS Jitter, 103.5-GHz Fundamental-Sampling PLL With an Extended Loop Bandwidth," in IEEE Solid-State Circuits Letters, vol. 6, pp. 201-204, 2023.

[BKJ+23] J. Bang, J. Kim, S. Jung, S. Park and J. Choi, "A 47fsrms-Jitter and 26.6mW 103.5GHz PLL with Power-Gating Injection-Locked Frequency-Multiplier-Based Phase Detector and Extended Loop Bandwidth," 2023 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023.

[BRB+22] J. -M. Baracco, P. Ratajczak, P. Brachat, J. -M. Fargeas and G. Toso, "Ka-Band Reconfigurable Reflectarrays Using Varactor Technology for Space Applications: A proposed design," in IEEE Antennas and Propagation Magazine, vol. 64, no. 1, pp. 27-38, Feb. 2022.

[BTO22] T. Basaklar, Y. Tuncel, and U. Y. Ogras, "tinyMAN: Lightweight energy manager using reinforcement learning for energy harvesting wearable IoT devices," arXiv preprint arXiv:2202.09297, 2022.

[CAA+18] P. Chen, S. Alsahali, A. Alt, J. Lees and P. J. Tasker, "Behavioral Modeling of GaN Power Amplifiers Using Long Short-Term Memory Networks," 2018 International Workshop on Integrated Nonlinear Microwave and Millimetre-wave Circuits (INMMIC), Brive La Gaillarde, France, 2018.

[CAL+22] S. Callender, A. Whitcombe, A. Agrawal, R. Bhat, M. Rahman, C. C. Lee, P. Sagazio, G. Dogiamis, B. Charton, M. Chakravorti, S. Pellerano and C. Hull, "A Fully Integrated 160-Gb/s D-Band Transmitter Achieving 1.1-pJ/b Efficiency in 22-nm FinFET," in IEEE Journal of Solid-State Circuits, vol. 57, no. 12, pp. 3582-3598, Dec. 2022.

[CBV01] P. Celka, N. J. Bershad and J. -M. Vesin, "Stochastic gradient identification of polynomial Wiener systems: analysis and application," in IEEE Transactions on Signal Processing, vol. 49, no. 2, pp. 301-313, Feb. 2001.

[CHK+17] J. Chen, Z. S. He, D. Kuylenstierna, T. Eriksson, M. Hörberg, T. Emanuelsson, T. Swahn, and H. Zirath, "Does LO Noise Floor Limit Performance in Multi-Gigabit Millimeter-Wave Communication?," in IEEE Microwave and Wireless Components Letters, vol. 27, no. 8, pp. 769-771, Aug. 2017, doi: 10.1109/LMWC.2017.2724853.

[CRD+10] J. Verlant-Chenet, J. Renard, J. -M. Dricot, P. De Doncker and F. Horlin, "Sensitivity of Spectrum Sensing Techniques to RF Impairments," 2010 IEEE 71st Vehicular Technology Conference, Taipei, Taiwan, 2010, pp. 1-5, doi: 10.1109/VETECS.2010.5493999.

[CSZ+20] M. Cavalcante, F. Schuiki, F. Zaruba, M. Schaffner, and L. Benini, “Ara: A 1-GHz+ scalable and energy-efficient RISC-V vector processor with multiprecision floating-point support in 22-nm FD-SOI,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst. , vol. 28, no. 2, pp. 530–543, Feb.2020.

[CXS+20] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, “A Survey of Accelerator Architectures for Deep Neural Networks”,Elsevier Engineering, Volume 6, Issue 3, 2020, Pages 264-274,

[DCS+21] C. Desset, N. Collaert, S. Sinha, and G. Gramegna, “InP / CMOS co-integration for energy efficient sub-THz communication systems,” In Globecom Workshop on Emerging Topics in 6G Communications, December 2021.

[DDL15] B. Debaillie, C. Desset and F. Louagie, "A flexible and future-proof power model for cellular base stations", In VTC Spring, Glasgow, Scotland, May 2015.

[Dem06] A. Demir, “Computing Timing Jitter From Phase Noise Spectra for Oscillators and Phase-Locked Loops With White and1/f Noise,” in IEEE Transactions on Circuits and Systems I: Regular Papers, Sept. 2006.

[DF03] C. Desset, and A. Fort, “Selection of channel coding for low-power wireless systems,” In The 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring, vol. 3, pp. 1920-1924. April 2003.

[DHC+17] H, Debin, W. Hong, J. Chen, P Yan, and Y. Xiong, “A compact d-band I/Q mixer with im- proved transformer balun,” Microwave and Optical Technology Letters, 2017.

[DPS20] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR - The Next Generation Wireless Access Technology*, 2nd Edition, 1. Chapter 26, §26.2: “mmW LO generation and phase noise aspects” - September 18, 2020 ISBN: 9780128223208.

[DWZ+20] C. Desset, P. Wambacq, Y. Zhang, M. Ingels, and A. Bourdoux, “A flexible power model for mm-wave and THz high-throughput communication systems,” In PIMRC Workshop on Enabling Technologies for Terahertz Communications (ETTCOM), London, UK, August 2020.

[DZM+04] L. Ding, G. T. Zhou, D. R. Morgan, Z. Ma, J. S. Kenney, J. Kim and C. R. Giardina, "A robust digital baseband predistorter constructed using memory polynomials," in IEEE Transactions on Communications, vol. 52, no. 1, pp. 159-165, Jan. 2004.

[Eis10] H. Eisele, ”480 GHz oscillator with an InP Gunn device,” Electronics Letters, vol. 46, no. 6, pp. 422-423, 2010, doi: 10.1049/el.2010.3362

[FFB+19] A. Frøytlog, T. Foss, O. Bakker, G. Jevne, M. Haglund, F. Li, J. Oller, G. Li, “Ultra-Low Power Wake-up Radio for 5G IoT”, IEEE Communications Magazine. PP. 1-7, 2019. 10.1109/MCOM.2019.1701288.

[Fri23] S. Friedrich, S. B. Sampah, R. Wittig, M. R. Vemparala, N. Fasfous, E. Matúš, W. Stechele, and G. Fettweis, “Lightweight Instruction Set for Flexible Dilated Convolutions and Mixed-Precision Operands,” in Proceedings of 24th International Symposium on Quality Electronic Design (ISQED 2023), San Francisco, USA, Apr 2023.

[FSC+11] M. Feiginov, C. Sydlo, O. Cojocari, and P. Meissner, “Resonant-tunnelling diode oscillators operating at frequencies above 1.1 THz,” Applied Physics Letters, vol. 99, no. 23, p. 233506, Dec. 2011

[FU98] G. D. Forney and G. Ungerboeck, “Modulation and coding for linear Gaussian channels”, IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2384-2415, October 1998.

[GH09] F. M. Ghannouchi and O. Hammi, "Behavioral modeling and predistortion," in IEEE Microwave Magazine, vol. 10, no. 7, pp. 52-64, Dec. 2009.

[GS91] A. Ghorbani and M. Sheikhan, “The effect of solid-state power amplifiers (SSPAs) nonlinearities on MPSK and m-QAM signal transmission,” in6th Int. Conf. on Digital Processing of Signals in Communications, Loughborough UK, p. 193–197, 1991.

[HA22] S. Haas, N. Asmussen: “A Trusted Communication Unit for Secure Tiled Hardware Architectures”, 29th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), 2022

[HB08] F. Horlin, A. Bourdoux, “Digital Compensation for Analogue Front-Ends : A New Approach to Wireless Transceiver Design”, 2008

[HBQ+21] L. Heim, A. Biri, Z. Qu, L .Thiele, "Measuring what really matters: Optimizing neural networks for tinyml," arXiv preprint arXiv:2104.10645, 2021.

[HEX20-D22] Hexa-X Deliverable D2.2 “Initial radio models and analysis towards ultra-high data rate links in 6G,” Eds. Marko E. Leinonen, Dec. 2021.

[HEX20-D23] Hexa-X Deliverable D2.3, “Radio models and enabling techniques towards ultra-high data rate links and capacity in 6G,” 2023.

[HEX21-D22] Hexa-X, “deliverable D2.2 Initial radio models and analysis towards ultra-high data rate links in 6G,” Dec. 2021.

[HEX223-D12] Hexa-X-II deliverable D1.2 “6G Use Cases and Requirements,” December 2023.

[HEX223-D52] Hexa-X-II deliverable D5.2 “Characteristics and classification of 6G device classes,” October 2023.

[HEX23-D23] Hexa-X deliverable D2.3 “Radio models and enabling techniques towards ultra-high data rate links and capacity in 6G,” March 2023.

[HEX23-D73] Hexa-X Deliverable D7.3 "Special-purpose functionalities: final solutions" 31/05/2023, available at https://hexa-x.eu/wp-content/uploads/2023/06/Hexa-X_D7.3_v1.0.pdf.

[HH97] M. Honkanen and S.-G. Haggman, “New aspects on nonlinear power amplifier modeling in radio communication system simulations,” in Proceedings of 8th International Symposium on Personal, Indoor and Mobile Radio Communications - PIMRC ’97, vol. 3, pp. 844-848 vol.3, 1997.

[HJY+17] W. Hong, Z. H. Jiang, C. Yu, J. Zhou, P. Chen, Z. Yu, H. Zhang, B. Yang, X. Pang, M. Jiang, J. Chen and S. He, "Multibeam Antenna Technologies for 5G Wireless Communications," in *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6231-6249, Dec. 2017.

[HL98] A. Hajimiri and T. H. Lee, “A general theory of phase noise in electrical oscillators,” in IEEE Journal of Solid-State Circuits, vol. 33, no. 2, pp. 179-194, Feb. 1998.

[HLY+22] X. Hu, Z. Liu, X. Yu, Y. Zhao, W. Chen, B. Hu, X- Du, Z. Li, M. Helaoui, W. Wang and F. M. Ghannouchi, "Convolutional Neural Network for Behavioral Modeling and Predistortion of Wideband Power Amplifiers," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 8, pp. 3923-3937, Aug. 2022.

[HPK+22] H. Han, J. Park, J. Kim, K. Bae and I. Na, "Baseband Phase Noise Modeling and Analysis for 140GHz THz DFT-s-OFDM System," 2022 IEEE Globecom Workshops (GC Wkshps), Rio de Janeiro, Brazil, 2022, pp. 1748-1753, doi: 10.1109/GCWkshps56602.2022.10008595.

[IEEE99] IEEE 802.11a-1999, "High-speed Physical Layer in the 5 GHz band", Nov. 1999

[IEEE06] IEEE 802.15-06-0477-01-003c, “RF impairment models for 60GHz-band SYS/PHY simulation”, Nov. 2006

[IME+23] Nadine Collaert, “Scaling up GaN- and InP-based technologies for 5G and 6G wireless communication”, https://www.imec-int.com/en/articles/scaling-gan-and-inp-based-technologies-5g-and-6g-wireless-communication , 2023

[INT+07] D. Inti, “Time-Varying Frequency Selective IQ Imbalance Estimation and Compensation”, master thesis, 2017

[ISA17] R. Izumi, S. Suzuki and M. Asada, “1.98 THz resonant-tunneling-diode oscillator with reduced conduction loss by thick antenna electrode.” 42nd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), 2017

[JPO+22] S. Jia, X. Pang, O. Ozolins, X. Yu, H. Hu, J. Yu, P. Guan, F. Da Ros, S.Popov, G. Jacobsen, M. Galili, T. Morioka, D. Zibar, and L. K. Oxenløwe,”0.4 THz Photonic-Wireless Link With 106 Gb/s Single Channel Bitrate,” Journal of Lightwave Technology, vol. 36, no. 2, pp. 610-615, 2018, doi:10.1109/JLT.2017.2776320

[KAW20] A. Al-Khalidi, K. H. Alharbi, J. Wang, R. Morariu, L. Wang, A. Khalid, J. Figueiredo, and E. Wasige “Resonant Tunnelling Diode Terahertz Sources with up to 1 mW Output Power in the J-Band”, IEEE Trans. on THz Science and Tech, Vol 10, Issue 2, March 2020, pp. 150-157.

[KK03] H. Ku and J. S. Kenney, "Behavioral modeling of nonlinear RF power amplifiers considering memory effects," in IEEE Transactions on Microwave Theory and Techniques, vol. 51, no. 12, pp. 2495-2504, Dec. 2003.

[KKP+14] M. R. Khanzadi, D. Kuylenstierna, A. Panahi, T. Eriksson and H. Zirath, “Calculation of the Performance of Communication Systems From Measured Oscillator Phase Noise” in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 61, no. 5, pp. 1553-1565, May 2014.

[KKY+22] Y. Koyama, Y. Kitazawa, K. Yukimasa, T. Uchida, T. Yoshioka, K. Fujimoto, T. Sato, J. Iba, K. Sakurai, and T. Ichikawa, “A High-Power Terahertz Source Over 10 mW at 0.45 THz Using an Active Antenna Array With Integrated Patch Antennas and Resonant-Tunneling Diodes”, IEEE Trans. on THz Science and Tech., vol. 12, no. 5, sept 2022.

[KSO13] Y. Koyama, R. Sekiguchi, and T. Ouchi, “Oscillations up to 1.40 THz from Resonant-Tunneling-Diode-Based Oscillators with Integrated Patch Antennas,” Applied Physics Express, vol. 6, no. 6, p. 064102, Jun. 2013.

[LAS+21] O. L. A. López, H. Alves, R. D. Souza, S. Montejo-Sánchez, E. M. G. Fernández and M. Latva-Aho, "Massive Wireless Energy Transfer: Enabling Sustainable IoT Toward 6G Era," in IEEE Internet of Things Journal, vol. 8, no. 11, pp. 8816-8835, 1 June1, 2021, doi: 10.1109/JIOT.2021.3050612.

[LBG04] T. Liu, S. Boumaiza and F. M. Ghannouchi, "Dynamic behavioral modeling of 3G power amplifiers using real-valued time-delay neural networks," in IEEE Transactions on Microwave Theory and Techniques, vol. 52, no. 3, pp. 1025-1033, March 2004.

[LH89] O. Lodge, J. L. Howard, “On the concentration of electric radiation by lenses”, in Nature, MacMillan and Co. 40:94, 1889.

[LIA+23] Y. Liang, C. C. Boon, G. Qi, G. Dziallas, D. Kissinger, H. J. Ng, P. I. Mak and Y. Wang, "A Low-Jitter and Low-Reference-Spur 320 GHz Signal Source With an 80 GHz Integer-N Phase-Locked Loop Using a Quadrature XOR Technique," in IEEE Transactions on Microwave Theory and Techniques, vol. 70, no. 5, pp. 2642-2657, May 2022.

[LKS+22] O. L. A. López, D. Kumar, R. D. Souza, P. Popovski, A. Tölli and M. Latva-Aho, "Massive MIMO With Radio Stripes for Indoor Wireless Energy Transfer," in IEEE Transactions on Wireless Communications, vol. 21, no. 9, pp. 7088-7104, Sept. 2022, doi: 10.1109/TWC.2022.3154428.

[LL20] X. Liu and H. C. Luong, "A Fully Integrated 0.27-THz Injection-Locked Frequency Synthesizer With Frequency-Tracking Loop in 65-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 55, no. 4, pp. 1051-1063, April 2020, doi: 10.1109/JSSC.2019.2954232.

[LPA+19] G. LaCaille, A. Puglielli, E. Alon, B. Nikolic, and A. Niknejad “Optimizing the lo distribution architecture of mm-wave massive mimo receivers,” arXiv: Signal Processing. November 2019.

[LPT+13] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, “Ambient backscatter: Wireless communication out of thin air,” ACM SIGCOMM computer communication review, vol. 43, no. 4, pp. 39–50, 2013.

[LRJ+23] J. Liao, K. Ruttik, R. Jäntti, and D.-T. Phan-Huy. 2023. “Demo: UE Assisted Ambient IoT in LTE Downlink, in real-time and open source,” In Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23), 588–589, June 2023.

[LRR+23] O. L. A. López, O. M. Rosabal, D. E. Ruiz-Guirola, P. Raghuwanshi, K. Mikhaylov, L. Lovén, and S. Iyer, "Energy-Sustainable IoT Connectivity: Vision, Technological Enablers, Challenges, and Future Directions," in IEEE Open Journal of the Communications Society, vol. 4, pp. 2609-2666, 2023, doi: 10.1109/OJCOMS.2023.3323832.

[LWR+23] J. Liao, X. Wang, K. Ruttik, R. Jäntti, D.-T. Phan-Huy, ”In-Band Ambient Backscatter Communications Leveraging LTE Cell-Specific Reference Signals,” IEEE Journal of Radio Frequency Identification, May 2023.

[LYZ+08] T. Liu, Y. Ye, X. Zeng and F. M. Ghannouchi, "Memory Effect Modeling of Wideband Wireless Transmitters Using Neural Networks," 2008 4th IEEE International Conference on Circuits and Systems for Communications, Shanghai, China, 2008.

[LZH+15] W. Li, Y. Zhang, L. -K. Huang, J. Cosmas, C. Maple and J. Xiong, "Self-IQ-Demodulation Based Compensation Scheme of Frequency-Dependent IQ Imbalance for Wideband Direct-Conversion Transmitters," in IEEE Transactions on Broadcasting, vol. 61, no. 4, pp. 666-673, Dec. 2015, doi: 10.1109/TBC.2015.2465138.

[Maz75] J. E. Mazo, “Faster-than-Nyquist signaling”, Bell System Technical Journal, vol. 54, no. 8, pp. 1451-1462, October 1975.

[Mau14] T. Maudoux, "Optimizing your power amplifier for predistortion with RF PA linearizer(RFPAL)", 2014

[MBA+22] A. Mahmood, L. Beltramelli, S. F. Abedin, S. Zeb, N. I. Mowla, S. A. Hassan, E. Sisinni, and M. Gidlund "Industrial IoT in 5G-and-Beyond Networks: Vision, Architecture, and Design Trends," in IEEE Transactions on Industrial Informatics, vol. 18, no. 6, pp. 4122-4137, June 2022, doi: 10.1109/TII.2021.3115697.

[MJS+16] M. Magno, V. Jelicic, B. Srbinovski, V. Bilas, E. Popovici, and L. Benini,"Design, implementation, and performance evaluation of a flexible low-latency nanowatt wake-up radio receiver," IEEE Trans. Ind. Informat.,vol. 12, no. 2, pp. 633-644, Apr. 2016.

[MK20] K. Mikhaylov and H. Karvonen, "Wake-up radio enabled BLE wearables: Empirical and analytical evaluation of energy efficiency," In 2020 14th International Symposium on Medical Information Communication Technology (ISMICT), pp. 1-5, May 2020.

[MLC+23] B. -T. Moon, S. -G. Lee and J. Choi, "24.2 A 264-to-287GHz, −2.5dBm Output Power, and −92dBc/Hz 1MHz-Phase-Noise CMOS Signal Source Adopting a 75fsrms Jitter D-Band Cascaded Sub-Sampling PLL," 2023 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023.

[MMK+06] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt and J. Pastalan, "A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers," in IEEE Transactions on Signal Processing, vol. 54, no. 10, pp. 3852-3860, Oct. 2006

[MSD+21] G. Mohiuddin, M. M. Shaikh, S. A. Dahri, F. Panhwar, K. A. Memon, and N. Madina, "Characterization of RF Power Amplifier for Narrow and Wide Band Memory Polynomial Implementations, " Quaid-e-Awam University Research Journal of Engineering Science & Technology 2021.

[MSL+17] I. Mehdi, J. V. Siles, C. Lee, and E. Schlecht, "THz Diode Technology: Status, Prospects, and Applications," Proceedings of the IEEE, vol. 105, no. 6, pp. 990-1007, 2017, doi: 10.1109/JPROC.2017.2650235

[NBM+23] S. Naser, L. Bariah, S. Muhaidat, and E. Basar. "Zero-energy devices empowered 6G networks: Opportunities, key technologies, and challenges." (2023).

[NDR16] T. Nagatsuma, G. Ducournau, and C. C. Renaud, "Advances in terahertz communications accelerated by photonics," Nature Photonics, vol. 10, no.6, pp. 371-379, 2016, doi: 10.1038/nphoton.2016.65.

[NND+18] N. Nishigami, Y. Nishida, S. Diebold, J. Kim, M. Fujita, and T. Nagatsuma, "Resonant Tunneling Diode Receiver for Coherent Terahertz Wireless Communication," 2018 Asia-Pacific Microwave Conference (APMC), Kyoto, 2018, pp. 726-728.

[NND+19] Y. Nishida, N. Nishigami, S. Diebold, J. Kim, M. Fujita, and T. Nagatsuma, "Terahertz coherent receiver using a single resonant tunnelling diode", Nature research-Scientific reports, 2019.

[NPH+23] I. Ndiaye, D.-T. Phan-Huy, A. Hassan, J. Liao, X. Wang, K. Ruttik, R. Jäntti "Zero-Energy-Device for 6G: First Real-Time Backscatter Communication thanks to the Detection of Pilots from an Ambient Commercial Cellular Network" accepted to 6GNet 2023, Paris, 18-20 Oct. 2023.

[NPK+08] J. Nishizawa, P. Płotka, T. Kurabayashi, and H. Makabe, "706-GHz GaAs CW fundamental-mode TUNNETT diodes fabricated with molecular layer epitaxy," physica status solidi (c), vol. 5, no. 9, pp. 2802-2804, 2008, doi: 10.1002/pssc.200779256

[NS22] S. G. Neelam and P. R. Sahu, "Digital Compensation of IQ Imbalance, DC Offset for Zero-Padded OTFS Systems," in IEEE Communications Letters, vol. 26, no. 10, pp. 2450-2454, Oct. 2022, doi: 10.1109/LCOMM.2022.3190047.

[NZ01] A. E. Nordsjo and L. H. Zetterberg, "Identification of certain time-varying nonlinear Wiener and Hammerstein systems," in IEEE Transactions on Signal Processing, vol. 49, no. 3, pp. 577-592, March 2001.

[ODC+16] J. Oller, I. Demirkol, J. Casademont, J. Paradells, G. U. Gamm, and L. Reindl, "Has time come to switch from duty-cycled MAC protocols towake-up radio for wireless sensor networks?" IEEE Trans. Netw., vol.24, no. 2, pp. 674-687, Apr. 2016.

[OHS+16] N. Oshima, K. Hashimoto, S. Suzuki, and M. Asada, "Wireless data transmission of 34 Gbit/s at a 500-GHz range using resonant tunnelling diode terahertz oscillator," Electron. Lett., vol. 52, no. 22, pp. 1897–1898, Oct. 2016.

[OHS+17] N. Oshima, K. Hashimoto, S. Suzuki and M. Asada, "Terahertz Wireless Data Transmission With Frequency and Polarization Division Multiplexing Using Resonant-Tunneling-Diode Oscillators," in IEEE Transactions on Terahertz Science and Technology, vol. 7, no. 5, pp. 593-598, Sept. 2017

[OKO+15] K. Okada, K. Kasagi, N. Oshima, S. Suzuki, and M. Asada, "Resonant-Tunneling-Diode Terahertz Oscillator Using Patch Antenna Integrated on Slot Resonator for Power Radiation", IEEE Trans. on THz Science and Tech.,vol. 5, No 4, July 2015.

[OML09] M. O'Droma, S. Meza, and Y. Lei, "New modified Saleh models for memoryless nonlinear power amplifier behavioural modelling," IEEE Communications Letters, vol. 13, no. 6, pp. 399–401, 2009.

[PHH23] F. Pauls, S. Haas, M. Hasler: "Trust-minimized Integration of Third-Party Intellectual Property Cores", 20th International SoC Design Conference (ISOCC), 2023

[PBR+22] D.-T. Phan-Huy, D. Barthel, P. Ratajczak, and R. Fara, "Ambient Backscatter Communications in Mobile Networks: Crowd-Detectable Zero-Energy-Devices," IEEE Journal of Radio Frequency Identification, June 2022.

[PCW+22] M. Perotti, M. Cavalcante, N. Wistoff, R. Andri, L. Cavigelli, and L. Benini, "A 'new Ara' for vector computing: an open source highly efficient RISC-V V 1.0 vector processor design," in IEEE 33rd Int. Conf. Appl.-Specif. Syst. Archit. Process. (ASAP), Gothenburg, Sweden, Jul. 2022, pp. 43–51.

[PCY+22] S. Park, S. Choi, S. Yoo, Y. Cho and J. Choi, "An Ultra-Low Jitter, Low-Power, 102-GHz PLL Using a Power-Gating Injection-Locked Frequency Multiplier-Based Phase Detector," in IEEE Journal of Solid-State Circuits, vol. 57, no. 9, pp. 2829-2840, Sept. 2022

[PJG+18] U. R. Pfeiffer, R. Jain, J. Grzyb, S. Malz, P. Hillger, and P. Rodriguez- Vazquez, "Current Status of Terahertz Integrated Circuits - From Components to Systems," IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS), pp. 1-7, 2018, doi: 10.1109/BCICTS.2018.8551068.

[PKM+15] J. Petäjäjärvi, H. Karvonen, K. Mikhaylov, A. Pärssinen, M. Hämäläinen, J. Iinatti, "WBAN energy efficiency and dependability improvement utilizing wake-up receiver;" IEICE Transactions on Communications, vol. 98, no. 4, pp. 535-542, 2015.

[PMR+22] S. Prasad, M. Meenakshi and P. H. Rao, "Hardware Impairments in mmWave Phased Arrays," 2022 IEEE Microwaves, Antennas, and Propagation Conference (MAPCON), Bangalore, India, 2022, pp. 891-896.

[PMV+16] J. Petäjäjärvi, K. Mikhaylov, R. Vuohtoniemi, H. Karvonen, and J. Iinatti, "On the human body communications: wake-up receiver design and channel characterization," EURASIP Journal on Wireless Communications and Networking, pp. 1-17, 2016.

[PW23] P. P. Puluckul and M. Weyn, "InfiniteEn: A Multi-Source Energy Harvesting System with Load Monitoring Module for Batteryless Internet of Things," in IEEE 9th World Forum on Internet of Things, Aveiro, Portugal, Oct. 2023.

[QLC+22] O. Quevedo-Teruel *et al*., "Geodesic Lens Antennas for 5G and Beyond," in *IEEE Communications Magazine*, vol. 60, no. 1, pp. 40-45, January 2022.

[Rat23] P. Ratajczak, "Design of a Dual Polarization Reconfigurable Intelligent Surface at 26.0 GHz for 5G Applications," 2023 17th European Conference on Antennas and Propagation (EuCAP), Florence, Italy, 2023, pp. 1-4, doi: 10.23919/EuCAP57121.2023.10133449.

[RBF+13] P. Ratajczak, P. Brachat, J. Fargeas and J. Baracco, "C-band active reflectarray based on high impedance surface," 2013 IEEE International Symposium on Phased Array Systems and Technology, 2013, pp. 570-576.

[RCB+21] K. Rasilainen, J. Chen, M. Berg and A. Pärssinen, "Dielectric Lens Antennas for 300-GHz Applications," *2021 15th European Conference on Antennas and Propagation (EuCAP)*, Dusseldorf, Germany, 2021.

[RMJ+20] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey of Machine Learning Accelerators," In Proceedings of the 2020 IEEE High Performance Extreme Computing Conference (HPEC), Greater Boston Area, MA, USA, 22–24 September 2020

[RMJ+21] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, “AI Accelerator Survey and Trends,” In Proceedings of the 2021 IEEE High Performance Extreme Computing Conference (HPEC), Virtual, 19–23 September 2022; pp. 1–9

[RIC12] Lyons, Richard. (2012). How Discrete Signal Interpolation Improves Digital-to-Analogue Conversion. Linear Audio Magazine.

[RIS23-D34] RISE-6G Deliverable 3.4: "Optimised RIS prototypes for PoCs and model assessment test"

[RIS21-23] RISING, Austrian funding project WAW 3041415: “RISING – Reflecitve Intelligent Surfaces for Reliable Wireless Communications in the 5G mmWave Band

[RK20] D. W. Rosolowski and P. Korpas, "IQ-imbalance and DC-offfset compensation in ultrawideband Zero-IF receiver," 2020 23rd International Microwave and Radar Conference (MIKON), Warsaw, Poland, 2020, pp. 209-213, doi: 10.23919/MIKON48703.2020.9253894.

[RKL+20] S. Rostami, P. Kela, K. Leppanen, and M. Valkama, “Wake-up radio-based 5G mobile access: Methods, benefits, and challenges,” IEEE Communications Magazine, vol. 58, no. 7, pp. 14-20, 2020.

[RLA+23] O. M. Rosabal, O. L. A. López, H. Alves and M. Latva-aho, "Sustainable RF Wireless Energy Transfer for Massive IoT: enablers and challenges," in IEEE Access, doi: 10.1109/ACCESS.2023.3337214.

[RPB+23] K. Rasilainen, T. D. Phan, M. Berg, A. Pärssinen and P. J. Soh, “Hardware Aspects of Sub-THz Antennas and Reconfigurable Intelligent Surfaces for 6G Communications," in *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2530-2546, Aug. 2023.

[RRM+23] D. E. Ruíz-Guirola, C. A. Rodríguez-López, S. Montejo-Sánchez, R. D. Souza, O. L. López, and H. Alves, “Energy-Efficient Wake-Up Signalling for Machine-Type Devices Based on Traffic-Aware Long Short-Term Memory Prediction,” IEEE Internet of Things Journal, vol. 9, no. 21, pp. 21620-21631, 2022.

[SAH+23] A. Sethi, R. Akbar, M. Hietanen, J. P. Aikio, O. Kursu, M. Jokinen, M. E. Leinonen, T. Rahkonen, and A. Pärssinen, "Chip-to-Chip Interfaces for Large-Scale Highly Configurable mmWave Phased Arrays," in IEEE Journal of Solid-State Circuits, vol. 58, no. 7, pp. 1987-2004, 2023, doi: 10.1109/JSSC.2023.3273502.

[SAS+23] A. Sabovic, M. Aernouts, D. Subotic, J. Fontaine, E. D. Poorter, and J. Famaey, "Towards energy-aware tinyML on battery-less IoT devices," in Internet of Things, vol. 22, 2023, doi: 10.1016/j.iot.2023.100736.

[SDF19] A. K. Sultania, C. Delgado, and J. Famaey, “Implementation of NB-IoT Power Saving Schemes in ns-3,” IoT, Jun. 2019.

[SJL02] I. H. Sohn, E. R. Jeong and Y. H. Lee, "Data-aided approach to I/Q mismatch and DC offset compensation in communication receivers," in IEEE Communications Letters, vol. 6, no. 12, pp. 547-549, Dec. 2002, doi: 10.1109/LCOMM.2002.806451.

[SLS+23] R. Valente da Silva, O. L. A. López and R. D. Souza, "Energy-Aware Federated Learning With Distributed User Sampling and Multichannel ALOHA," in IEEE Communications Letters, vol. 27, no. 10, pp. 2867-2871, Oct. 2023, doi: 10.1109/LCOMM.2023.3312793.

[SMB+12] M. Själander, S. A. McKee, P. Brauer, D. Engdal and A. Vajda, "An LTE Uplink Receiver PHY benchmark and subframe-based power management," 2012 IEEE International Symposium on Performance Analysis of Systems & Software, New Brunswick, NJ, USA, 2012, pp. 25-34

[SNM18] K. Sengupta, T. Nagatsuma, and D. M. Mittleman, ”Terahertz integrated electronic and hybrid electronic-photonic systems,” Nature Electronics, vol. 1, no. 12, pp. 622-635, 2018, doi: 10.1038/s41928-018-0173-2.

[SSJ+03] D.F. Sievenpiper, J.H. Schaffner, H. Jea Song, R.Y. Loo, and G. Tangonan, “Two-Dimensional Beam Steering Using an Electrically Tunable Impedance Surface”, IEEE trans. AP, pp 2713-2722, vol 51, n°10, October 2003.

[STP+23] M. Sarajlić, N. Tervo, A. Pärssinen, L. H. Nguyen, H. Halbauer, K. Roth, V. Kumar, T. Svensson, A. Nimr, S. Zeitz, M. Dörpinghaus, and G. Fettweis, Waveforms for sub-THz 6G: Design guidelines,” in 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 168–173, 2023.

[SW21] Z. Sha and Z. Wang, "Channel Estimation and Equalization for Terahertz Receiver With RF Impairments," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 6, pp. 1621-1635, June 2021, doi: 10.1109/JSAC.2021.3071824.

[SZB+99] D. F. Sievenpiper, L. Zhang, R. Broas, N. Alexopoulos, and E. Yablonovitch, “High-Impedance electromagnetic surfaces with a forbidden frequency band”, IEEE trans. MTT, pp 2059-2074 vol 47, n°11, November 1999.

[TAZ+13] A. Tzanakaki, M. P. Anastasopoulos, G. S. Zervas, B. R. Rofoee, R. Nejabati and D. Simeonidou, “Virtualization of heterogeneous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services”. IEEE Communications Magazine, vol. 51, no. 8, pp. 155-161, August 2013.

[TJD+19] C. Tarver, L. Jiang, A. Sefidi and J. R. Cavallaro, "Neural Network DPD via Backpropagation through a Neural Network Model of the PA," 2019 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2019, pp. 358-362, doi: 10.1109/IEEECONF44664.2019.9048910.

[TZP+20] C. -H. Tsai, Z. Zong, F. Pepe, G. Mangraviti, J. Craninckx and P. Wambacq, "Analysis of a 28-nm CMOS Fast-Lock Bang-Bang Digital PLL With 220-fs RMS Jitter for Millimeter-Wave Communication," in IEEE Journal of Solid-State Circuits, vol. 55, no. 7, pp. 1854-1863, July 2020

[VAK+22] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, Wide-Area, and Non-Terrestrial IoT: A Survey on 5G Advances and the Road Toward 6G," in IEEE Communications Surveys & Tutorials, vol. 24, no. 2, pp. 1117-1174, Secondquarter 2022, doi: 10.1109/COMST.2022.3151028.

[VHV16] H. Vangala, Y. Hong, and E. Viterbo, “Efficient algorithms for systematic polar encoding,” IEEE communications letters, vol 20, no. 1, pp. 17-20, 2015.

[Viz22] P. Vizcaino, F. Mantovani, R. Ferrer, J. Labarta, “Acceleration with long vector architectures: Implementation and evaluation of the FFT kernel on NEC SX-Aurora and RISC-V vector extension,” Concurrency and Computation: Practice and Experience, Vol. 35, Issue 20, Wiley online library, 2023, https://doi.org/10.1002/cpe.7424

[WAH+19] D. Wang, M. Aziz, M. Helaoui and F. M. Ghannouchi, "Augmented Real-Valued Time-Delay Neural Network for Compensation of Distortions and Impairments in Wireless Transmitters," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 1, pp. 242-254, Jan. 2019.

[WAW+18] J. Wang, A. Al-Khalidi, L. Wang, R. Morariu, A. Ofiare and E. Wasige, "15-Gb/s 50-cm Wireless Link Using a High-Power Compact III–V 84-GHz Transmitter," in IEEE Transactions on Microwave Theory and Techniques, vol. 66, no. 11, pp. 4698-4705, Nov. 2018.

[WBJ03] G. White, A. Burr, and T. Javornik, “Modelling of nonlinear distortion in broadband fixed wireless access systems,” Electronics Letters, vol. 39, pp. 686–687(1), April 2003

[WHM+21] H. Wang, F. Wang, H. T. Nguyen, S. Li, T. Y. Huang, A. S. Ahmed, M. E. D. Smith, N. S. Mannem and J. Lee, "Power Amplifiers Performance Survey 2000-Present," [Online]. Available: https://gems.ece.gatech.edu/PA_survey.html

[WSL+22] W. Wang, L. Sun, H. Liu and Y. Feng, "LSTM-CNN for Behavioral Modeling and Predistortion of 5G Power Amplifiers," 2022 IEEE 9th International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications (MAPE), Chengdu, China, 2022.

[WTJ+20] Z. Wen, Y. Tang, Y. Jia, H. Zhu, B. Chen, H. Yue and Z. Deng, “A Wideband Nonlinear Behavioral Model for D-Band Power Amplifiers Including Memory Effects," 2020 IEEE MTT-S International Wireless Symposium (IWS), Shanghai, China, 2020.

[WWH+23] H. Wang, F. Wang, H. T. Nguyen, S. Li, T. Y. Huang, A. S. Ahmed, M. E. D. Smith, N. S. Mannem and J. Lee, “Power Amplifiers Performance Survey 2000-Present,” [Online]. Available: https://gems.ece.gatech.edu/PA_survey.html

[YJH+16] X. Yu, S. Jia, H. Hu, M. Galili, T. Morioka, P. U. Jepsen, and L. K. Oxenløwe, ”160 Gbit/s photonics wireless transmission in the 300- 500 GHz band,” APL Photonics, vol. 1, no. 8, pp. 081301, 2016, doi: 10.1063/1.4960136.

[YSW16] S. Yan, W. Shi and J. Wen, "Review of neural network technique for modeling PA memory effect," 2016 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), Beijing, China, 2016.

[ZGL+17] J. Zhang, X. Ge, Q. Li, M. Guizani and Y. Zhang, "5G Millimeter-Wave Antenna Array: Design and Challenges," in IEEE Wireless Communications, vol. 24, no. 2, pp. 106-112, April 2017, doi: 10.1109/MWC.2016.1400374RP.

[ZHC+14] Z. Zhu, X. Huang, M. Caron and H. Leung, "Blind Self-Calibration Technique for I/Q Imbalances and DC-Offsets," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 61, no. 6, pp. 1849-1859, June 2014, doi: 10.1109/TCSI.2013.2290826.

[ZKG+20] J. Zhao, B. Korpan, A. Gonzalez, and K. Asanovic, “Sonicboom: The 3rd Generation Berkeley Out-of-Order Machine,” in Fourth Workshop on Computer Architecture Research with RISC-V (CARRV 2020), May 2020

[ZSK+20] S. Zeinolabedinzadeh, I. Song, M. Kaynak and J. D. Cressler, "A Wide Locking-Range, Low Phase-Noise and High Output Power D-Band SiGe PLL," 2020 IEEE 20th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF), San Antonio, TX, USA, 2020, pp. 35-38, doi: 10.1109/SIRF46766.2020.9040189.

[ZTY+22] S. Zhang, Y. Tian, P. Ye, L. Guo, H. Zeng and H. Li, "A Novel Calibration Method of DC-Offsets in Direct-Conversion Transmitter," in IEEE Communications Letters, vol. 26, no. 11, pp. 2745-2749, Nov. 2022, doi: 10.1109/LCOMM.2022.3192260.

[ZYZ+13] C. Zhang, S. Yan, Q. -J. Zhang and J. -G. Ma, "Behavioral modeling of power amplifier with long term memory effects using recurrent neural networks," 2013 IEEE International Wireless Symposium (IWS), Beijing, China, 2013, pp. 1-4, doi: 10.1109/IEEE-IWS.2013.6616831.