Horizon 2020 Program (2014-2020)
FET-Open – Novel ideas for radically new technologies
FETOPEN-01-2018-2019-2020

Architecting More than Moore – Wireless Plasticity for Massive Heterogeneous Computer Architectures [†]

D3.2: Wireless communication performance and cost models

WP3 - Wireless Communications within Package

Contractual Date of Delivery	30/09/2021
Actual Date of Delivery	01/10/2021
Deliverable Dissemination Level	Public
Editor	Sergi Abadal (UPC)
Contributors	UPC (leader), RWTH, UNIBO, EPFL
Quality Assurance	Davide Rossi (UNIBO) Peter Haring Bolívar (UoS)

[†]This project is supported by the European Commission under the Horizon 2020 Program with Grant agreement no: 863337.

Document Revisions & Quality Assurance

Deliverable Number	D3.1
Deliverable Responsible	UPC
Work Package	WP3
Main Editor	Sergi Abadal

Internal Reviewers

1. Davide Rossi (UNIBO)
2. Peter Haring Bolívar (UoS)

Revisions

Version	Date	By	Overview
1.4.0	30/09/2021	*Editor*	Synthesized final version.
1.3.0	29/09/2021	*Reviewers*	New version including comments from internal reviewers.
1.2.0	26/09/2021	*UPC Team*	Third draft with results of Chapter 5, 6, 7, plus appendices.
1.1.0	18/08/2021	*Editor*	Second draft with structure and text.
1.0.0	30/06/2021	*Editor*	First draft. Outline.

Legal Disclaimer

Executive Summary

The WiPLASH project aims to develop wireless-enabled architectures that deliver an improvement of 10$\times$ over existing architectures in multi-chip environments. In this context, accurate models of the wireless communication circuitry enabling the cost-benefit analysis of the wireless-enabled architectures is crucial. With these models, one can not only assess the impact of wireless on the proposed architectures, but also incorporate them into a larger framework with which design space explorations can be performed at the system architecture level –in pursuit of the 10$\times$ improvement. In this deliverable, we provide such models through a cross-cutting analysis at the wireless channel, physical layer, and link layer of design. The first provides the attenuation and dispersion suffered by signals as they travel through the package, based on results of Deliverable D3.1. With these figures, one can estimate the transmission rate and power consumption achievable in such an environment, as well as the area occupied by the associated transceiver circuits. For this, we develop our own bottom-up models combining literature analysis, link budget calculations, and modeling of specific circuits. Finally, at the link layer, one can model the impact of sharing a channel (or channels) among multiple wireless interfaces on the performance of the communication. In this respect, we propose a new link-layer protocol and perform a comparative evaluation of its performance through a comprehensive simulation campaign. Then, we model performance through data fitting of the simulations. In summary, with the analysis of this three aspects, we deliver models of the performance (latency and throughput) and cost (area, power) of wireless communications within computing packages, ready to be incorporated in the system-level simulators of WP5.

Abbreviations and Acronyms

NoC Network-on-Chip

WNoC Wireless Network-on-Chip

WNiP Wireless Network-in-Package

NiP Network-in-Package

MCM Multi-chip Module

mmWave millimeter-Wave

THz terahertz

FDTD Finite-Difference Time-Domain

PDP Power-Delay Profile

EM Electromagnetic

AlN Aluminum nitride

SiP System-in-Package

I/O Input/Output

BER Bit Error Rate

SNR Signal-to-Noise Ratio

SoC System-on-Chip

MAC Medium Access Protocol

TDP Thermal Design Power

CSMA Carrier-Sensing Multiple Access

PCB Printed Circuit Board

PHY Physical Layer

PLL Phase-Locked Loop

OOK On-Off Keying

PSK Phase-Shift Keying

QAM Quadrature Amplitude Modulation

PAM Pulse Amplitude Modulation

ACK ACKnowledgment

GRS Ground-Referenced Signaling

SPP Surface Plasmon Polariton

EMIB Embedded Multi-die Interconnect Bridge

LoS Line of Sight

SAR Successive Approximation Register

SDCT Sigma-Delta Converter

PAE Power-Added Efficiency

ADC Analog-Digital Converter

DAC Digital-Analog Converter

WPAN Wireless Personal Area Network

WSN Wireless Sensor Network

RZ Return-to-Zero

ENOB Effective Number of Bits

The WiPLASH consortium is composed by

UPC	Coordinator	Spain
IBM	Beneficiary	Switzerland
UNIBO	Beneficiary	Italy
EPFL	Beneficiary	Switzerland
AMO	Beneficiary	Germany
UoS	Beneficiary	Germany
RWTH	Beneficiary	Germany

Contents

List of Figures

List of Tables

1. Introduction

Efficient on-chip communication to enable the exchange of data between the processing elements of a multicore processor or System-on-Chip (SoC) is a prerequisite for high performance. Nowadays, most of these computing systems incorporate a Network-on-Chip (NoC) composed of a fabric of integrated routers and intra-chip wired links [9]. However, recent trends in computer architecture are leading to extreme scaling (using many processor cores), specialization (putting together multiple hardware accelerators to boost performance), and disintegration (interconnecting multiple smaller dies within a System-in-Package (SiP) instead of within a very large monolithic SoC). This places unprecedented bandwidth and reconfigurability requirements to the interconnect fabric which now has to also extend beyond the limits of a single chip [10, 11]. New paradigms are thus required in the manycore era, which is the hypothesis over which the WiPLASH project unfolds.

Among the different emerging alternatives, wireless chip-scale communications stand as a promising contender as advocated by WiPLASH [12, 13]. In this type of communication, on-chip transceivers modulate the information, which is radiated by on-chip antennas in the form of Electromagnetic (EM) waves that propagates through the chip package, possibly beyond the boundaries of the chip, until reaching on-chip antennas placed across the SiP. These wireless links provide unique system-wide low latency, inherent broadcast capabilities, and possibilities for network reconfigurability that wired alternatives cannot offer due to the need of a path infrastructure and possibly many hops (and chip/clock domain crossings) to reach distant locations [12–14]. Hence, the concepts of Wireless Network-on-Chip (WNoC) and Wireless Network-in-Package (WNiP) are conceived as the combination of wired and wireless intra-/inter-chip links within a computing package.

To illustrate the WNoC paradigm, Figure 1.1 represents a possible scenario with wireless links within an heterogeneous architecture, together a possible protocol stack that defines the communication process. Information coming from the processor or the memory modules are first routed to the wired or wireless network (network layer); once entering the wireless network, data is momentarily stored in an intermediate buffer while the Medium Access Protocol (MAC) protocol determines which channel is used to transmit the information and when (link layer). Once data is bound to be transmitted, data is serialized and modulated by the transceiver before being radiated (physical layer). Radiation is typically assumed to be omnidirectional and within a single fixed frequency band, although WiPLASH also proposes to use miniaturized tunable antenna arrays to produce field concentrations in certain areas of the chip in different frequency channels. Fields at the receiving antennas are picked up and decoded, deserialized, and passed to the corresponding core or memory module.

As demonstrated in the literature and within WiPLASH, the unique features of WNoC and WNiP networks can become key enablers of radically new architectures

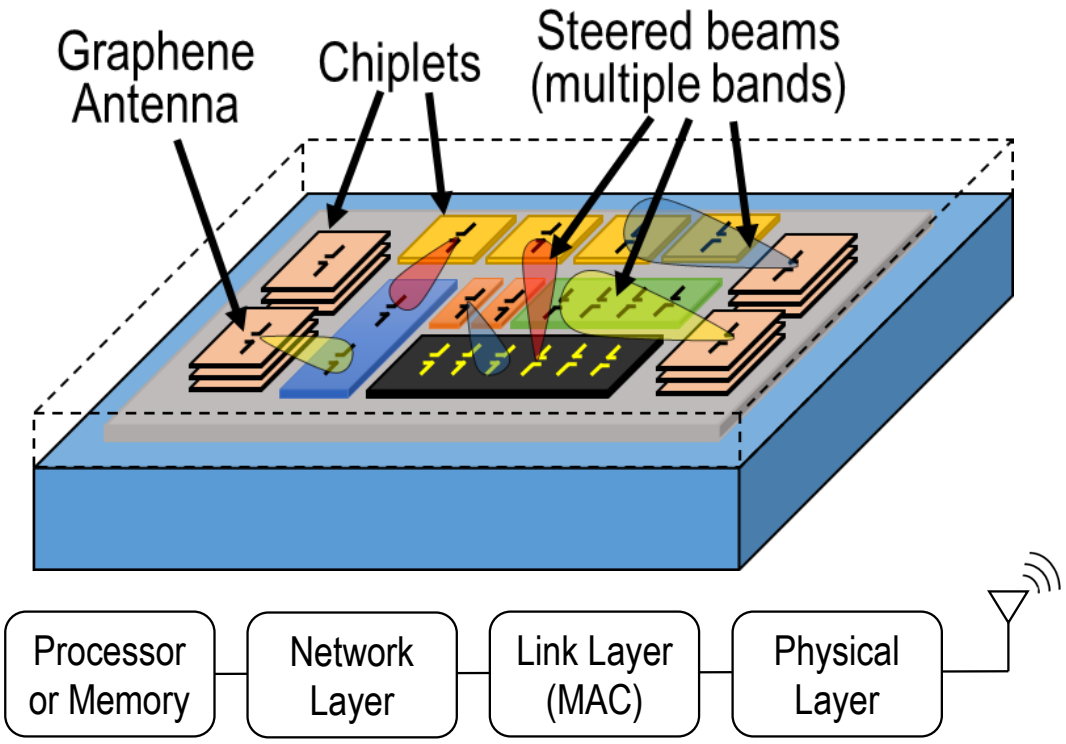


Figure 1.1: A general view of the WiPLASH vision on wireless communications at the chip scale within a heterogeneous computer architecture and multiple frequency-tunable, beam-steerable antennas. At the bottom, we show the logical structure of the wireless network with its network, link, and physical layer protocols.

capable of pushing the scalability limits of nowadays SoCs/SiPs [15–17]. However, by its nature, wireless communications also bring certain disadvantages. Indeed, the low latency, broadcast and reconfiguration capabilities generally come at the cost of (i) moderate energy efficiency stemming from the fact that bits need to be serialized, modulated, and amplified to compensate for the relatively high attenuation of the wireless channel; non-negligible area overhead produced by the analog and digital components required to implement the links, and low aggregate bandwidth resulting from the need to share a few channels among all antennas.

Aware of the pros and cons of wireless chip-scale communication, a challenge resides in assessing the actual impact of these networks on future manycore architectures. The main reason is that wireless transceivers in the literature have not been developed specifically for this scenario, thus reducing their suitability for this scenarios and introducing distorted performance and efficiency metrics in the architectural simulation. Another possible reason is that proposals for architectures that may fully benefit from the characteristics of wireless communications are scarce [5, 18–20] and have used generic models for a preliminary assessment of their performance; as a result, the demand for such models is relatively low. We found that these reasons are entangled and often reinforce each other.

In WiPLASH, for which developing wireless-enabled architectures that offer an improvement of 10$\times$ over existing architectures is among the main objectives, accurate models of the wireless network enabling the cost-benefit analysis of the wireless-enabled architectures is crucial. With these models, one can not only accurately as-

sess the impact of wireless on the proposed architectures, but also incorporate them into a larger framework with which design space explorations can be performed at the system architecture level – in pursuit of the 10× improvement.

Providing accurate models of the performance and resource consumption (energy, area) of wireless communication links at the chip scale is a challenge in itself for a number of reasons, which forces the WiPLASH project to go beyond the state of the art in the following ways:

- **Wireless channel:** performance and power models of wireless communications are generally grounded on knowledge of the channel and an adequate link budget formulation. However, most of the characteristics of the wireless channel within a system package have remained largely unknown until a few years back. Therefore, accurate modeling of power and performance require, in turn, accurate models of the channel as provided in WiPLASH. In this deliverable, we provide a summary of channel studies from Deliverable D3.1 that can guide the modeling at upper layers.
- **Physical layer:** models of the physical layer are needed to calculate the area and power consumption of a wireless links. The issue is that these models have never been done before at the level of accuracy and flexibility that is demanded in WiPLASH. There are surveys of specific components (by others, [2,3]) or even of complete transceivers (by the authors of this deliverable, [6,21]). However, these have been generally made in a top-down approach, including a heterogeneous compilation of technologies and applications, which are often not compatible with the design drivers and constraints of the chip scenario. Here, we propose to go one step beyond and build the models bottom up as a complement to the top-down strategy, based on a single transceiver architecture and a set of constraints and requirements which can be considered plausible for the chip-scale communications scenario.
- **Link layer:** there is a plurality of MAC protocols leading to different latency-throughput curves that could serve as performance models in WiPLASH. However, these protocols are usually single-channel or assign the multiple channels statically or quasi-statically [19, 22–27]. Moreover, evaluations seldom consider spatiotemporally imbalanced traffic and therefore are insufficient. Since WiPLASH considers the existence of multiple quickly reconfigurable channels, be it in space, time, or frequency, we would need to provide performance models for dynamic and multi-channel MAC protocols. Ideally, in light of the typically imbalanced traffic in multiprocessor workloads [28, 29], models should include hotspot and bursty traffic besides the typical Poisson and uniformly distributed models. In this deliverable, we provide the tools to address both challenges.

In summary, the main focus of this deliverable is *to build performance and resource consumption models for wireless chip-scale communications relevant to the targets and specifications set in the project*. In this direction, and aiming to place the deliverable within the bigger context of the project, the main contributions reported in this document are:

- An outline the formulation of wireless chip-scale communication models within the bounds of a design space exploration framework, to be interfaced with the entire system evaluation framework.

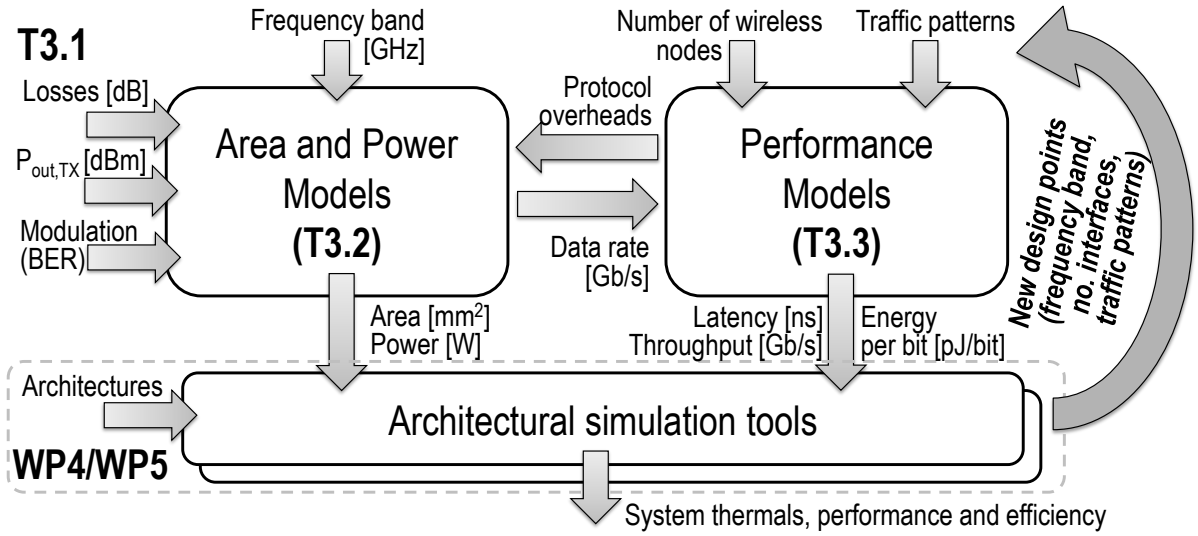


Figure 1.2: Graphical abstract of this deliverable (D3.2). Channel modeling is performed over simulations performed in T3.1 [1], which provides propagation losses. The losses, given a modulation and BER requirement, together with the frequency band of choice, determine the area, power, and data rate of the associated transceiver as modeled in task T3.2. These figures, together with the MAC protocols designed and simulated in T3.3 for a number of wireless nodes and given a certain traffic pattern, determine the latency, throughput, and final energy per bit of a link. All these results are inputs for WP4 in architecture design and WP5 in simulation. The loop is closed through system-level design space exploration, which will propose new requirements (traffic patterns, bandwidth, number of wireless nodes) to be assessed.

- A survey the state of the art in wired multi-chip interconnects, i.e. Multi-chip Module (MCM), interposers, and bridges. The survey also studies their underlying physical layers in terms of performance and cost, based on the analysis of a set of recent proposals. This provides baselines for comparison with the wireless physical layer.
- Performance and resource consumption (area, power) models for the wireless communication alternative, relating three different layers:
  - **Wireless channel:** taking EM simulations of different packages and obtaining attenuation and delay spread models;
  - **Physical layer:** creating area and power models in a bottom-up approach, based off a given transceiver architecture parameterized to multiple context variables and design decisions; and
  - **Link layer:** presenting a new family of dynamic MAC protocols extensible to multi-channel scheduling approaches, and evaluating the performance in terms of latency and throughput for a variety of traffic patterns representing the variety of workloads that multi-chip architectures may face.

Within the project, and as summarized in Figure 1.2, the specification of the wireless architecture from Task T1.1 has provided fundamental inputs in terms of target frequency ranges and package design. These are valuable insights to understand what a plausible transceiver architecture would be, as well as the variable ranges to consider in the models. Moreover, these specification ranges have guided Task T3.1 on channel

models (Deliverable D3.1 [1]), whose outputs are extremely relevant here. The modeling of area/power and performance are enabled by work in two tasks, namely Task T3.2 on physical layer design and T3.3 on link layer design, respectively. The outputs of the study are relevant to the WP4 on architecture design and WP5 on multi-scale simulation framework. In particular, these models can be coupled with design space exploration frameworks where a number of wireless interfaces are integrated with the architecture. The models allow for a fast exploration of the performance, area, and power of the entire system including the wireless network.

The remainder of this document is organized as follows. We first provide a description of the modeling methodology from the channel up to the physical and link layers of design in Chapter 2. The modeling results are presented in the three subsequent chapters: brief summary of channel models in Chapter 3, physical layer modeling in Chapter 4 and link layer modeling in Chapter 5. We finally summarize the main findings and discuss possible future lines of research in Chapter 6. Beyond that, the deliverable includes a set of appendices with background information about protocol design at the chip scale (Appendix Chapter A), a survey of the state of the art in multi-chip interconnects, including wireless (Appendix B), and more detailed results for channel models and link layer models (Appendix C and D.

2. Methodology

This chapter summarizes the methods employed in subsequent chapters to model the different aspects impacting the performance and resource consumption of wireless links within a computing package. Figure 2.1 shows a graphical schematic of the methodology, which details inputs and outputs, how the different parts interact with each other, as well as the software used in each step. To see how this methodology fits within the context of a possible system-level design space exploration framework, we refer the reader back to Figure 1.2.

In essence, we model the wireless channel using simulations from Deliverable D3.1 [1] together with some new simulations, which describe the propagation of wireless signals within computing packages. These serve to obtain attenuation and dispersion of signals as a function of distance, which can later be modeled with a few modeling parameters that can be extracted from the data. More details on the methods to obtain these parameters are given in Section 2.1.

The channel model parameters can be then fed to physical layer models for which attenuation and dispersion are critical. We obtain these models through a bottom-up approach that fixes the transceiver architecture and uses aggregate data from individual components published in the literature, together with theoretical design trends. The models relate performance in data rate and error rate with area overhead and power consumption. More details on the methodology are given in Section 2.2.

Assuming a given data rate and error rate, a link layer analysis delivers the performance of a wireless link in terms of latency and throughput. To that end, event-driven simulations are conducted which consider different types of traffic, different number of antennas sharing a link, and different number of available channels. More details on the simulation methods are depicted in Section 2.3.

2.1 Channel Modeling

Here, we summarize the methodology used in Deliverable D3.1 [1] to characterize the channel within the computing package, and then detail which procedure has been followed to model it.

2.1.1 Simulation

We first provide 3D models that capture the geometry and materials of the different packages under study. In particular, we model the flip-chip, interposer, and wirebond packages depicted earlier in Section A.3 using datasheets and schematics from real packages. The detailed composition of these packages is given in Chapter 3.

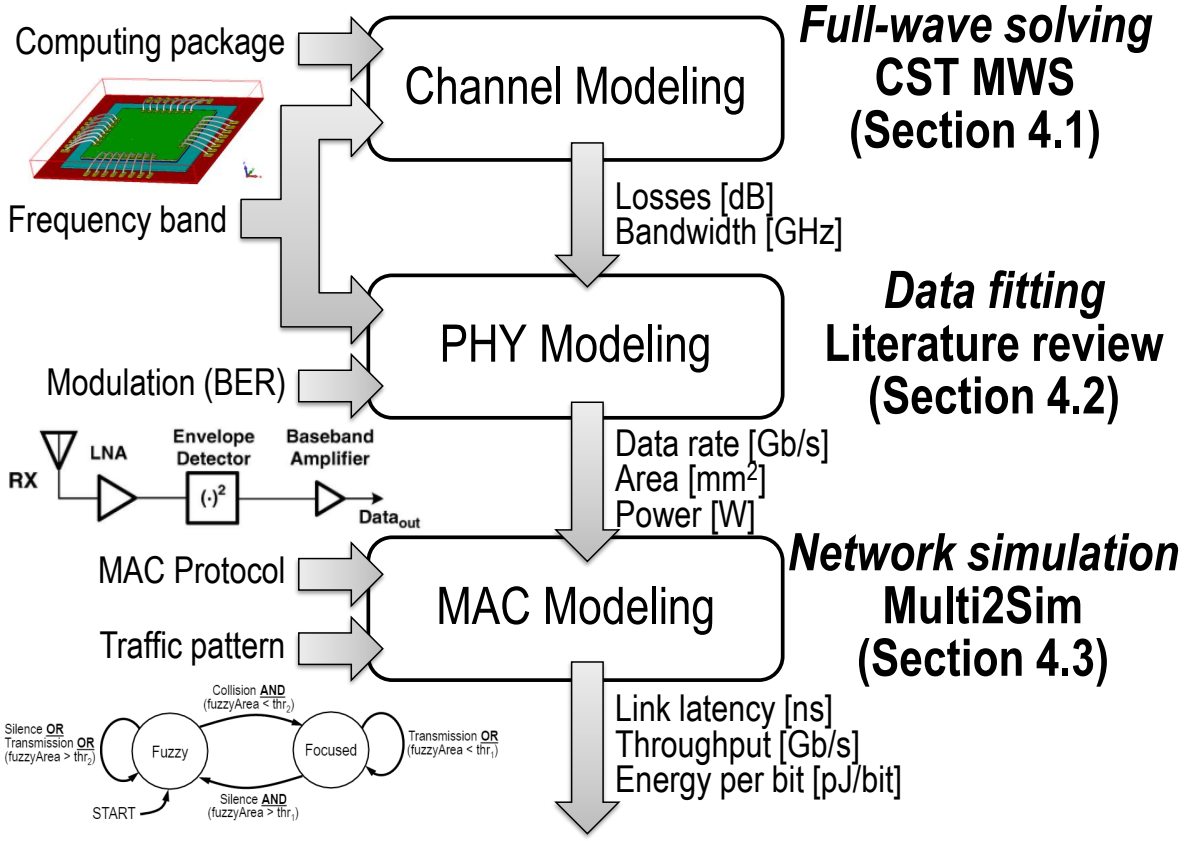


Figure 2.1: General view of the evaluation methodology used in this deliverable for the modeling of the wireless channel, physical layer, and link layer of design.

The different packages are simulated in a particular frequency band or using broadband pulses in the time domain by means of a full-wave solver. In our case, we employ CST Microwave Studio [30], which hosts a variety of methods for frequency and time domain analysis. We considered a homogeneous distribution of 4×4 antennas within the die(s) of the package. To minimize the impact of the antenna on the channel characterization procedure, we employ electrically small antennas as transmitters and receivers. In our case, simulations have been performed in two workstations, namely, a quad-core CPU at 3.90 GHz with 32 GB of RAM and a GeForce GTX 1080Ti GPU to accelerate time-domain simulations, and a 16-core CPU at 2.16 GHz with 128 GB of RAM. Several approximations have been performed to reduce the computational burden as described in [31, 32].

The outcome of the simulations are a set of S-parameters or time signals relating the output at the receiving antenna as a function of the input at the transmitting one. These parameters are then fed to custom MATLAB scripts that obtain the path loss characteristics out of the S-parameters, as elaborated in Section 2.1.1.1, and the delay spread scaling out of the time signals, as described in Section 2.1.1.2.

2.1.1.1 Frequency Domain Analysis

Once the S parameters are obtained, the channel frequency response $H_{ij}(f)$ is evaluated for each antenna pair as

$$G_i G_j |H_{ij}(f)|^2 = \frac{|S_{ji}(f)|^2}{(1-|S_{ii}(f)|^2)\cdot(1-|S_{jj}(f)|^2)}, \tag{2.1}$$

where G_i and G_j are the transmitter and receiver antenna gains, S_{ji} is the coupling between transmitter i and receiver j, whereas S_{ii} and S_{jj} are the reflection coefficients at both ends as obtained from the simulations. With respect to the antenna gain, we evaluate the gain over the complete solid angle as signals may be picked up from any direction.

2.1.1.2 Time Domain Analysis

In the time domain, we define an input excitation $x_i(t)$ at the input of the transmitting antenna i. CST employs the Finite-Difference Time-Domain (FDTD) method to calculate the output signal $y_j(t)$ at the receiving antenna j. Hence, the impulse response $h_{ij}(t)$ between transmitter i and receiver j can be derived with the classical formulation

$$y_j(t) = x_i(t) \star h_{ij}(t), \tag{2.2}$$

where $\star$ denotes the convolution operator. Our simulations consider a Gaussian cycle whose bandwidth spans the whole spectrum from 10 GHz to 1 THz, leading to a very short impulse at the input so that $x(t) \to \delta(t)$. Therefore, the output signal approximates the impulse response $y(t) \to h(t)$. Once $h(t)$ is calculated, we can evaluate the Power-Delay Profile (PDP) that describes the intensity of a signal received through a multipath channel as a function of time delay τ, as

$$P_{ij}(\tau) = |h_{ij}(t,\tau)|^2, \tag{2.3}$$

between transmitter i and receiver j.

2.1.2 Modeling

The evaluation of the response of the channel among all possible antenna pairs across the computing package allows to obtain the matrix of responses in the frequency and time domains, denoted as $\mathbf{H} = (H_{ij})$ and $\mathbf{P} = (P_{ij})$, respectively.

2.1.2.1 Attenuation

With the matrix of frequency responses $\mathbf{H}$ and their respective distances, a path loss analysis can be performed by fitting the average attenuation L over distance d over Equation (A.2), repeated here:

$$PL = PL_0 + 10\gamma \log_{10} \frac{d}{d_0} + X_g. \tag{2.4}$$

In this case, a linear fitting over the logarithmic plot provides the values of PL_0, d_0, and γ that minimize the average squared error between the fitted line and the measurement points. In our case, $d_0 = 2$ mm unless otherwise stated. The error X_g can be

understood as the effects of multipath propagation due to, for instance, some waves simply traversing the chip and other waves coupling into the package and back to the chip and can be modeled into X_g, which traditionally models fading or other effects. As for the other parameters, $\gamma < 2$ generally implies waveguided propagation in layers of such enclosed structures, whereas $\gamma > 2$ implies that propagation in the lossy silicon dominates. We also report the worst-case attenuation PL_{max} across the entire package, as it may not coincide with the attenuation between the two most distant antennas.

2.1.2.2 Dispersion

With the matrix of PDP responses $\mathbf{P}$, one can evaluate the multipath richness of the channel with the delay spread τ_{rms} as

$$\tau_{rms}^{(i,j)} = \sqrt{\frac{\int (\tau - \overline{\tau_{ij}})^2 P_{ij}(\tau)\, d\tau}{\int P_{ij}(\tau)\, d\tau}}, \tag{2.5}$$

where $\overline{\tau_{ij}}$ is the mean delay of the channel, which is calculated as the first moment of the PDP. In other words,

$$\overline{\tau_{ij}} = \frac{\int \tau P_{ij}(\tau) d\tau}{\int P_{ij}(\tau)\, d\tau}. \tag{2.6}$$

As we will see in Chapter 3, the delay spread increases with distance in a trend that can be fitted in a linear function. Hence, we will model the delay spread with fitted parameters τ_{rms}(2mm) which represents the value at 2mm, and γ_t as the slope of the linear function in [ns/mm].

Finally, we will assume that the transmission rate of all nodes are dimensioned to the worst case across all links and, therefore, they should be operated at the lowest speed ensuring correct decoding at all nodes. As a result, we will take the worst delay spread across all pairs of transmitters-receivers (i.e., across all distances) as limiting case and use it to evaluate the coherence bandwidth B_c, as follows

$$\tau_{rms} = \max_{i,j \neq i} \tau_{rms}^{(i,j)} \Rightarrow B_c \propto \frac{1}{\tau_{rms}}. \tag{2.7}$$

For simplicity, we will take $B_c = \frac{1}{\tau_{rms}}$.

2.2 Physical Layer Modeling

The methodology that we follow to model the physical layer of design uses a bottom-up approach, unlike our prior efforts. The problem of the top-down approach used in [5,21] and depicted in Section B.3 is that transceivers are difficult to compare because of the large variance in terms of specifications, technologies, and communication scenarios. Moreover, those models often do not account for the Digital-Analog Converter (DAC) and SerDes components crucial to implement the communication. To address these issues, in our bottom-up strategy we consider a fixed transceiver architecture. Then, with theoretical models and literature reviews for specific components, we extract area and power tendencies applicable to the WNoC scenario.

We note that the proposed methodology could be coupled to the wireless channel models described herein to guide the specification of certain components. The ultimate goal is to provide models that relate resource consumption, i.e. area and power, with performance, i.e. output transmission rate and error rate. Next, we first summarize the aspects analyzed in the literature review in Section 2.2.1 and then further detail the modeling approach in Section 2.2.2.

2.2.1 Literature review

In this section, we describe the methodology used in a few works, including ours, to obtain performance and efficiency trends from published works on RF design. Driven by the lack of a single approach to analog or mixed signal design, with a variety of technologies, applications with wildly different design drivers, and circuit architectures, several attempts have been made at capturing the evolution of certain analog components present in all transceivers. These are analyzed in Sections 2.2.1.1 and 2.2.1.2 for the cases of data converters and power amplifiers, respectively.

2.2.1.1 Data Converters

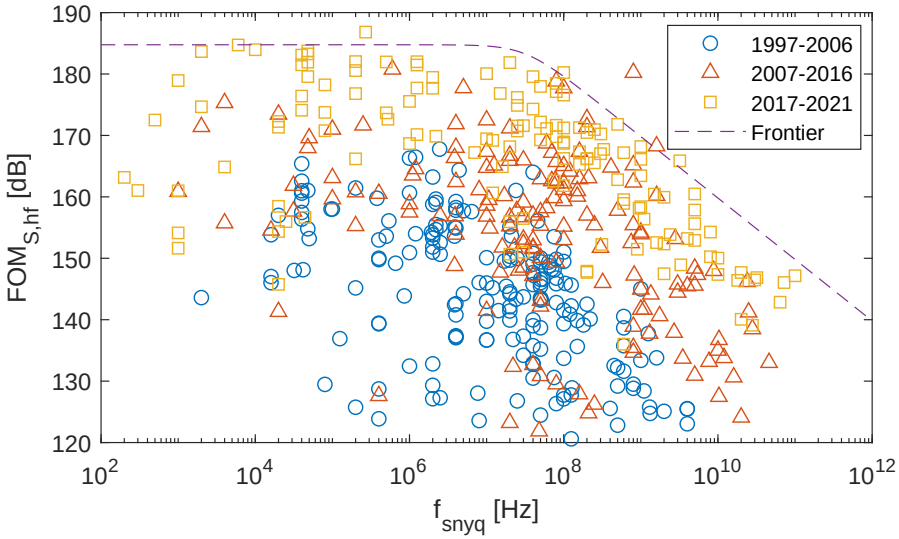
One of the most relevant examples of component-centric survey is that of *analog-digital converters* performed by Murmann at Stanford [2]. The survey covers over 600 designs from 1997 that have been appearing in the *International Solid-State Circuits Conference* and the *VLSI Symposium*, including all different architectures such as Successive Approximation Register (SAR) or Sigma-Delta Converter (SDCT). The assessment includes performance metrics such as the sampling frequency at Nyquist rate f_{snyq} and the Effective Number of Bits (ENOB), resource consumption metrics such as the area A_{DAC} or the sampling density $\delta_S = \frac{A_{DAC}}{f_s}$, and energy consumption metrics such as the Walden or Schreier figures of merit, expressed as

$$\begin{aligned} FOM_W &= \frac{P}{2^{ENOB} \cdot f_s} \\ FOM_S &= SNDR + 10\log(\tfrac{f_s/2}{P}) \end{aligned} \tag{2.8}$$

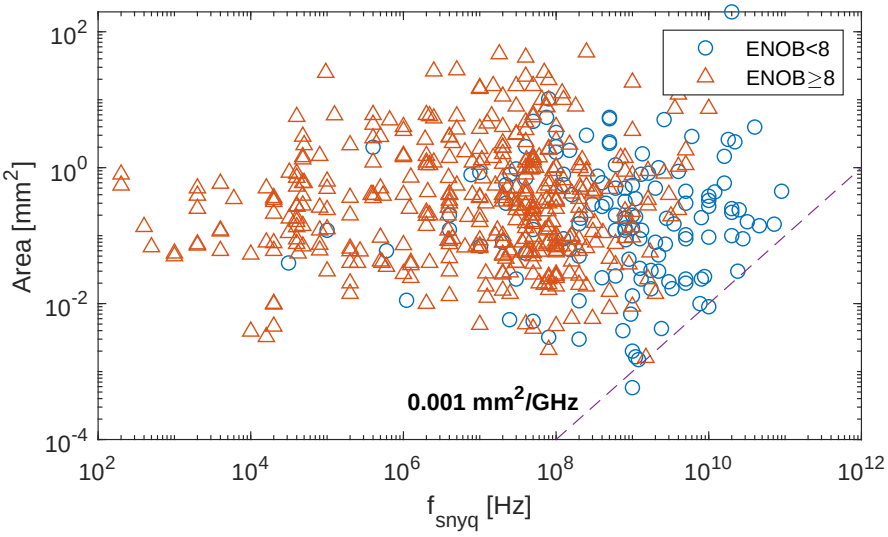
where P is the power consumption, f_s is the sampling rate, and $SNDR = 6.02 \cdot ENOB + 1.76$ is the Signal-to-Noise/Distorsion-Ratio.

Such an analysis has allowed to derive trends and performance-efficiency frontiers when pushing the limits of modulation frequency or technology. This is useful for the case of WNoC since, in the pursue of simplicity and low power with low-order modulations, achieving high speeds entails increasing the modulation rate to very high levels. This means that, even for such simple modulations, Analog-Digital Converters (ADCs) shall not be taken for granted as they may consume a significant amount of area and power to achieve very high sampling rates.

Figure 2.2 shows two examples of the assessments that can be made with Murmann's survey. On the top plot, we observe the Schreier's figure of merit, which relates the performance in both resolution $ENOB$ and sampling frequency f_s, with the power consumption. The envelope shown in the figure provides an upper bound on efficiency, and we see maximum efficiency is expected up to 100 MHz, after which the efficiency degrades significantly. On the bottom plot, the area occupied by the ADC is expressed as a function of the sampling frequency and distinguishing between low-resolution and



(a) Energy



(b) Area

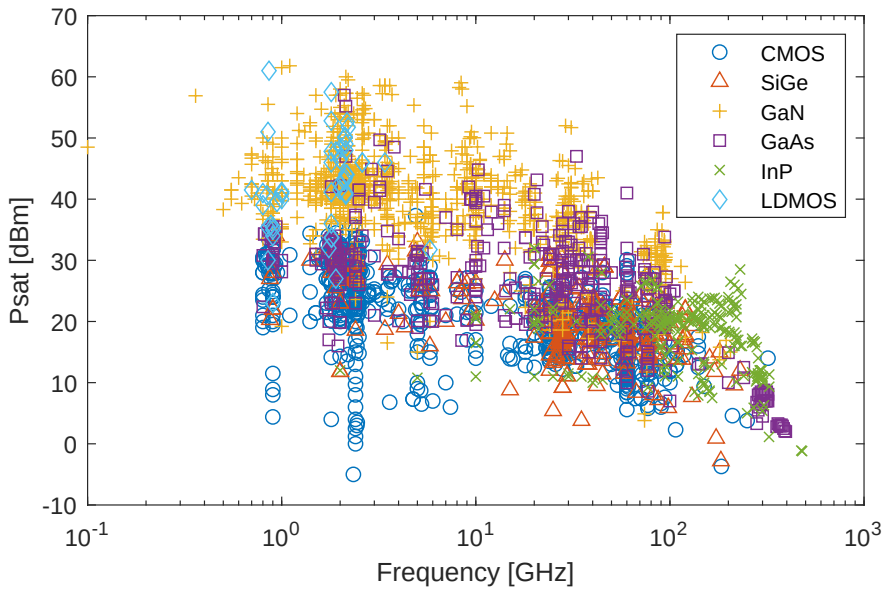
Figure 2.2: Examples of performance-cost curves for ADCs with data from [2].

high-resolution converters. We see that higher sampling rate correlates with lower resolution, as expected, but not necessarily with higher area. Two reasons may be the smaller size of passives at higher frequencies and of digital circuits at newer technologies. We also identify a frontier of 0.001 mm^2/GHz in the most area efficient designs.

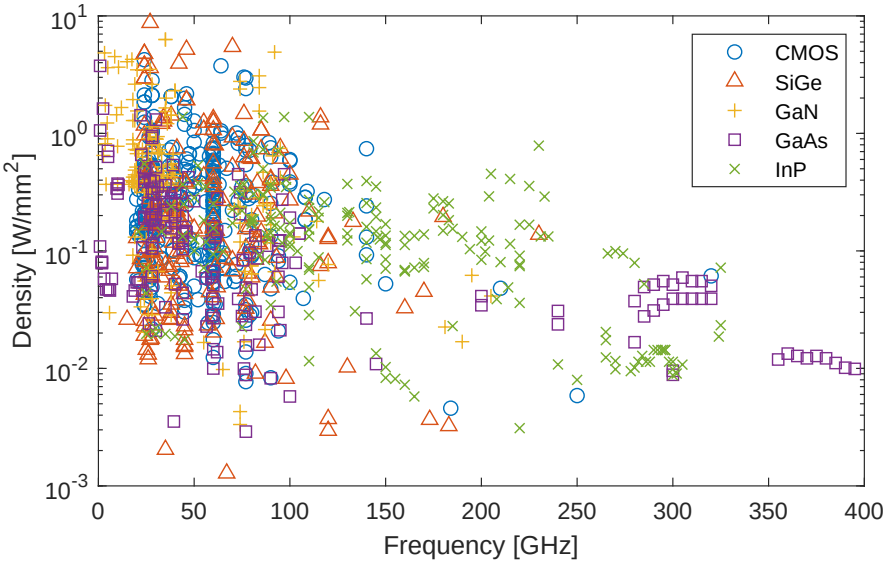
2.2.1.2 Power Amplifiers

Another survey relevant to this deliverable is that of power amplifiers led by Wang at Georgia Tech [3]. At the time of this writing, the survey had more than 3800 points covering a wide range of technologies such as CMOS, SiGe, GAN, or InP, recently including oscillators and multipliers at terahertz (THz) frequencies. Analyzed metrics include frequency, saturation power P_{sat}, gain G, or the Power-Added Efficiency (PAE)

$$PAE = \frac{P_{out} - P_{in}}{P_{DC}} \tag{2.9}$$



(a) Energy



(b) Area

Figure 2.3: Efficiency trends for power amplifiers with data from [3].

where P_{out} and P_{in} are the RF power at the output and input of the amplifier, whereas P_{DC} is the power consumed to provide the amplification. In the case of WNoC, due to the non-negligible path loss, signals may need to be amplified significantly and, at high frequencies, the power amplifier can become the most consuming element of the entire transceiver.

Figure 2.3 provides two examples of trends that can be extracted from the survey's data. On the top plot, we can clearly see how different technologies occupy a different region of the amplification spectrum: CMOS generally saturate earlier and for lower frequencies, while GaN is ideally suited thanks to its support of high power at moderated frequencies, whereas GaAs has a good performance at 60–100 GHz while InP is dominant higher frequencies. On the bottom plot, the area density defined as $\delta_P = \frac{P_{sat}}{A_{PA}}$ where A_{PA} is the area of the amplifier. We observe that the density tends to decrease

at higher frequencies, meaning that since the amplifiers saturate earlier, it takes more area to provide similar amplification. As such, the best designs of any technology tend to see their density reduced as they approach f_T and f_{MAX} of the technology.

2.2.2 Modeling

2.2.2.1 Area Modeling

In wireless chip-scale communication, physical links are not needed in order to convey the information from the transmitter to the receiver. Therefore, the only components that occupy chip area are the antennas and the transceivers needed to modulate the data and to drive the signals to the antenna. One can also include the area occupied by the data converters, i.e. ADC and DAC, the area required serialization/deserialization circuits. Therefore the area of a wireless interface A_{wi} becomes

$$A_{wi} = A_{serdes} + A_{conv} + A_{trx} + A_{ant} \tag{2.10}$$

which includes, from digital to analog, the serialization circuits, the data converters, the transceiver, and the antenna.

The area will be calculated as a function of multiple design parameters that are of interest in the design space exploration of complete computing systems. In particular, we are interested in seeing how the area scales with the data rate R and the channel losses PL. These, in turn, imply an indirect scaling in terms of (a) operation frequency f, because higher data rates may require shifting to higher frequency bands pursuing larger bandwidths; and (b) the transmission range d_{max} as it determines the worst-case losses. Each of these input parameters affects each component of the transceiver chain differently. Next, we describe how the components are modeled.

Serialization and deserialization: we will take base on numbers from specific designs that cover the speeds required in our analysis [33–35]. For instance, data from [34] indicates an area of 0.04 mm^2 to achieve a line rate of 14 Gb/s at 65nm CMOS. Being digital logic, technology scaling reduces the size of the circuitry proportionally to the feature size. Hence, scaling down to 14nm CMOS would lead to an area on the order of $\sim$0.01 mm^2, this is, around 0.001 mm^2/Gb/s of binary data. One can assume that area scales linearly with data rate due to the need to employ larger multiplexers and demultiplexers (or more stages). Therefore, the area per serialized symbol remains constant.

Data converters: we will use data from Murmann's survey, described in previous subsections, to model the area occupancy through a few steps. We will assume two extreme cases. On the one hand, we model the best-case scenario of the interface not requiring a DAC is not needed because the transceiver uses direct modulation [36]. On the other hand, we model the worst-case scenario of having a DAC of a similar complexity than the ADC and thus consuming similar area. In either case, the area will be obtained as a function of the assumed $ENOB$ and f_s requirement. Namely, we assume a constant area of 0.01 mm^2 for frequencies below 10 GHz, and a density of 0.001 mm^2/GHz otherwise as long as the $ENOB$ is lower than 8. For higher $ENOB$ values, the assumed density at high frequencies worsens by at least an order of magnitude. These are calculated based on the envelop of Figure 2.2.

Transceiver: for such a central part of the model, we will be using a bottom-up approach. In particular, we will aggregate the area of individual components required

to support a certain modulation. Each of the components will have its own model, namely:

- **Power amplifier:** the area is evaluated as a function of the central frequency of the transceiver, which determines the maximum achievable amplification density and, as such, the minimum area occupancy. From Figure 2.3, one can obtain a Pareto frontier that passes through all technologies and describes the best density than one can expect in power amplifiers. This is an optimistic but realizable approximation to the area of a power amplifier. The area is obtained with this figure and the expected required gain.
- **Low-noise amplifier:** to simplify the analysis, we assume that a low-noise amplifier takes a similar area than the power amplifier.
- **PLL:** although a survey exists that covers a wide variety of PLLs [37], the reported area results do not indicate the frequency of the PLL. Then, rather than obtaining a model, we use active area data from a specific design at 60 GHz with 28-nm FD-SOI technology [38] and scale it to higher bands by using area figures for frequency doublers and triplers from [7].
- **Mixers:** we use data from specific transceivers at 60 GHz [7] and at 240 GHz, the latter being a tapeout by RWTH Aachen for WiPLASH, as explained in [39], and then interpolate values at intermediate frequencies.
- **Filters:** filters use analog components that may lead to non-negligible area occupancy. We use a fixed amounts from works in the mmWave band [40, 41] and assume that similar areas can be occupied at higher frequencies by the size reduction of the passive elements but higher complexity of the circuit.

Antenna: we will make the assumption that a resonant patch antenna or a broadband antenna design of similar area, e.g. bow tie or fractal [42], is able to provide the bandwidth required by the transceiver. In such case, both the width and length are comparable to half a wavelength $\lambda/2$, leading to $A_{ant} \approx \frac{\lambda^2}{4}$, where $\lambda = \frac{c_0}{\varepsilon_{eff} f}$ for a metallic antenna resonating at a frequency f within a medium of effective permittivity ε_{eff}. For a graphene antenna, the dimensions are commensurate to the Surface Plasmon Polariton (SPP) wavelength $\lambda_{SPP} = \frac{\lambda}{K}$ where K is the compression factor of the SPP wave in the antenna.

2.2.2.2 Energy Modeling

Unlike in traditional wireless networks, the network nodes in a WNoC are integrated within the same platform and share a common power supply. At the physical layer, we will assume a single transmitter and a single receiver within range. In this context, the power consumption during a transmission is given by

$$P_{wi} = P_{tx} + P_{rx} = (P_{ser} + P_{dac} + P_{mod}) + (P_{rx} = P_{des} + P_{adc} + P_{demod}), \tag{2.11}$$

where P_{tx} and P_{rx} are the power consumed at the transmitter and receiver side, respectively, which include the serializer, DAC and the modulator (including any power amplifiers) on the one hand, and the deserializer, ADC and demodulator (including any

amplifiers) on the other hand. Equivalently, we define the bit energy or energy per bit of a transmission as

$$E_{bit} = E_{b,tx} + E_{b,rx}, \tag{2.12}$$

where $E_{b,tx}$ and $E_{b,rx}$ are the mean energy consumption spent in transmission and reception, respectively. Leakage currents of the inactive transmitters or receivers are neglected. For a wireless interface with complete transmitter chain power P_{tx} and receiver chain power P_{rx}, both for a data rate R, the energies are given by $E_{b,tx} = \frac{P_{tx}}{R}$ and $E_{b,rx} = \frac{P_{rx}}{R}$.

The energy will be calculated as a function of multiple design parameters that are of interest in the design space exploration of complete computing systems. In particular, we are interested in seeing how the energy scales with the data rate R and the channel losses PL. These, in turn, imply an indirect scaling in terms of (a) operation frequency f, because higher data rates may require shifting to higher frequency bands pursuing larger bandwidths; and (b) the transmission range d_{max} as it determines the worst-case losses. An important remark needs to be done regarding the channel losses and its relation with distance: authors in [43] propose and discuss a figure of merit for wireless transceivers which encompasses both their energy efficiency E_b and transmission range d_{max} by means of the following expression: $\Phi = \frac{E_b}{\sqrt{d_{max}}}$. Such a figure may be consistent with the intuition that the channel within computing systems may yield a loss exponent lower than 2. As in the case of area modeling, each of the suggested exploration parameters affects differently to the various analyzed components, whose power is calculated as follows.

Serialization and deserialization: for this we will take base on numbers from specific designs that cover the speeds required in our analysis [33–35]. For instance, data from [34] indicates a power of around 10 mW to achieve a line rate of 14 Gb/s at 65nm CMOS, with supply voltage of 1.0–1.15V, and assuming that the serializer and deserializer take the same power. Being digital logic mostly, consumption scales down with technology. Thus, at the same clock speed, reducing to 14nm CMOS with a supply voltage of 0.7–0.8V would lead to a power of around 7 mW or less, this is, 0.5 pJ/symbol. We assume that power consumption scales linearly with data rate (in other words, constant energy per symbol) as it implies faster switching of the multiplexers and demultiplexers. We also note that this data is used as a baseline, as we can expect further power reductions with newer technology or with circuit optimization.

Data converters: we will use data from Murmann's survey, described in previous subsections, which provide figures of merit directly related to the energy per converted bit. In particular, we will use a mixture of two figures of merit. According to Murmann [2], the Walden figure of merit FOM_W is more suitable for low-resolution designs, whereas Schreier's figure of merit FOM_S is better for assessing high-resolution designs. Hence, we will use the former if the transceiver requires an ADC with less than 8 bits of $ENOB$, and the latter otherwise. In particular, we will obtain the figure of merit as the value of the envelope line at the sampling frequency demanded by the transceiver. Then, the power is estimated by isolating it from the appropriate figure of merit. If needed, we will assume that the power of a DAC is similar than that of an ADC.

Transceiver: for such a central part of the model, we will be using a bottom-up approach. In essence, we aggregate the area of individual components required to support a certain modulation. Each of the components will have its own model, namely:

- **Power amplifier:** we calculate the required output power P_t through a link budget analysis that takes into consideration the frequency, SNR requirements, the assumed losses in the channel, and a given noise figure at the receiver. It is then considered that an amplifier only needs to be designed so that $P_{sat} \approx P_{out}$. Then, we check a well-known graph shown in [3] which relates the maximum PAE that a power amplifier of saturation power P_{sat} at high frequencies (> 50 GHz) can provide. Once the PAE is obtained, we can calculate the consumed power by isolating P_{DC} from Equation (2.9) with $P_{out} = P_{sat}$, provided that P_{in} (coming from the modulator) is known.

- **Low-noise amplifier:** in this case, we use a figure of merit which is widely considered in the papers on low-noise amplifiers, namely $FOM_{amp} = \frac{G \cdot BW}{(NF-1) \cdot P_{DC}}$ where $G = 10^{G_{dB}/10}$ is the gain magnitude, BW is the bandwidth in GHz, NF is the noise figure magnitude, and P_{DC} is the power consumption. Different works in the literature have reported designs with FOM values of around 10 at 60 GHz [44] and up to 1 at 240 GHz [45, 46]. We fix the G and NF values and leave BW as input parameter to sweep.

- **PLL:** although a survey exists that covers a wide variety of PLLs [37], the efficiency results are reported through figures of merit that do not indicate the frequency of the PLL. Instead, we use a model provided by RWTH Aachen [39] which calculates the power consumption of a PLL based on a specific design and timings, capacitances, or voltages obtained for a specific SiGe technology node. This model allows to obtain the power as a function of the PLL frequency.

- **Mixers:** we assume a fixed power consumption of 2 mW without conversion gain in light of the results observed in various transceivers in the literature [7, 47] or the 240 GHz transceiver tapeout presented in Deliverable D1.1 [39].

- **Filters:** the power consumption of a filter is evaluated by fixing the quality factor Q of the filter, the SNR of the baseband signal, and the proportionality constant n. The power consumption is then calculated as a function of the bandwidth of the analog signal BW, which is in turn function of the modulation rate and index, as $P_{filter} = n \cdot K_B \cdot T \cdot Q \cdot SNR \cdot BW$.

2.3 Link Layer Modeling

The link layer of design assumes the existence of a physical layer providing a raw transmission rate R through a number of channels. These channels may be shared among a number N of wireless interfaces, whose MAC protocol will determine when to send and how to manage collisions, if any. As described in Section A.4, this impacts on the latency of transmissions as it adds a delay τ_{MAC} and scales the throughput by a factor of $\mu < 1$, which is the fraction of the channel capacity effectively used for transmissions. Therefore, the ideal protocol would manage access fairly across all nodes sharing the channel while $\tau_{MAC} \to 0$ and $\mu \to 1$.

The characterization of the pair $\{\tau_L, \mu\}$ will form our model of the link layer. To obtain both, one can resort to analytical models like those of the original works on Carrier-Sensing Multiple Access (CSMA) [48]. However, these depend on multiple assumptions that do not necessarily hold in the WNoC scenario, e.g. Poisson arrivals.

Instead, we resort of event-driven simulation of the WNoC to obtain the performance metrics as we depict in Section 2.3.1. We then model the performance of the wireless link as described in Section 2.3.2.

2.3.1 Simulation

The characterization of the performance and efficiency at the link level requires obtaining the latency and throughput of a link shared by a variable number of nodes, for different types of traffic, and increasing the load from a few packets per second up to levels where the link is expected to saturate. This requires implementing the MAC protocols and traffic generators within a network or architecture simulator that replicates the WNoC scenario.

In our case, evaluations are carried out in the cycle-accurate architecture simulator Multi2sim [49]. Multi2sim has been augmented with wireless on-chip communication modules that model collisions and multiple MAC protocols. Multi2sim admits synthetic traffic and multithreaded applications. As we will see in Chapter 5, in this deliverable we propose a new family of MAC protocols, called FUZZY TOKEN. To evaluate its performance, we implement it within Multi2sim and and compare the average packet latency and throughput against that of two baseline protocols: a CSMA-like protocol called BRS [26] and a token passing protocol. Next, we describe the traffic patterns used in the simulations.

2.3.1.1 Traffic Patterns

Typically, NoCs are evaluated with synthetic traffic models that have, as main parameter, the injection rate λ in packets/cycle. Widespread simple models assume a Poisson process with the same average injection rate for all cores. However, we have seen in Section A.1 that traffic in the multicore scenario shows a clear self-similarity caused by the data dependencies within the applications. Moreover, common memory patterns such as producer-consumer lead to some cores transmitting more often than others. Our traffic model takes these aspects into account.

To account for the effect of self-similarity, we model a heavy-tailed distribution of traffic via a Pareto distribution [50]. In more detail, injection is composed by bursts of length t_{ON} followed by periods of silence of length t_{OFF}. Bursts and silences are expressed as

$$\begin{aligned} t_{ON} &= \frac{b_{ON}}{(1-U)^{1/a}} \\ t_{OFF} &= \frac{b_{OFF}}{(1-U)^{1/a}} \end{aligned} \tag{2.13}$$

where $b_{ON} = 1$, $b_{OFF} = b_{ON}(\frac{1}{\lambda} - 1)$, U is a random generator following a uniform probability distribution of values between 0 and 1, and $a = 3 - 2H$. The value of $H \in [0.5, 1)$, the Hurst exponent, leads to increasing degrees of self-similarity as $H \to 1$ [28]. Moreover, to model an uneven injection of traffic across nodes, we make use of the hotspotness parameter σ proposed in [28], where σ represents the standard deviation of the spatial injection distribution. Low values of σ represent higher concentrations of traffic around a few cores. These are consistent with the definitions given in Section A.1 and of extreme importance in the evaluation of MAC protocols, since their performance is very sensitive of these spatiotemporal characteristics of traffic.

2.3.1.2 Performance Metrics

The MAC protocols are evaluated on the basis of average packet latency and throughput as described in Section A.4. Latency is defined as the time between the generation of a packet and its correct reception at all the intended destinations, measured in clock cycles. Throughput is measured in transmitted bits per clock cycle. Another important metric is the energy consumed per transmitted bit, which may increase due to collisions and also depends on the number of nodes sharing the channel. Assuming enough transmission power to reach the N nodes sharing the channel, and considering the collisions that may occur, we calculate the energy per bit $E_{bit,link}$ at the link layer as

$$E_{bit,link} = E_{mac} + E_{b,N}\left(1 + \frac{L_{pre}}{L_{tx}} N_{re}\right), \tag{2.14}$$

where E_{mac} is the energy consumed by the MAC protocol to react to collisions or pass the token, and the term $(1 + \frac{L_{pre}}{L_{tx}} N_{re})$ models the energy wasted in collisions. Here, L_{pre} and L_{tx} are the length of a collided and a successful transmission, which may differ depending on the protocol, and N_{re} is the average number of retransmissions per successfully transmitted packet. Finally, $E_{b,N}$ is energy of a non-colliding message transmitted to N nodes, given by

$$E_{b,N} = E_{b,tx} + N \cdot E_{b,rx}. \tag{2.15}$$

We note that N refers to the number of wireless interfaces that are tuned into the channel, which may differ from the total number of wireless interfaces. If $N = 1$, the transmission is unicast. The energy and area overheads of the MAC protocol are assessed via estimation of the buffering requirements of the protocol and other digital components. Finally, all parameters except N_{re}, which is obtained from the simulation statistics, come from the system/MAC specifications and do not require simulation.

2.3.2 Modeling

2.3.2.1 Performance Modeling

The characterization of the performance and efficiency of a given MAC protocol are generally describe via its latency-throughput characteristic, which is built by evaluating the protocol for an increasing load. The canonical behavior of the protocol, in average, is a gradual increase of the latency as the load increases until the link saturates. Saturation means that the throughput will not increase further even if the load keeps rising. Beyond saturation, packets are lost if the queue is finite; otherwise, the latency tends to infinity. Before saturation, the latency tends to increase following a parabolic or even exponential function, increasing the slope as the load increases. Hence, assuming an infinite buffer, we can describe the average latency with a quadratic model as

$$\begin{aligned} \tau_L = \alpha\lambda + \beta\lambda^2 + \tau_{ZL} \quad &\text{for} \quad \lambda \in [0, \lambda_{sat}] \\ \tau_L = \infty \quad &\text{for} \quad \lambda > \lambda_{sat} \end{aligned} \tag{2.16}$$

Similarly, we can describe the throughput M as a function of the load in packets per second and the average length of a packet L as

$$\begin{aligned} M = \lambda L \quad &\text{for} \quad \lambda \in [0, \lambda_{sat}] \\ M = \lambda_{sat} L \quad &\text{for} \quad \lambda > \lambda_{sat} \end{aligned} \tag{2.17}$$

In the model, the parameters are thus as follows:

- **Saturation throughput λ_{sat}:** referring to the load or throughput at which the latency exceeds a given latency threshold, e.g. 5$\times$ the zero-load latency. The ideal value is R, this is, that the link saturates only when it reaches is maximum capacity.
- **Zero-load latency τ_{ZL}:** referring to the link latency obtained with infinitesimally small load. The ideal value corresponds to the transmission time with $\tau_{MAC} = 0$.
- **Latency function coefficients α and β:** referring to the gradual increase of latency observed at moderate, non-negligible loads. The ideal value is $\alpha = 0, \beta = 0$, this is, that latency remains small no matter the offered load.

To fully model the performance and efficiency of the MAC protocol, simulations shall be repeated for different number of wireless interfaces N sharing the channel and different types of traffic, i.e. different values of H and σ.

2.3.2.2 Area and Energy Modeling

Finally, the area and energy overheads of the protocol should be simulated by synthesizing the control circuits and evaluating their power consumption and silicon area. However, given the simplicity of the protocols, these overheads can be generally be neglected. We provide some figures later in Chapter 5.

3. Channel Models

This chapter is devoted to the modeling of wireless channels within computing packages using the methods and results from [1] as a solid ground. The chapter is divided in several sections, each of which describes and evaluates a given package: flip-chip in Section 3.1, interposer in Section 3.2, and wirebond in Section 3.3. In more detail, these sections make a brief summary of the results for each package and the main model that can be extracted out of them. A more comprehensive list of models is given in Appendix C.

3.1 Flip-chip package

3.1.1 Environment Description

An instance of a complete flip-chip package with solder bumps is shown in Figure 3.1. During the manufacturing process, the solder bumps are deposited on the chip pads and, then, the chip is flipped over and its solder bumps are aligned precisely to the pads of the package carrier external circuit.

The layers are described from top to bottom as summarized in Table 3.1. On top, the heat sink and heat spreader dissipate the heat out of the silicon chip, as they both have good thermal conductivity. Bulk silicon serves as the foundation of the transistors. This layer has low resistivity (10 Ω·cm), which is convenient for the operation of transistors, but not for electromagnetic propagation [51]. The interconnect layers, which occupy the bottom of the silicon die as shown in the inset of Fig. 3.1, are generally made of copper and surrounded by an insulator such as silicon dioxide (SiO_2) [52]. Finally, we find a package substrate or PCB below the bump array. Although the material of the carrier may be alumina or similar, we model it as perfect electrical conductor due to the existence of a dense metallic redistribution layer within it.

The bulk silicon used in the chip substrate generally has low resistivity, and therefore a thin substrate is preferred [31]; whereas materials used as heat spreaders have low electrical losses [51] and rather thick layers are desirable. To evaluate this impact in our simulations, we assume that both the substrate and the heat spreader, Aluminum nitride (AlN) in our case, can have a thickness of either 0.1 or 0.5 mm each. On the sides of the die, we assume an empty space of variable size filled with air or epoxy. The package is laterally enclosed with a metallic lid.

Table 3.1: Characteristics of the layers in a flip-chip package and default dimensions.

	Thickness	**Material**	ε_r	tan(δ)	ρ
Heat sink	0.1–0.5 mm	Aluminum	PEC	PEC	PEC
Heat spreader	0.1–0.5 mm	Aluminum Nitride	8.6	$3{\cdot}10^{-4}$	–
Silicon die	0.5 mm	Bulk Silicon	11.9	–	10 Ω·cm
Insulator	10 μm	SiO_2	3.9	0.025	–
Bumps	87.5 μm	Cu and Sn	PEC	PEC	PEC
Redistribution layer	3 μm	Copper	PEC	PEC	PEC
PCB	0.5 mm	Epoxy resin	4	–	–

Table 3.2: Package parameters for flip-chip.

Parameter	**Default Value**	**Variations**	**Units**
Die size	8	12, 16, 20	mm
Silicon thickness	0.1	0.5	mm
Heat spreader thickness	0.5	0.1	mm
Lateral space material	Vacuum	Epoxy	N/A
Lateral space dimensions	1	1.4, 1.8	mm
Frequency	60	120, 180, 240	GHz

3.1.2 Summary of Results

3.1.2.1 Frequency Analysis

The starting model quantifies the path losses of a flip-chip at 60GHz for all the default dimensions given at Table 3.1. Different combinations of silicon and heat spreader are simulated to obtain the dependence of the path losses with the distance. Fig. 3.2 plots the path losses for combination of silicon and AlN. For all of the combinations its seen that the path losses points are scattered, which can mean that the energy is not received from a fixed point, but from many reflections. Hence, the path loss depends more on the position of transmitter and receiver. Still, a linear regression seems to suggest an upward trend with distance for all of the combinations. Also, it is seen that the benefits of thinning the silicon layer down are significant. A 100-μm chip has a path loss ranging between 20 and 45 dB, whereas packages with thick silicon have an extra 30 dB of path loss for the worst case. The AlN thickness also has a little impact on the path loss, affecting distant links mostly. This effect is more noticeable when the

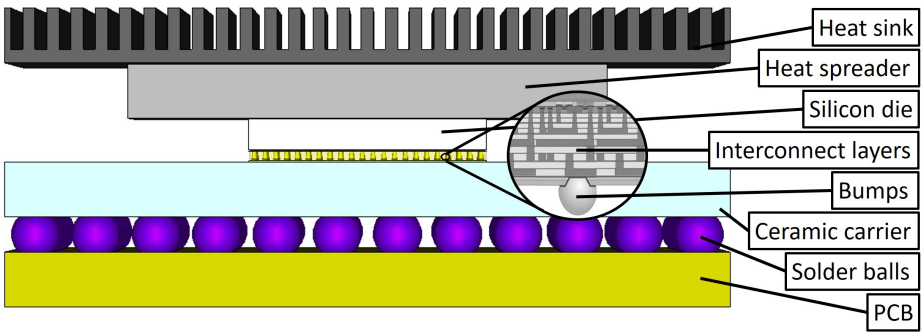


Figure 3.1: Schematic of the layers of a flip-chip package.

Table 3.3: Models for flip-chip package channel in the frequency domain with multiple component thicknesses.

Frequency	Die-side	Si	AIN	Margin space	PL_0	γ
60GHz	8	0.1	0.1	1	28.5	2.037
60GHz	8	0.1	0.5	1	32.14	1.0243
60GHz	8	0.5	0.1	1	32.18	4.8502
60GHz	8	0.5	0.5	1	30.91	3.5080

silicon die is also thick, because the waves entering the silicon suffer very significant losses as compared to when the silicon is thin, whereas those travelling through the AIN layer are less attenuated.

As it is observed in Table 3.3, the effect of the silicon and AIN thickess is notable, especially for thin silicon. The effect is clearly more visible at longer distances, as PL_0 does not change significantly, but γ does. The waveguiding effect of thin silicon and thick AIN is notable, leading to $\gamma \approx 1$.

After a first round of simulations on 60 GHz, the frequency is gradually increased until 240 GHz. This is the frequency band at which the test Silicon-Germanium (SIGe) transceivers will operate in co-integration with the graphene antennas in WiPLASH.

For the frequency scaling analysis we observe on Fig. 3.3 that the path losses increases with the frequency. At 60 GHz, the path loss ranges between 30-–40 dB, approximately. The path loss rises up to 55 dB for 240 GHz. Since the size of the ports is not modified when changing frequency, and since the antenna and mismatch losses are removed from the channel response, this increase in path loss is not due to the antenna, so it can be due to an increase of losses in the material.

The models extracted from this frequency scaling experiment are given in Table 3.4 and show how the increase of frequency leads to increase of either the base PL_0 or the exponent γ, except for $f = 120$ GHz, which is already documented in Deliverable D3.1 [1].

For more graphical summaries of the channel characteristics of a flip-chip package in the frequency domain, we refer the reader to Deliverable D3.1 [1]. The resulting models for all the simulations are tabulated in Appendix C.

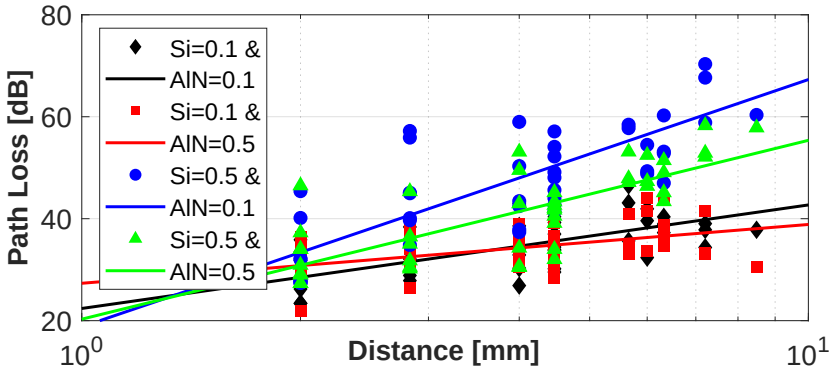


Figure 3.2: Path losses for a flip-chip at 60GHz for different silicon and AIN thicknesses.

Table 3.4: Models for flip-chip package channel in the frequency domain with multiple frequencies.

Frequency	Die-side	Si	AIN	Margin space	PL_0	γ
60GHz	8	0.1	0.5	1	32.14	1.0243
120GHz	8	0.1	0.5	1	18.79	0.7214
180GHz	8	0.1	0.5	1	31.69	2.7625
240GHz	8	0.1	0.5	1	44.49	1.3667

3.1.2.2 Time Analysis

In the time domain the scale of the delay spread with distance is assessed. The antennas are excited with an extremely short Gaussian pulse whose spectrum spans all frequencies between 10GHz and 1THz. Figure 3.4 shows the delay spread for all the combinations of substrate and heat spreader thickness. The delay spread is larger when thick layers are used. The main components of the signal get weaker and appear more delayed reflections that create a longer tail. From the figure we have the best design point is thick silicon and thin AIN, this leads to a delay spread below 0.05 ns. this may occur because the thick silicon layer kills all long multipath components.

Table 3.5 depicts the model parameters for the linear scaling of the delay spread in the different flip-chip variations. We observe how, indeed, the waveguiding effect of thin silicon and thick AIN leads to lower coherence bandwidth, whereas the cases of thin AIN favour having a better channel in terms of dispersion.

For more graphical summaries of the channel characteristics of a flip-chip package in the time domain, we refer the reader to Deliverable D3.1 [1]. The resulting models for all the simulations are tabulated in Appendix C.

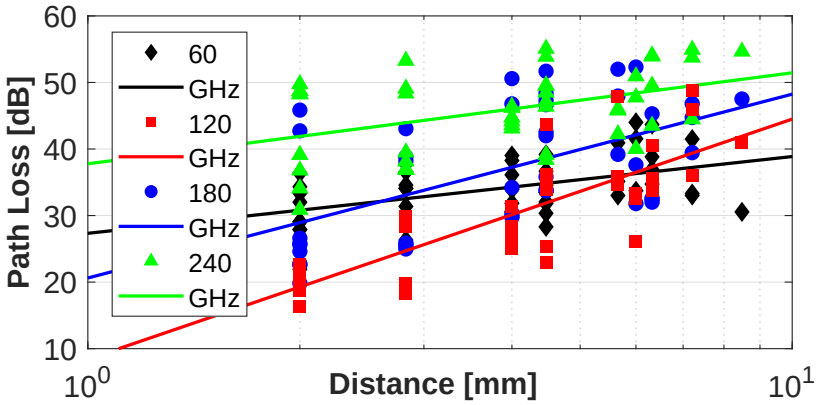


Figure 3.3: Path losses for a flip-chip different frequencies for silicon thickness of 0.1mm and AIN thickness of 0.5mm.

Table 3.5: Models for flip-chip package channel in the time domain with multiple component thicknesses.

Die-side	Si	AIN	Margin	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
8	0.1	0.1	1	0.02165	0.0036	0.0617	16.213
8	0.1	0.5	1	0.05953	0.0038	0.0890	11.23
8	0.5	0.1	1	0.00922	0.0039	0.0467	21.4
8	0.5	0.5	1	0.02495	0.0085	0.0960	10.4

3.2 Interposer package

3.2.1 Environment Description

Figure 3.5 shows a schematic representation of an interposer-based package. The process of integration here is similar to that of flip-chip, but with a few extra added steps. In particular, the interposer is a thin silicon chip that interfaces the PCB/carrier with its array of solder bumps at a similar granularity than a flip-chip. On top, however, the contacts are patterned at a finer granularity. The top side of the interposer interfaces with the chiplets, which are integrated using a flip-chip technique. Therefore, the chiplets have the same structure that the one summarized in Section 3.1. As for heat dissipation, we can consider that each chip is added its heat spreader individually and then covered by a common heat sink.

Table 3.6 depicts the layers from top to bottom, whereas Table 3.7 lists the different variants that we model here. On top, the heat sink and heat spreader dissipate the heat out of the silicon chip. Bulk silicon (10 Ω·cm) serves as the foundation of the transistors in each chiplet. The interconnect layers reside within the silicon dioxide (SiO_2) insulator. Then, below the fine array of micro-bumps, we find the silicon interposer. Interposers can be (i) active, which include active devices and are implemented in bulk silicon, and (ii) passive, which can be implemented in high-resistivity silicon [53]. Below the interposer, we model an interposer-wide bump array, and below it, a PCB.

Laterally, the cross-section of the interposer package resembles that of flip-chip, with the exception that void now appears not only between the chiplets and package

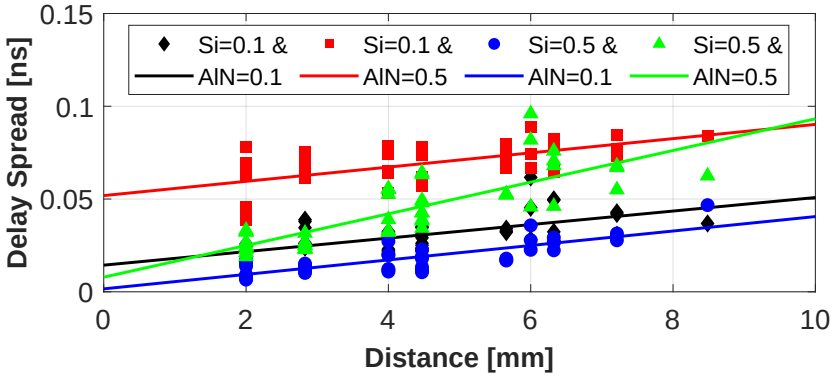


Figure 3.4: Delay spread of a flip-chip for different for silicon and AIN thicknesses.

Table 3.6: Characteristics of the layers in an interposer-based package.

	Thickness	**Material**	ε_r	tan(δ)	ρ
Heat sink	0.1–0.5 mm	Aluminum	PEC	PEC	PEC
Heat spreader	0.1–0.5 mm	Aluminum Nitride	8.6	$3{\cdot}10^{-4}$	–
Silicon die	0.5 mm	Bulk Silicon	11.9	–	10 Ω·cm
Insulator	10 μm	SiO_2	3.9	0.025	–
Microbumps	40 μm	Cu and Sn	PEC	PEC	PEC
Interposer	0.1 mm	High-Res Silicon	11.9	–	0.1 Ω·cm
Bumps	0.1 mm	Lead	PEC	PEC	PEC
Redistribution layer	3 μm	Copper	PEC	PEC	PEC
PCB	0.5 mm	Epoxy resin	4	–	–

Table 3.7: Package parameters for interposer.

Parameter	**Default Value**	**Variations**	**Units**
Interposer size	20	–	mm
Interposer resistivity	0.1	1, 10	Ω·cm
Number of chiplets	4	16	–
Chiplet silicon thickness	0.1	0.5	mm
Heat spreader thickness	0.5	0.1	mm
Chiplet separation	2	1, 4	mm
Filling material	Vacuum	Epoxy	N/A
Frequency	60	120, 180, 240	GHz

limits, but also between chiplets and between the interposer and the package limits. The simulated antennas are distributed among the chiplets homogeneously.

3.2.2 Summary of Results

3.2.2.1 Frequency Analysis

As was done for the flip-chip, the assessment begins quantifying the path losses and delay spread in the interposer packages at 60GHz for the default dimensions given at 3.7. Then, different combinations of silicon and heat spreader are also simulated.

When we evaluate the combinations of the silicon and AIN thicknesses Fig. 3.6, the result shows that thin silicon is preferable because it minimizes the losses of waves

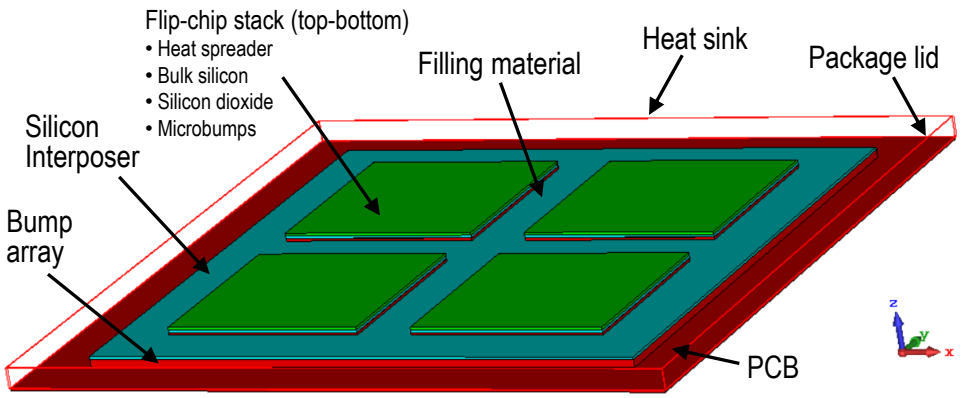


Figure 3.5: Schematic of the layers of an interposer package.

Table 3.8: Models for interposer package channel in the frequency domain with multiple component thicknesses.

Frequency	Die-side	Si	AIN	Separation	PL_0	γ
60GHz	20	0.1	0.1	2	31.42	3.3606
60GHz	20	0.1	0.5	2	31.15	2.5834
60GHz	20	0.5	0.1	2	39.77	4.5022
60GHz	20	0.5	0.5	2	31.06	5.0369

propagation through it.A thick AIN seems to aid reducing the path losses a little a bit further, but with marginal differences. For the best design point out of the simulated, the average path loss is around 40dB, with a worst case value of 50dB.

The modeling results, shown in Table 3.8 confirm the results above and describe a rather lossy environment due to the change of medium besides the presence of lossy silicon. The path loss exponent is consequently large.

Since the objective of the project is to get higher in frequency, the next stage is repeat the simulations to get closer to 240GHz. Fig 3.7 suggest that lower frequencies are preferable. Scaling the frequency from 60GHz to 180GHz has a cost of around 10dB. The reason may be the increase in losses of the different material found along the path.

Models in Table 3.9 confirm the results at different frequencies, leading to a rather lossy behavior that is exacerbated at higher frequencies.

For more graphical summaries of the channel characteristics of an interposer package in the frequency domain, we refer the reader to Deliverable D3.1 [1]. The resulting models for all the simulations are tabulated in Appendix C.

3.2.2.2 Time Analysis

The dispersion is evaluated first within an interposer package of 20mm for different substrate and heat spreader thicknesses. To this end we use a picosecond-long impulse signal covering the spectrum from 0.01 to 1THz. The results of the simulation shows that the delay spread does not exceed 0.25ns in any of the evaluated scenarios.

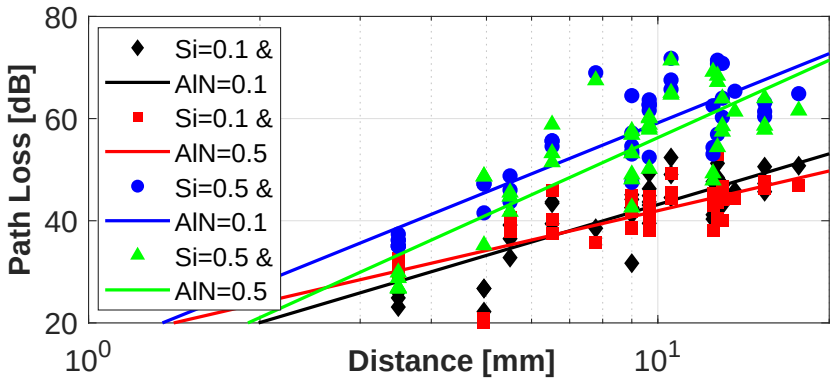


Figure 3.6: Path loss in an interposer at 60GHz for different combinations of substrate and heat spreader thicknesses.

Table 3.9: Models for interposer package channel in the frequency domain with multiple frequencies.

Frequency	Die-side	Si	AIN	Separation	PL_0	γ
60GHz	20	0.1	0.5	2	31.15	2.5834
120GHz	20	0.1	0.5	2	15.35	5.3794
180GHz	20	0.1	0.5	2	23	4.6952
240GHz	20	0.1	0.5	2	35.48	3.6252

In this case was noted that thick AIN leads to relatively higher delay spreads at short distances and better values at longer distances. On the other hand, a thin AIN layer lead to better results at short distances, but worse at long distances.The reason may be that the extra propagation length of having to go through the AIN layer, reflect on the heat sink, and propagate back to the receiving antenna, is proportionally larger at short co-planar distances. At longer distances, this extra thickness at the AIN layer actually aids propagation through waveguiding. Since we calculate the coherence bandwidth based on the worst-case delay spread, then it seems that thick AIN are preferable in this scenario.

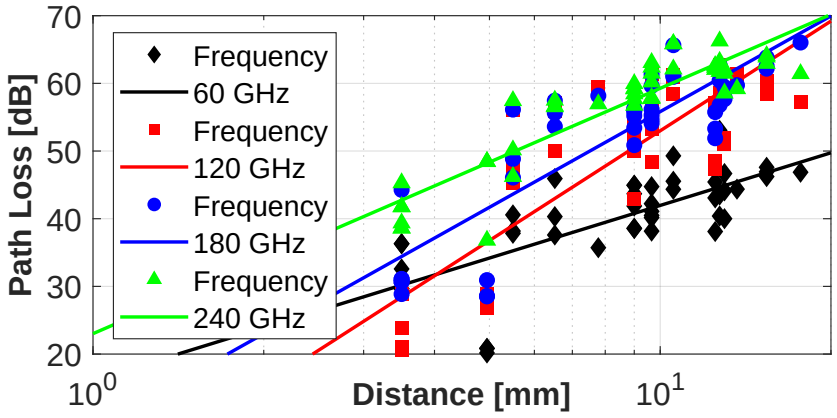


Figure 3.7: Path loss in an interposer at different frequencies.

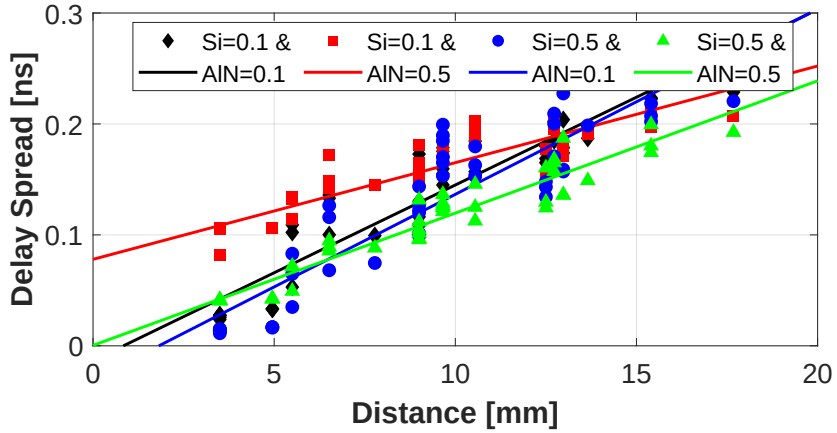


Figure 3.8: Delay spread in an interposer at Si and AIN thicknesses.

Table 3.10: Models for interposer package channel in the time domain with multiple component thicknesses.

Die-side	Si	AIN	Separation	Chiplets	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
20	0.1	0.1	2	16	0.1396	0.0068	0.2563	3.9
20	0.1	0.5	2	16	0.161	0.0075	0.3142	3.18
20	0.5	0.1	2	16	0.09765	0.0095	0.2566	3.89
20	0.5	0.5	2	16	0.0988	0.0075	0.2176	4.59

Table 3.10 shows how besides being a lossy environment, the interposer package is also rather dispersive. One reason is the inherently larger structure, and another one could be the multiple changes of medium that waves perform. This generates reflections and refractions that could reduce the coherence bandwidth significantly.

For more graphical summaries of the channel characteristics of an interposer package in the time domain, we refer the reader to Deliverable D3.1 [1]. The resulting models for all the simulations are tabulated in Appendix C.

3.3 Wirebond package

3.3.1 Environment Description

Figure 3.9 shows a three-dimensional schematic of a wirebond package. The key of this option is that it is a surface-mount technology that does not require any holes or vias to connect the external die to the system. The die is mounted in the upright position, with the insulator facing up and placed on top of an underfill material that fixes the chip mechanically to a metallic frame. The role of this frame is to mechanically interface the chip with the PCB. The electrical Input/Output (I/O) connections, on the other hand, are performed by means of bond wires stemming directly from the top metallization layers of the die and reaching the contacts in the PCB or ceramic carrier. Finally, the chip and the bond wires are covered by a mold compound and, on top, a ceramic enclosure.

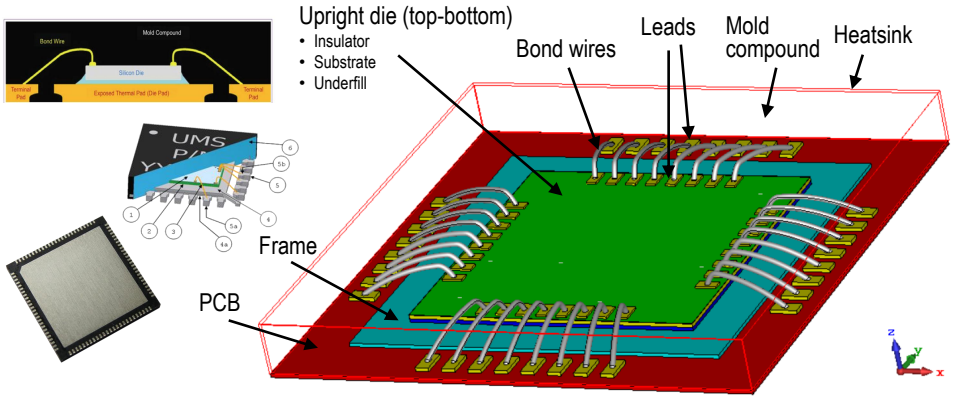


Figure 3.9: Schematic of the layers of an wirebond package, together with a top view and cross-section diagrams.

Table 3.11: Characteristics of the layers in a wirebond package.

	Thickness	**Material**	ε_r	tan(δ)	ρ
Enclosure	50 μm	Alumina	9.9	10^{-4}	–
Mold compound	0.45–1 mm	Epoxy resin	4	–	–
Insulator	10 μm	SiO_2	3.9	0.025	–
Silicon die	0.5 mm	Bulk Silicon	11.9	–	10 Ω·cm
Underfill	0.1–0.5 mm	Aluminum Nitride	8.6	$3{\cdot}10^{-4}$	–
Frame	0.1 mm	Copper	PEC	PEC	PEC
Leads	0.1 mm	Copper	PEC	PEC	PEC
Redistribution layer	3 μm	Copper	PEC	PEC	PEC
PCB	0.5 mm	Epoxy resin	4	–	–

Table 3.12: Package parameters for wirebond.

Parameter	**Default Value**	**Variations**	**Units**
Die size	8	12, 16, 20	mm
Bond wires	32	64, 128	–
Molding compound margin	0.1	0.05, 0.5	mm
Silicon thickness	0.1	0.5	mm
Heat spreader thickness	0.5	0.1	mm
Enclosure material	Alumina	PEC	N/A
Frequency	60	120, 180, 240	GHz

Table 3.11 depicts the layers from top to bottom, whereas Table 3.12 lists the different variants that we evaluate. On top, the ceramic enclosure and mold compound cover the entire system. Then we find the chip, where the silicon dioxide (SiO_2) insulator appears first hosting the different interconnect layers. Below, a thick layer of bulk silicon (10 Ω·cm) serves as the foundation of the transistors. Directly below the silicon die, we have the thermal interface material that leads to the metallic frame and the PCB below it.

Dies connected through bond wires are generally relatively small because only the periphery of the chip can be used to implement I/O connectors. The package extends laterally beyond the die first through the frame. There is another space between the frame and the limit of the package, which is necessary to host the PCB-side leads of the bond wires. The number of bond wires used by default is 32 and their pitch is calculated based on the specifications of the widespread QFN64 package.

3.3.2 Summary of Results

3.3.2.1 Frequency Analysis

As before, the first step is to quantifying the path loss in the wirebond with the default configuration at 60GHz. Then, the effect of modifying the silicon and heat spreader thickness is assessed. The path loss for the first configuration seems to be constant at different distances, with large values around 50 dB. We also observe that a few short links have an extremely high attenuation around 80–90 dB. Our hypothesis is that either (i) these are an artifact of the simulation, and should be ignored, or (ii) the package structure and the presence of resonating bond wires creates directions of minimum radiation, which leads to low lateral coupling at short distances. At more

distant links, energy may still come from reflections coming from the bonding wires or the end of the package.

Fig. 3.10 shows the path losses for all the combinations of Silicon and AIN. The main observation to make here is that, as usual, thick silicon harms the wireless channel by introducing significant losses. This is clearly observable at high distances. Another observation is that having a thick piece of AIN does not necessarily help reduce losses. The reason is that in the wirebond package the die is mounted upright, leaving the thermal pad at the bottom.

As in the previous sections, once the default configuration is evaluated, we raise the frequency up to 240GHz to reproduce a scenario relevant to the objectives of the WiPLASH project. Figure 3.11 suggest that an increase in the antenna frequency may have a negative impact on the path loss. We see a that the path loss reaches an average amount of upto 80 dB when reaching 240 GHz. Moreover, the upscaling in frequency does not avoid the the presence of extremely attenuated links at short distances, which maintain a similar range of values around 100–120 dB.

For more graphical summaries of the channel characteristics of a wirebond package in the frequency domain, we refer the reader to Deliverable D3.1 [1]. The resulting models for all the simulations are tabulated in Appendix C.

3.3.2.2 Time Analysis

Next, the dispersion is evaluated for a wirebond package of 8mm for different substrate and heat spreaders thicknesses. To make sure that the dispersion limits are given by the channel and not the excitation port, we use an picosecond-long impulse signal covering the whole spectrum from 0.01 to 1 THz. From this plot we observe that the wirebond package has a reasonable delay spread, with worst-case values well below below 0.15 ns. We also see how thin silicon alternatives are again preferable by a long margin as they reduce the worst-case delay spread by around 30%. For thin silicon, the impact of AIN is higher at short distances. At high distances, its impact becomes marginal.

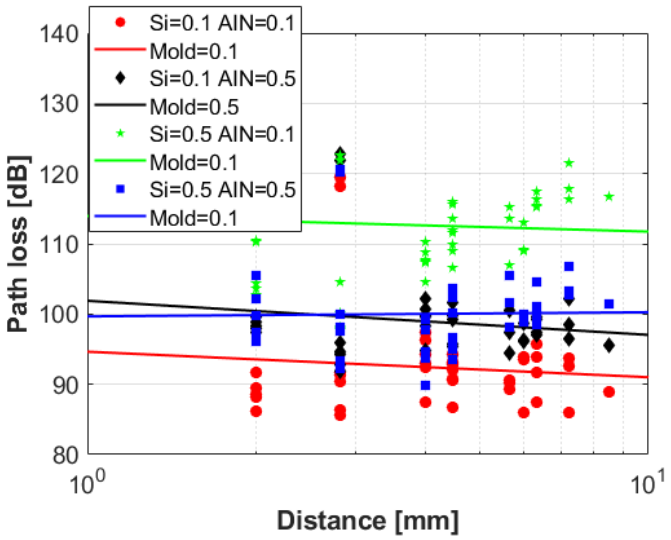


Figure 3.10: Path losses for wirebond package at for different substrates 60GHz

For more graphical summaries of the channel characteristics of a wirebond package in the time domain, we refer the reader to Deliverable D3.1 [1]. The resulting models for all the simulations are tabulated in Appendix C.

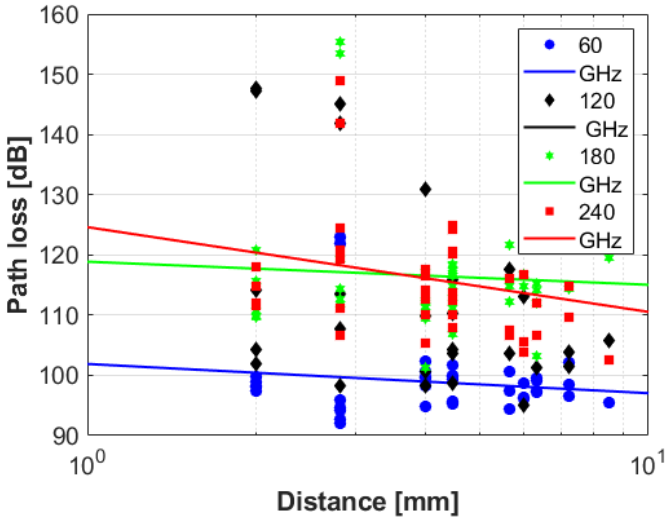


Figure 3.11: Path losses for wirebond package at different frequencies

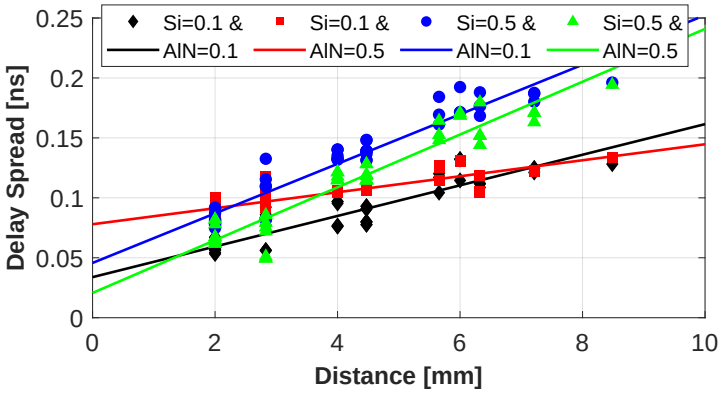


Figure 3.12: Path losses for wirebond for different substrate and heat spreader thicknesses

4. Physical Layer Models

The appeal of high-speed wireless communication for personal and chip-area applications, together with the availability of the millimeter-Wave (mmWave) and THz spectrum and technology to exploit its ample bandwidth, has been the main driver for the design of highly integrated antennas [54] and transceivers [7, 8]. As a result, recent years have seen the emergence of designs at multi-Gb/s rates with an area and power compatible with the wireless in-package communications.

While the myriad of transceiver proposals is useful to estimate the benefits of wireless in-package communications in specific architectures, assessing the potential of the approach on a broader scope is difficult due to the lack of models relating performance and resource consumption. Such models are generally not available due to the complexity of integrated RF transceiver design, where the performance of the multiple components needs to be carefully balanced across the transceiver chain to fulfill with a set of specifications. Design space explorations like those proposed in the work packages WP4 and WP5 of WiPLASH require, instead, models not bound to a set of specifications, but rather to a relatively wide range of values for each performance of resource consumption metric. In the pursuit of those models, we first describe and quantify the main requirements of the scenario in Section 4.1, which guides in the setting of certain performance parameters such as the SerDes speed, the $ENOB$ of the data converters, or the target modulations. Then, we detail the path to area and power models for wireless chip-scale communications in Sections 4.2.1 and 4.2.2.

4.1 Requirements of the Scenario

Figure A.1 from Chapter A provided a summary of the area and power requirements at different scales of an in-package network. Here, we detail the case for a regular manycore chip (an analysis of the multi-chip scenario would yield similar results). Besides the already known data and error rate requirement on the tens of Gb/s and 10^{-12}–10^{-15} ranges, area and power constraints are evident given the dimensions of a chip, 20×20 mm^2 or smaller, and the heat dissipation problems in manycore architectures leading to dark silicon.

Assuming a 100-core processor in a 450 mm^2 chip with the Thermal Design Power (TDP) of a Xeon Phi (210 W), we will thus have that each core can only take 4.5 mm^2 and at most 2.1 W of sustained power including the processor, memory, and communication sub-systems. Optimistically assuming the same budget for the three sub-systems, the Network-on-Chip (NoC) (including the wireless part, if any) should not exceed 1.5 mm^2 and not take more of 700 mW per core. Assuming again an equitable distribution of resources and neglecting network interface and MAC overheads, we would estimate the WNoC or WNiP to have a budget of around 0.75 mm^2 and 350

mW per core. Let this estimation serve as reasonable limits for the cost of a WNoC, noting that they would be increased or reduced depending on the actual distribution of resources and the number of cores or chiplets in the system.

From the perspective of data serialization, one needs to interface the system clock (which ranges from hundreds of MHz in embedded systems to a few GHz in high-end processors) with the modulation rate of the wireless link. Since high rates are expected due to the use of low-order modulations, one can expect serialization 8:1, 16:1, or even 32:1. These are already encountered in serial links with OOK or 4-PAM modulations at high speeds [34, 55]. We finally note that the serialization requirements could be alleviated if multiple channels (with multiple modulator chains) are employed.

From the perspective of data conversion, evident area and power constraints render equalization and other advanced signal processing methods prohibitive. Fortunately, the DACs may be completely bypassed if direct modulation is used. At the ADC, though, the sampling frequency will be pushed to speeds over tens of GS/s to comply with the modulation rate requirements. In contrast, the required ENOB will be relatively low as very few bits per sample are required, simply to differentiate signals of different power, i.e. coming from different distances. Finally, we reckon that moderate oversampling and additional bits per sample may be useful to improve performance with Return-to-Zero (RZ) techniques or adaptive decision circuits [56].

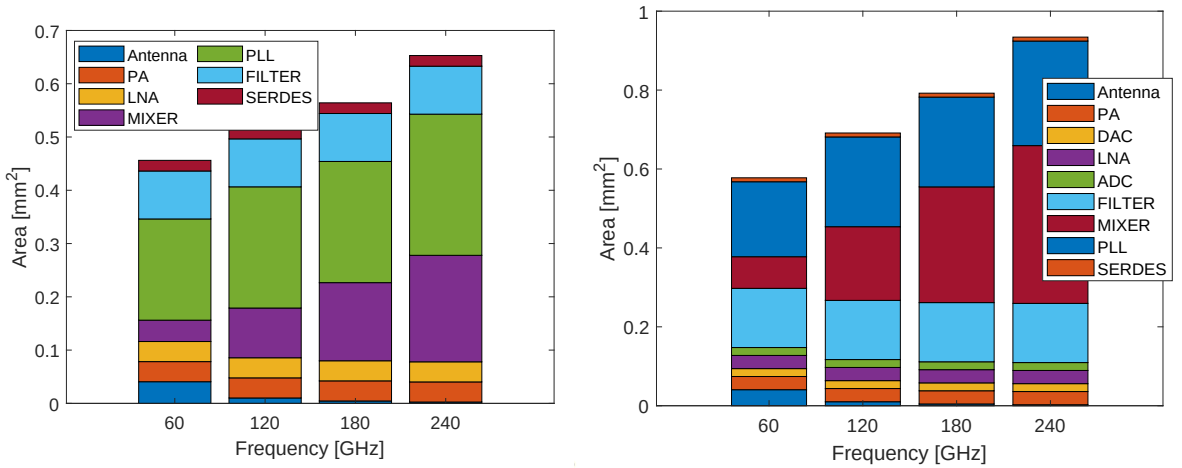
As for the transceiver, we have seen in prior chapter that the link budget might need to take into consideration relatively large path loss, thereby requiring amplifiers of significant gain. The use of low-order modulations sets the SNR requirement for $BER = 10^{-15}$ to 18 dB for OOK and 4-QAM, or 15 dB for BPSK. Beyond this, another stringent requirement is the use of a carrier frequency in the 60–300 GHz range, which may not be trivial to achieve with small area and power.

4.2 Resource Models

In this section, we assess how the different components scale as functions of relevant input parameters, to then put them together on a single metric following Equation (2.10) and (2.12) for area and energy, respectively. We consider two different modulations, On-Off Keying (OOK) and 4-Quadrature Amplitude Modulation (QAM), implemented via coherent modulation schemes as summarized in [7]. Relevant to the models, we assume $NF = 10$, $T = 300$, $ENOB = 6$, and $P_{in,PA} = -10$ dBm.

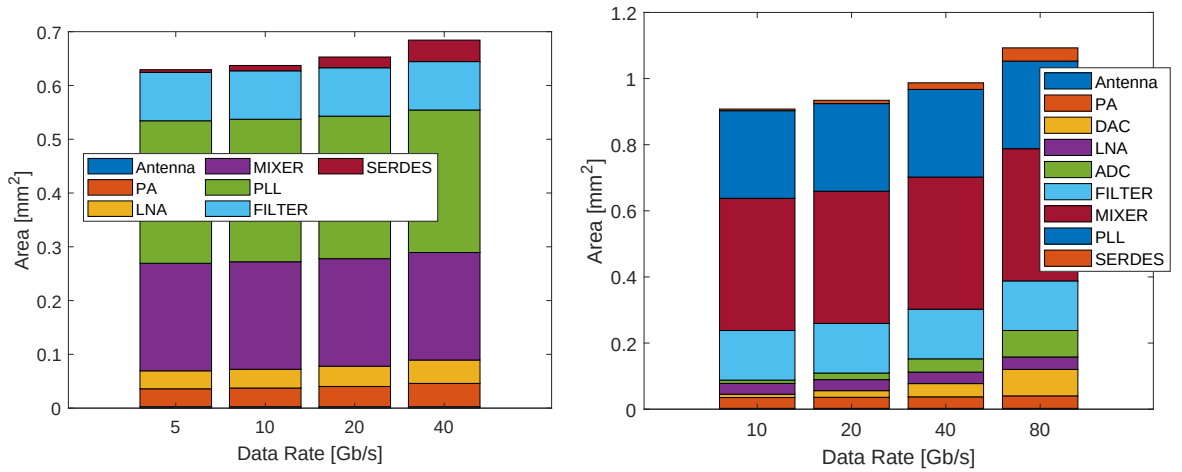
4.2.1 Area Models

Figure 4.1 shows the area consumed by the OOK and 4-QAM transceivers as a function of the operation frequency, from 60 GHz to 240 GHz. As expected, OOK consumes less area than 4-QAM due to the simpler transceiver architecture and not needing a DAC, for instance. In both cases, the area increases with frequency despite the antenna becoming negligible. The main contribution comes from the PLL and mixers, whose area increases with frequency (in the former case through the need of multipliers, and in the latter case due to the specific components chosen as models). From these observations, and while 4-QAM can be extended relatively easy to higher modulation orders for higher speed, we argue that OOK is a good choice due to the $\sim 30\%$ lower area for the same transmission speed. Moreover, OOK is compatible



(a) OOK (b) 4-QAM

Figure 4.1: Area as a function of the transceiver frequency assuming a data rate of 20 Gb/s and a loss of 40 dB.



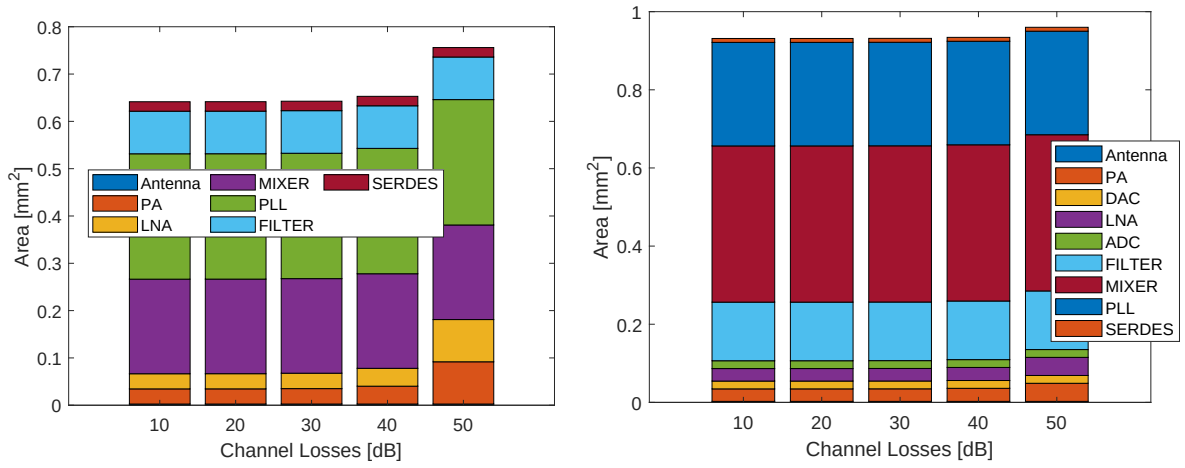
(a) OOK (b) 4-QAM

Figure 4.2: Area as a function of the transceiver data rate assuming a frequency of 240 GHz and a loss of 40 dB.

with non-coherent receivers that would remove the need for a PLL, using a much more compact VCO instead and reducing the area further towards achieving the goal of 100 Gb/s/mm^2.

Figure 4.2 illustrates how the area scales as a function of the transceiver data rate. We observe a rather large value in both cases, stemming from the fact that the default frequency is 240 GHz here. We also observe that the increase of area is marginal as compared to the increase of the data rate, and that the increase is more acute in the case of 4-QAM, mostly because of the area increase of the data converters –whose increase in sampling frequency has consequences. The rest of components do not increase considerably according to our model, yet we cannot discard that transceivers might actually get larger by the need of having higher bandwidth components.

Finally, Figure 4.3 shows the scaling of the transceiver with the channel losses. Here, the logic is simple: the only component affected is the power amplifier. For



(a) OOK (b) 4-QAM

Figure 4.3: Area as a function of the loss assuming a data rate of 20 Gb/s and a frequency of 240 GHz.

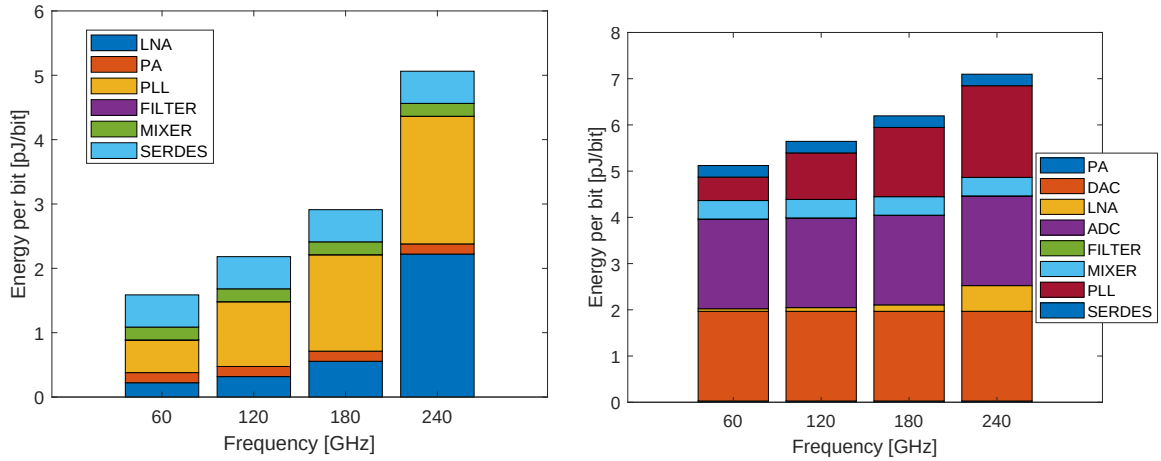
larger attenuation values, a larger number of amplification stages or a more complex designs may be required to compensate for the channel losses. The impact, though, is hard to grasp until the channel losses (which include the antennas as well) goes beyond 50 dB, which is well in line with the channel models shown in Chapter 3.

4.2.2 Power Models

Figure 4.4 shows the energy pr bit of the two modulations as a function of the frquency for both OOK and 4-QAM. A first observation is that energy is more proportional to frequency in the case of OOK, due to the large and increasing contribution of the PLL and LNA. In the former case, power scales linearly with frequency, whereas in the latter case, our model foresees that amplifiers become less efficient at high frequencies as they approach f_T and f_{MAX}. For 4-QAM, the presence of hefty components such as the data converters makes the impact of the PLL and LNA to be diluted. We also conclude that reaching the target of $\sim$1 pJ/bit is complex as the frequency increases. Yet still, such increase in frequency may be justified by the need of components with much higher bandwidth than current realizations.

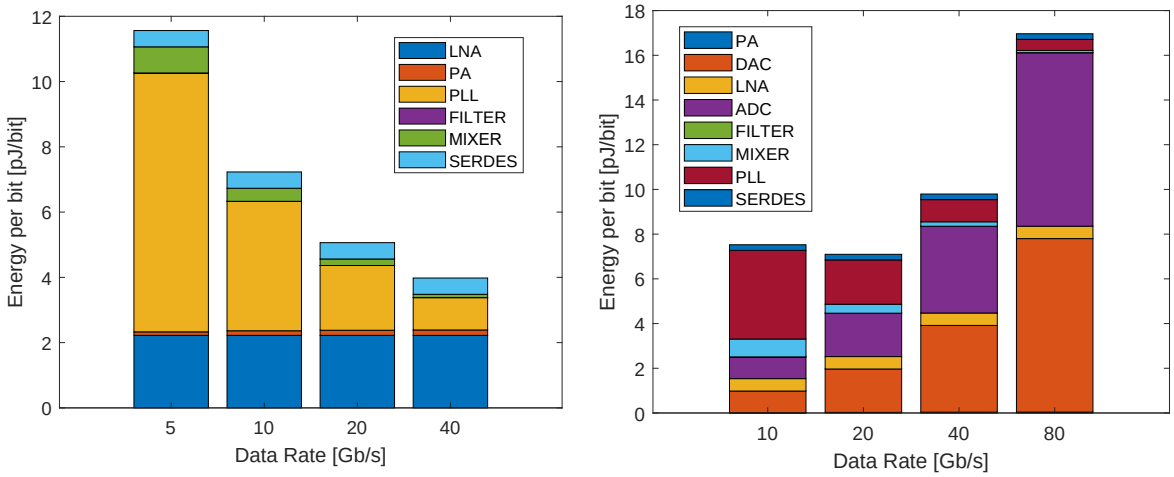
Figure 4.5 illustrates how the bit energy scales with the data rate. This analysis leads to two opposite tendencies. On the one hand, the bit energy of OOK is improved significantly as the data rate increases. This is because the PLL power consumption, which does not change, is amortized more and more at higher transmission rates. Other components such as the LNA or PA see their power to increase for larger bandwidth requirements, which leads to a rather constant bit energy across the board. On the other hand, the energy of 4-QAM first decreases due to the amortization of the PLL, but soon increases again due to the need of faster data converters. In fact, as data converters are pushed to the limit, their energy efficiency decreases, therefore becoming more expensive at higher speeds. From this analysis, we can conclude that OOK is a better option for moderate area and power at high modulation rates.

Finally, we plot the bit energy as a function of the channel losses in Figure 4.6. With our proposed models, the bit energy is rather constant for low loss values and



(a) OOK (b) 4-QAM

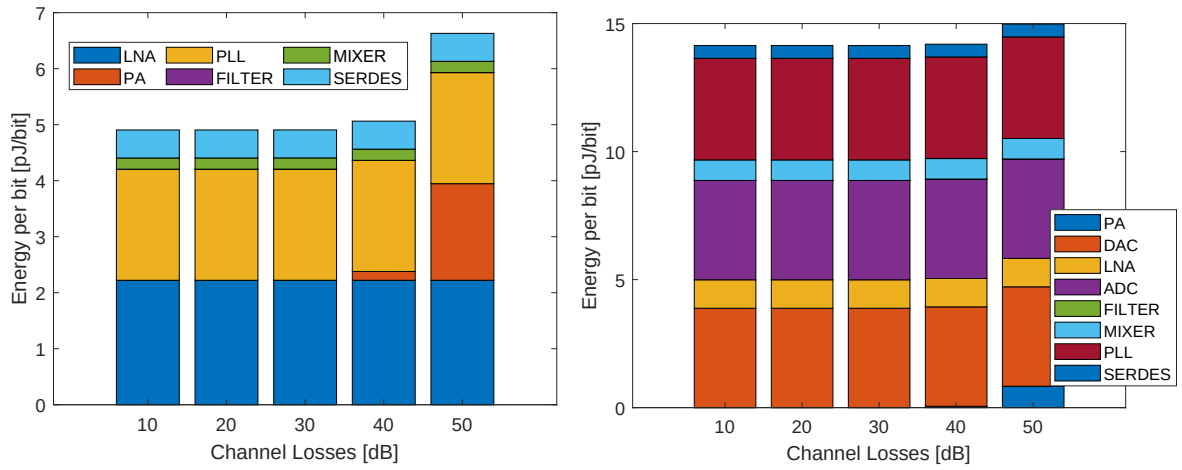
Figure 4.4: Energy per bit as a function of the transceiver frequency assuming a data rate of 20 Gb/s and a loss of 40 dB.



(a) OOK (b) 4-QAM

Figure 4.5: Energy per bit as a function of the transceiver data rate assuming a frequency of 240 GHz and a loss of 40 dB.

then increases significantly. This is because the effect of the power amplifier kicks in: for high channel losses, the power at the output of the power amplifier need to be very large. Increasing the saturation power implies more complex designs and possibly a reduction of the PAE as indicated in Chapter 2, thereby producing this increase at the right part of the plots.



(a) OOK (b) 4-QAM

Figure 4.6: Energy per bit as a function of the loss assuming a data rate of 20 Gb/s and a frequency of 240 GHz.

5. Link Layer Models

While physical layer models can provide the area and power consumption as a function of the transmission rate, error rate, and distance, these are only for a single pair of transmitter and receiver. In a WNoC, however, transmission channels are shared among many nodes as multicast/broadcast is a highly appreciated feature. Thus, link layer protocols become of paramount importance to ensure that the channels are shared fairly and efficiently. In particular, two overlapping transmissions in the same channel will fail with high probability. It can also happen that, in the attempt to avoid collisions, nodes miss opportunities to transmit. This will affect the latency and the throughput of the link.

Modeling of performance and overheads at the MAC layer is only found in seminal works in the area [48, 57] for very specific conditions such as Poisson traffic. Hence, we need to perform simulations specifically suited to the WNoC domain and extract performance models from those simulations. Next, in Section 5.1 we describe the protocols that serve as a baseline for comparison further in the chapter. Then, in Section 5.2, we propose a MAC protocol for the WiPLASH scenario. Finally, in Section 5.3, we present the results of our evaluation for a single channel, with views to extend it to multiple channels in future work.

5.1 Baseline Protocols

In an attempt to model different types of MAC protocols, our baselines consider a pure contention-based protocol resembling CSMA, a token passing variant as collision-free protocol, and a possible implementation of what would be an ideal protocol. More specifically, our simulations consider these protocols:

Carrier Sensing (`BRS`) with which we aim to represent contention-based protocols. We model the slotted version of the BRS-MAC protocol [26], using non-persistence and adopting the NACK burst mechanism to reduce the control overhead. The preamble size is fixed to 20 bits, which implies that the preamble accounts for a variable portion of the transmission. A packet will be considered lost in the unlikely case that it exceeds the maximum number of retries (8). Note that the network will most likely be saturated when this happens.

Token Passing (`W-TOKEN`) this category aims to represent a design family that relies in rigid strategies to avoid contention. In token passing, only the core that possesses the token is able to transmit [57]. One full packet can be transmitted in each round. We do not split long messages into flits here as the packet latency would be unacceptable, whereas bulk transmissions are not allowed for fairness reasons. Upon completion, or in case there is nothing to transmit, the token is handed off to the

next core. We assume that the token passing is performed implicitly through a virtual ring.

Centralized Buffer Arbitration (`W-CBUF`) for the sake of comparison, we also study the performance of a centralized MAC scheme. With unlimited resources, it would be possible to have an arbiter connected to every core with one-cycle bidirectional links and that works as follows. When a node is ready to transmit, it sends a request to the arbiter with its identity and the packet size. The arbiter stores this in a FIFO buffer and grants access to the node whose request is in the buffer head, waiting exactly the wireless transmission time between consecutive grants. Contention only appears when multiple nodes request access at the same clock cycle. This is resolved by the arbiter. This scheme therefore provides fair, ordered and contention-free access in a flexible way, with resources that are not available in other wireless networks. The main reason for evaluating this scheme is to motivate unconventional MAC designs and to quantify the improvement margin of the different protocols.

5.2 Proposed MAC Protocol: Fuzzy Token

In light of the existing gap between CSMA-like protocols and variants of token passing, which hampers the achievement of high performance across workloads, here we present a new family of protocols called FUZZY TOKEN. We limit our analysis to the single-channel version of this protocol, yet the algorithm is easily extensible to multiple channels with techniques similar to those used in CSMA and token protocols.

FUZZY TOKEN is a new MAC protocol capable of dynamically adapting to the demands of the WNoC for every application, minimizing the transmission latency of all nodes and increasing the network throughput regardless of the traffic patterns being served. To this end, the FUZZY TOKEN algorithm combines the strengths of token passing and random access.

Figure 5.1 illustrates the basic operation of FUZZY TOKEN. For a detailed description of the protocol, along with a list of design decisions and a walkthrough example of operation, we refer the reader to [58].

Time is divided into two kinds of periods or modes: *focused* token mode and *fuzzy* token mode. During focused token periods, only the token holder can transmit. If the token holder transmits, the channel is occupied during a number of cycles with guaranteed no collision, and the protocol remains in focused token mode. If the token holder does not transmit, then the protocol switches to fuzzy token mode. During such a period, the nodes within what we call the *fuzzy area* will contend for the channel. If a node wins the contention, it transmits successfully, and the protocol remains in fuzzy token mode. Collisions can occur in this period, which are resolved by using the NACK mechanism from [26]. However, colliding nodes do not back off; instead, they just wait for another opportunity to transmit in forthcoming cycles. In the case of a collision, the protocol switches back to the focused token mode. By alternating between the two modes, the protocol aims to take advantage of the capabilities of a fair and collision-free token passing, which works well for high, bursty, and evenly distributed loads; and of a contention-based protocol that performs better for moderate loads and hotspot traffic.

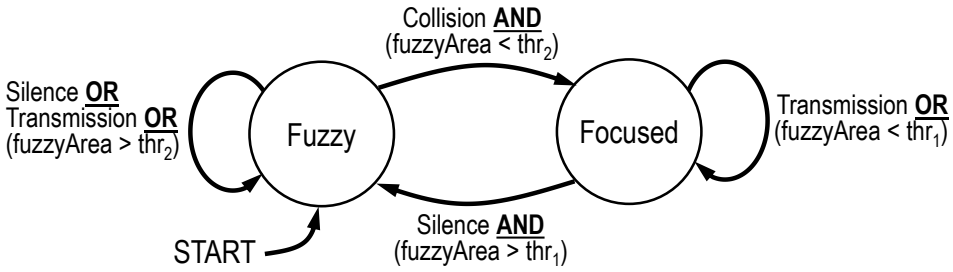


Figure 5.1: Basic state diagram of FUZZY TOKEN.

The amount of contention during the fuzzy token periods is controlled with the fuzzy area. The fuzzy area marks the nodes around the token holder that may be able to transmit in a given cycle. The main idea is to increase the fuzzy area when the load is low to give rapid access to the few nodes that want to transmit, and decrease it otherwise to minimize collisions. To this end, we increase the fuzzy area when a silence is detected and decrease it when there is a collision. In extreme cases, it is advisable to stop alternating between the focused and fuzzy modes. This is exemplified in Figure 5.2. When the fuzzy area is below a given threshold after many collisions, the network will benefit from staying in the focused token mode. On the contrary, when the fuzzy area is over a certain threshold after many silences, it is considered that the load is low enough to stay in fuzzy token mode. The normal alternating operation of the protocol is reinstated when the value of the fuzzy area increases/decreases again, respectively.

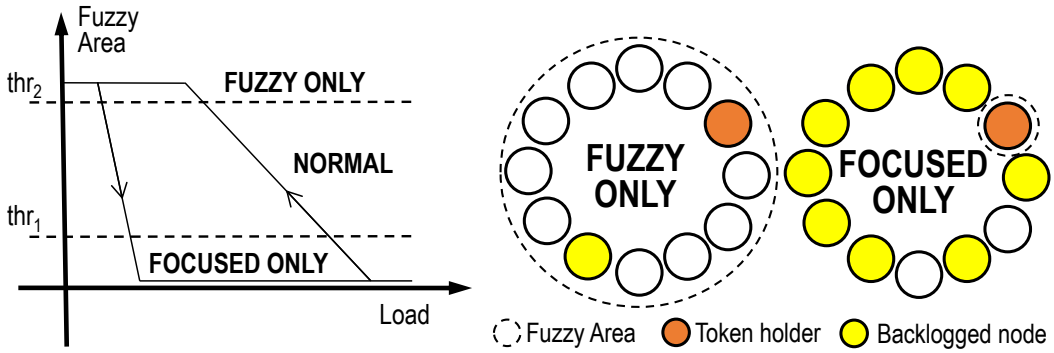


Figure 5.2: Transition chart (left) and extreme cases (right) of FUZZY TOKEN.

FUZZY TOKEN ensures fairness by circulating the token within the virtual ring. In more detail, the token is passed implicitly at every event (silence, collision, or successful transmission) regardless of the protocol mode. This is important because latency tails generated by unfair access will significantly slow down computation, even if they are infrequent. It is worth noting that, thanks to the unique characteristics of the on-chip scenario, all nodes have a consistent view of all events. This allows to pass the token implicitly and to update the fuzzy area values without explicit messages or centralized control.

The default FUZZY TOKEN configuration is: increase the fuzzy area by 1 in each silence, decrease the fuzzy area to 1 in each collision, thr_1 = 10%, thr_2 = 90%, no probability decay, and static token ring. This combination of parameters was found to be the optimum across the different workloads. For more details and a walkthrough example of the protocol, we refer the reader to [58].

5.3 Performance Models

The architecture and application parameters are summarized in Table 5.1. To evaluate its performance, we implement it within Multi2sim [49] and compare the average packet latency and throughput against that of the two baseline protocols. For a fair comparison, we optimize the token passing protocol with the same assumptions than in FUZZY TOKEN. namely, that all nodes have a consistent view of the wireless channel. Thus, the passing of the token can be made implicitly, with zero delay after transmission and one-cycle delay after silence. Our centralized buffer protocol, by being a simple upper-bound of performance, is not implemented nor evaluated in this section.

Table 5.1: Characteristics of simulated protocols and applications.

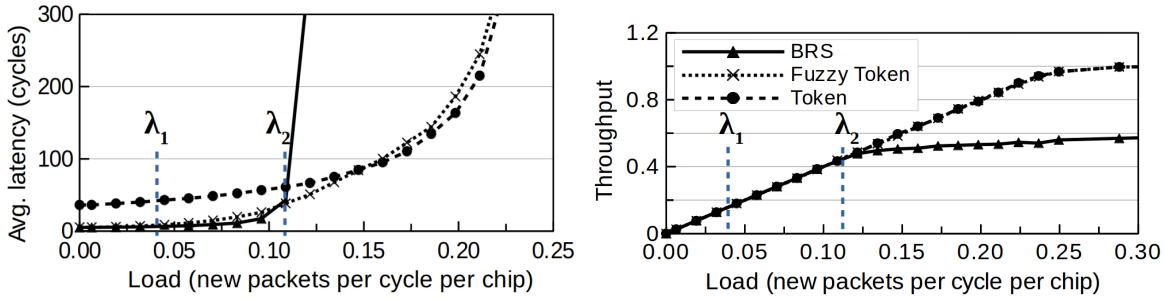
Wireless NoC Parameters	
Application	Synthetic traffic, H=0.5–0.9, σ=0.1–100
System	16–1024 cores, one antenna per core
Network	80-bit (4-cycle) packets (preamble: 20 bits, 1 cycle)
Link	BRS [26], Token, FUZZY TOKEN (NACK delay: 1 cycle)
Physical	OOK, 20 Gb/s

5.3.1 Evaluation

We start by comparing FUZZY TOKEN against the baseline protocols with synthetic traffic. By default, arrivals are Poisson and the injection process is equidistributed across all cores, to then evaluate hotspot and bursty traffic. Unless noted, the number of nodes is 64, which is later increased to higher numbers to study scalability. All results are included in Appendix D, and here we show a selection of plots that explain the performance and scalability of FUZZY TOKEN as compared to more traditional protocols.

The results of the first analysis with Poisson traffic are summarized in Figure 5.3. It is observed in Fig. 5.3(a) how FUZZY TOKEN is able to deliver the low latency of BRS at low loads and almost match the latency of token at high loads. In intermediate loads, FUZZY TOKEN can even outperform both BRS and token thanks to its intelligent management of contention. In Fig. 5.3(b), we see that FUZZY TOKEN achieves the same throughput than token, leaving BRS behind. In fact, at very high loads, FUZZY TOKEN ends up converging to regular token passing by design.

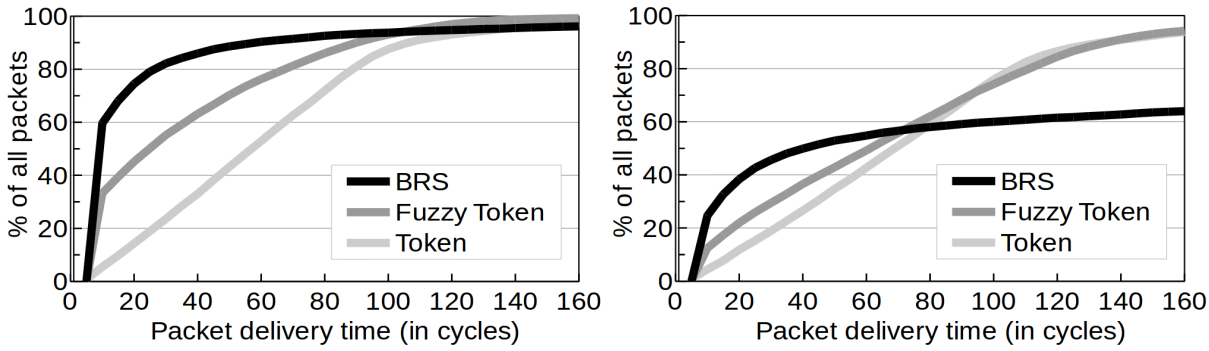
To complete the analysis, we plot the latency distribution of the three protocols in Figure 5.4 for two different loads. At moderate loads, Figure 5.4(a), it is observed that most transmissions with BRS take less than 30 cycles, with FUZZY TOKEN less than 60, and with Token less than 90. The caveat, however, is that due to collisions, 1.29% of the packets with BRS take more than 500 cycles, with a worst-case of ~3400 cycles. On the other hand, FUZZY TOKEN has a worst-case of ~330 cycles (the best among the three protocols), while still serving most of the packets within 60 cycles. At intermediate loads, Figure 5.4(b) shows how BRS still delivers many packets within the first 60 cycles, however now due to the higher amount of collisions, 28.9% of the packets take more than 500 cycles to be delivered, with a worst-case of ~110,000 cycles. On the other hand, FUZZY TOKEN also delivers most of the packets within the first 60 cycles, while providing a worst-case latency of ~390 cycles (again the best



(a) Latency (b) Throughput

Figure 5.3: Performance comparison for different MAC protocols over increasing load.

among the three protocols). The difference betweenBRS and FUZZY TOKEN, then, is shown in the re-distribution of the latency. While the latency distribution of FUZZY TOKEN gradually becomes the best between BRS and Token, BRS generates a very long tail that can take values up to several thousand cycles, which would clearly be a bottleneck for the execution of the application.



(a) Moderate load, $\lambda_1 = 0.045$. Tail: 1.29% BRS, 0% others.

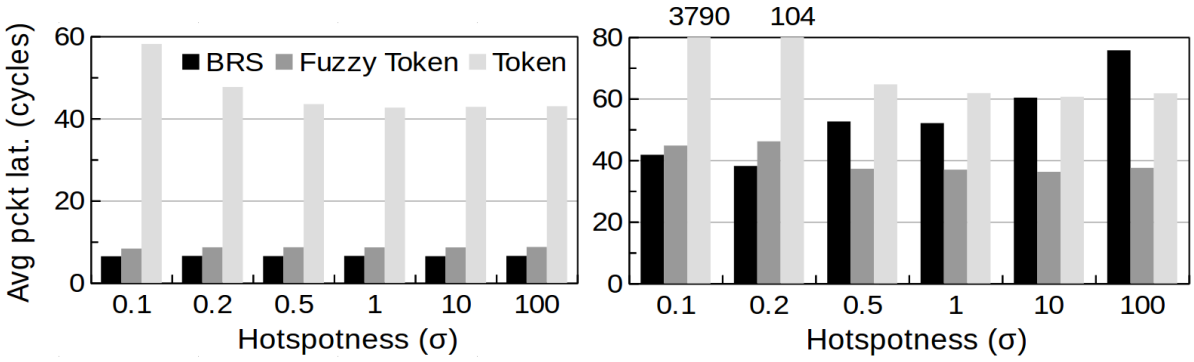
(b) Intermediate load, $\lambda_2 = 0.11$. Tail: 28.9% BRS, 0% others.

Figure 5.4: Cumulative distribution function (CDF) of the latency for the three protocols at different loads. Tail defined as delivery time over 500 cycles.

5.3.1.1 Hotspot Traffic

In hotspot workloads, a few processors inject most of the traffic. In that situation, it has been shown that contention-based protocols such as BRS outperform more rigid collision-free alternatives such as token [25]. To confirm the hypothesis that FUZZY TOKEN can deliver the best of the two types of protocols, we increase the spatial concentration of traffic via the σ parameter mentioned in Section 2.3.1.1 (i.e. low σ means hotspot traffic). The inter-arrival time is kept exponential.

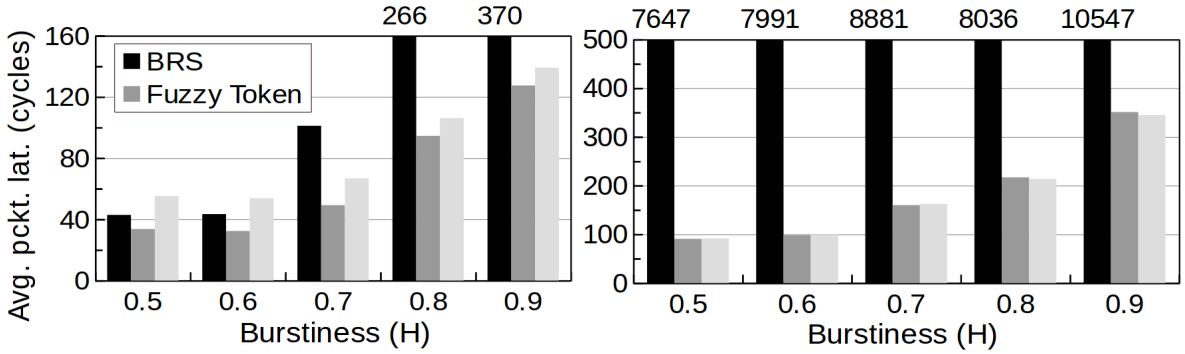
Figure 5.5 shows the results of the analysis at $\lambda_1 = 0.045$ and $\lambda_2 = 0.11$ packets/cycle. Fig. 5.5(a) illustrates the former case, where the load is moderate, contention is low, and an aggressive protocol is more appropriate. We confirm how FUZZY TOKEN is just a couple of cycles slower than BRS, which maintains a very low latency regardless of the value of σ. This is because nodes can transmit as soon as they generate the packets irrespective of their location. Token has a high latency, which worsens for low σ as the few transmitting nodes have to wait for their turn for



(a) Moderate load, $\lambda_1 = 0.045$ (b) Intermediate load, $\lambda_2 = 0.11$

Figure 5.5: Latency for hotspot traffic with different σ values. Low σ means that a few nodes inject most of the traffic.

each and every packet. Fig. 5.5(b) shows the results in the latter case, where contention starts becoming important (Fig. 5.3 shows how BRS starts to saturate at that load). In this case, BRS benefits from the decreasing number of contenders as traffic becomes more concentrated ($\sigma = 0.1$). Token passing performs poorly in such a situation because many clock cycles are wasted passing the token between a few greatly backlogged nodes. FUZZY TOKEN is capable of maintaining a low latency across all situations, outperforming the two other options by up to 100$\times$ with respect to token and up to 47% with respect to BRS. It is worth noting that similar tendencies are observed for loads beyond λ_2, but not shown in the sake of brevity.



(a) Moderate load, $\lambda_1 = 0.045$ (b) Intermediate load, $\lambda_2 = 0.11$

Figure 5.6: Latency for different H values. High H means more intense bursts.

5.3.1.2 Bursty Traffic

We repeat the same set of experiments now changing the temporal distribution of traffic, assuming $\sigma = 100$. Burstiness is modeled via the Hurst exponent H as depicted in Section 2.3.1.1, with higher H values leading to longer bursts and longer intervals between bursts.

Figures 5.6(a) and 5.6(b) illustrate the impact of burstiness on the MAC performance at moderate and intermediate loads, respectively. The first observation is that burstiness is detrimental for most mechanisms, especially for contention-based protocols. This is already patent at low loads: injections are infrequent, but bursty, and

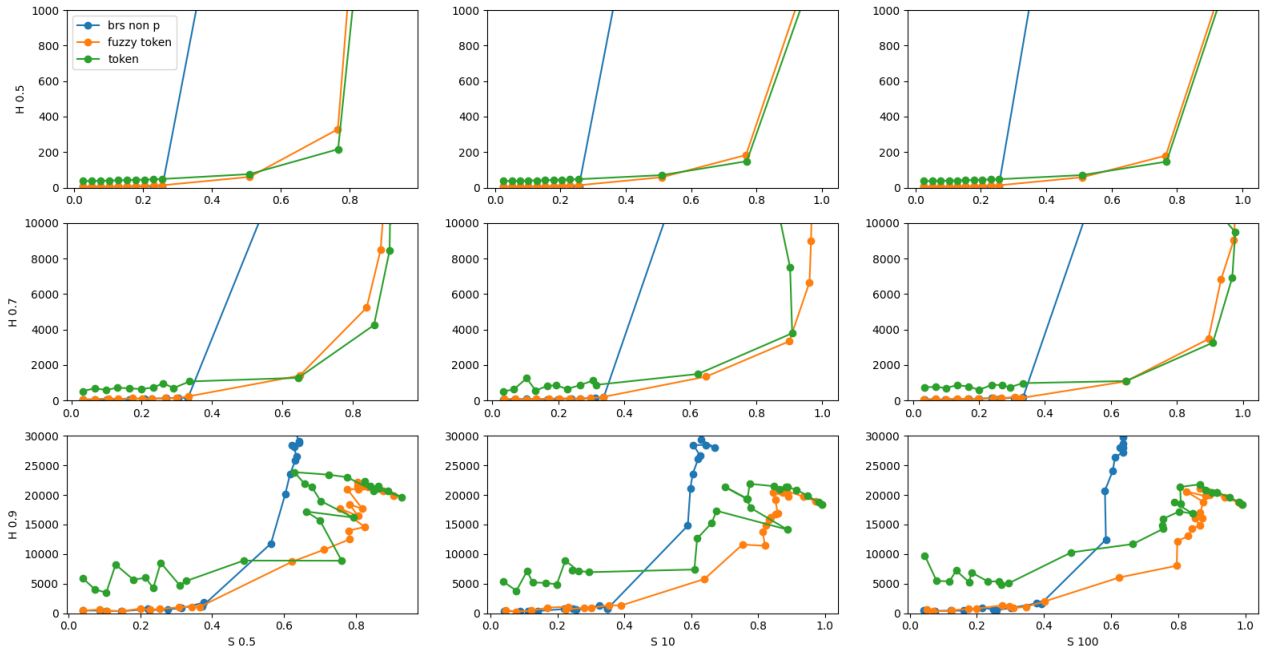


Figure 5.7: Latency-throughput characteristic for BRS, Token, and FUZZY TOKEN as functions of the Hurst coefficient for burstiness, and the σ parameter for hotspot behavior. for a system of 64 antennas.

hence collisions cannot be avoided. The latency of token passing also increases, but its collision-free nature better absorbs the bursts. Even so, FUZZY TOKEN is capable of achieving the best performance at all burstiness levels. This is because the first collisions occurring at the burst onset make FUZZY TOKEN to become pure token passing. As the burst is being served, the fuzzy area grows gradually and allows the last nodes of the burst to access the channel earlier. Figure 5.6(b) serves to confirm that increasing burstiness leads to early saturation, especially for BRS. As a result, FUZZY TOKEN avoids contention and converges to token passing to better absorb the intense bursty traffic.

5.3.1.3 Scalability

To study the impact of the number of participating stations in the communication on the performance of the different protocols, we present an overview of the simulation results for two system sizes, namely, $N = 64$ and $N = 256$. The plots are shown in Figures 5.7 and 5.8. Note that the rest of results from $N = 16$ up to $N = 1024$ are also included in Appendix D.

Essentially, we observe that the BRS protocol is largely unaffected by a change in the number of nodes because it is the overall load that determines the collision probability. Token, instead, is clearly affected negatively by the increase of nodes, as the token round trip time rises proportionally to the antenna count. FUZZY TOKEN continues to work well in all situations, although it tends become closer to token as the number of nodes grows. This is because the reduction of fuzzy area is fixed to an additive increase; at very high core counts, such a fuzzy area update is too slow to keep up with changes in the load.

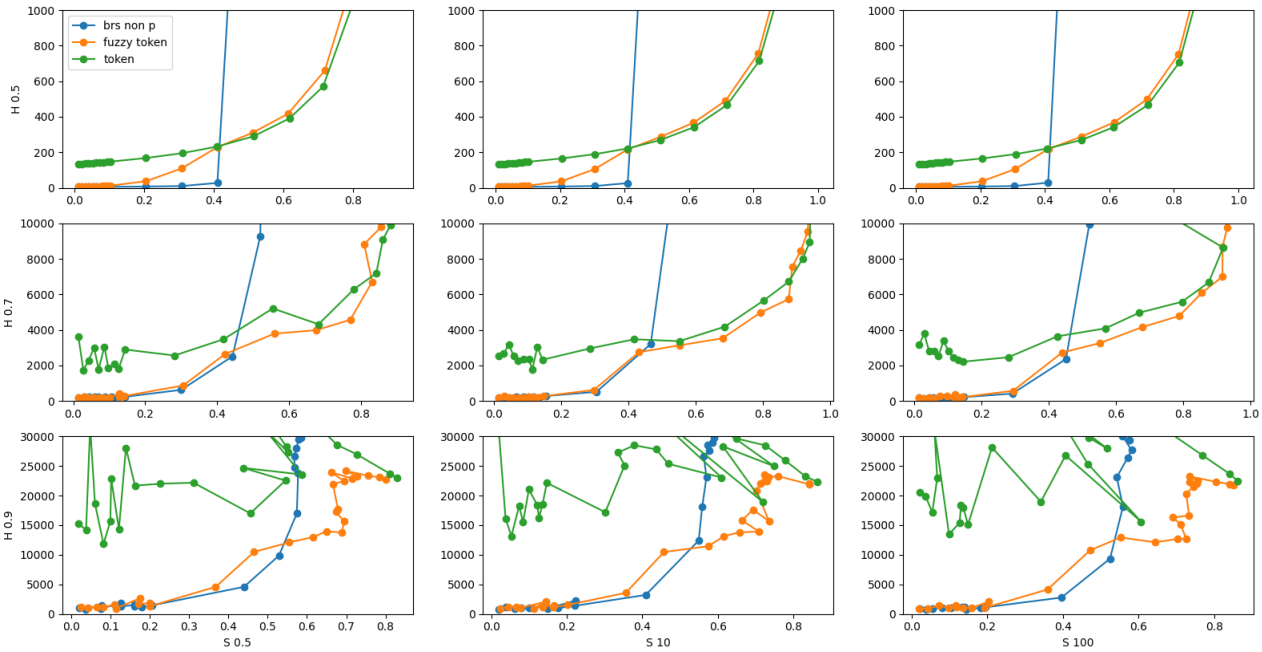


Figure 5.8: Latency-throughput characteristic for BRS, Token, and FUZZY TOKEN as functions of the Hurst coefficient for burstiness, and the σ parameter for hotspot behavior. for a system of 256 antennas.

5.3.2 Modeling

The methodology explained in Section 2.3 describes the modeling approach that is applied to the latency-throughput characteristic plots. In short, the results are truncated discarding those simulations whose latency is more than a given threshold, which we define as relative to the zero-load latency. This also marks the point of saturation load, λ_{sat}, Then, a quadratic fitting is applied to obtain the different parameters α, β, and τ_{ZL}. We note that the quadratic fitting has yielded an R^2 coefficient over 0.9 in most simulations, with some exceptions for high Hurst exponent values leading to very bursty traffic. Tables 5.2, 5.3, 5.4 list the different parameters for BRS, Token, and FUZZY TOKEN, respectively, assuming $N = 64$ and $N = 256$. The models for the rest of system sizes are given in Appendix D. It is finally worth noting that, for bursty traffic, token passing leads to extremely large latencies across the board. In those cases, fitting fails to capture the very low performance of the protocol; we mark them as *n/a* in the different mo

5.4 Resource Consumption Models

Unlike other hybrid protocols [19, 59], FUZZY TOKEN does not collect any utilization statistics. Because of this, FUZZY TOKEN only requires a small memory, placed in the transceiver, to store the *tokenID*, token ring order, *fuzzy area*, *periodMode*, and threshold values. In light of the simplicity of the FUZZY TOKEN algorithm, which is summarized in Fig. 5.1, and of the already small overheads reported by more complex protocols (e.g. less than 0.4 mW, 0.003 mm^2, 0.15 ns in [59]), we argue that the area

Table 5.2: BRS model parameters for different workloads and $N = \{64, 256\}$.

BRS, 64 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.35	0.28	0.28
	α	1.1	0.5	0.27
	β	40	42.6	43.5
	τ_{ZL}	5.1	5.1	5.1
0.6	λ_{sat}	0.35	0.2	0.2
	α	-153.3	-39.1	-6.4
	β	476.7	711.6	622
	τ_{ZL}	34	30	24.3
0.7	λ_{sat}	0.35	0.2	0.2
	α	176.2	-360	-93.6
	β	476.7	1892.8	1274.6
	τ_{ZL}	34	70.4	51.8
0.8	λ_{sat}	0.3	0.15	0.18
	α	-1025.7	-584.7	-332.4
	β	17174	5158.7	6773.8
	τ_{ZL}	120.5	196	181.6
0.9	λ_{sat}	0.2	0.1	0.05
	α	1060.5	1660	172.5
	β	-11620.6	1477.1	5516.2
	τ_{ZL}	421.8	182.5	356.2

BRS, 256 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.42	0.42	0.41
	α	4	4.5	4.6
	β	25	21.1	19
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.4	0.3	0.25
	α	124.7	-1234	-155.1
	β	-447	9165	2055
	τ_{ZL}	131	136	100
0.7	λ_{sat}	0.2	0.15	0.17
	α	232.5	-145	-365.9
	β	-286	1727	3055
	τ_{ZL}	202	205	182
0.8	λ_{sat}	0.1	0.01	0.1
	α	-580.7	-1333	6347
	β	8543	16276	-37356
	τ_{ZL}	656	630	435
0.9	λ_{sat}	0.01	0.01	0.05
	α	7511.3	-7552	4483
	β	-13210.6	54802	-16849
	τ_{ZL}	761	1146	724

Table 5.3: Token model parameters for different system sizes and workloads.

Token, 64 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.4	0.4	0.4
	α	32	31	30
	β	62	52	53
	τ_{ZL}	35	35	35
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Token, 256 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.01	0.01	0.01
	α	132	130	126
	β	200	177	196
	τ_{ZL}	131	130	131
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Table 5.4: Fuzzy token parameters for different system sizes and workloads.

Fuzzy Token, 64 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.5	0.5	0.5
	α	-1	-2	-2
	β	120	123	123
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.35	0.2	0.2
	α	92	-56	-128
	β	227	819	836
	τ_{ZL}	29	32	39
0.7	λ_{sat}	0.35	0.2	0.2
	α	-348	-388	-63
	β	1925	1870	1205
	τ_{ZL}	89	93	62
0.8	λ_{sat}	0.3	0.15	0.18
	α	3268	-188	3916
	β	-21099	5028	-12427
	τ_{ZL}	161	240	68
0.9	λ_{sat}	0.2	0.1	0.05
	α	6597	2184	-3506
	β	-4100	1374	22912
	τ_{ZL}	252	263	561

Fuzzy Token, 256 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.25	0.25	0.25
	α	6	3.77	10
	β	571	593	520
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.2	0.2	0.25
	α	1006	233	-26
	β	-3936	3162	2823
	τ_{ZL}	81	79	98
0.7	λ_{sat}	0.2	0.15	0.17
	α	-1468	-821	382
	β	1513	6810	7217
	τ_{ZL}	218	215	148
0.8	λ_{sat}	0.1	0.1	0.1
	α	6184	3142	6006
	β	-14870	-8602	-21973
	τ_{ZL}	349	490	360
0.9	λ_{sat}	0.01	0.01	0.05
	α	-11806	-1620	13381
	β	92880	31607	-65840
	τ_{ZL}	1419	1026	567

and energy overheads of FUZZY TOKEN's circuitry are negligible when compared to those of the transceiver itself [7, 60, 61].

Another source of overhead are collisions. As Figure 5.9 illustrates, FUZZY TOKEN achieves high performance with a very moderate energy overhead over the regular token passing (less than 12%) due to collisions. Energy-wise, there are two additional points that are worth remarking. First, the energy consumption in regular token passing is not affected by the load. This is because of the implicit passing of the token, which does not consume energy. This, however, comes at the cost of high latency at low loads. The second point is that the energy of BRS increases with the load because of the appearance of collisions, but then decreases. This effect is due to the finite population of the chip scenario: at very high loads, the backoff reaches huge values and reduces the probability of collisions at the expense of unacceptable latency. We see that FUZZY TOKEN achieves the low-load latency of BRS while avoiding its high energy expense at higher loads.

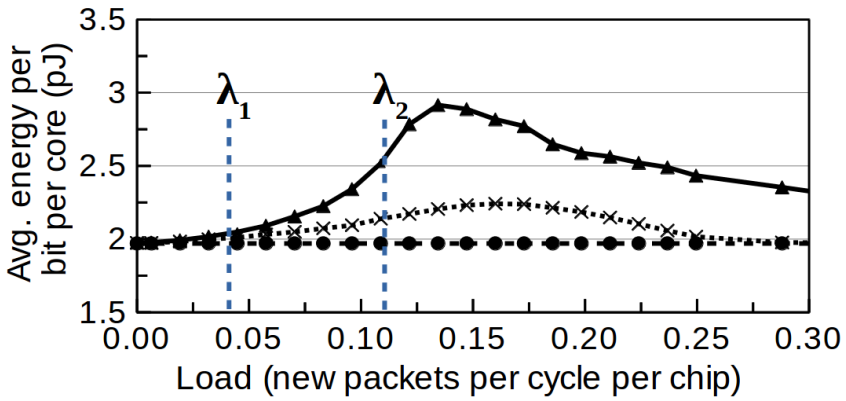


Figure 5.9: Energy consumption comparison for different MAC protocols over increasing load.

6. Discussion and Concluding Remarks

This deliverable has addressed the need for accurate area, power, and performance models for their incorporation in system design space exploration frameworks. In particular, we modeled (i) the wireless channel within packages by means of fitting of a comprehensive simulation campaign performed within WiPLASH; (ii) the area and power of the wireless transceivers, by building our own models via a bottom-up approach and component modeling by a combination of literature analysis and circuit modeling; and (iii) the latency and throughput of a wireless link based on our own simulation of three possible link-layer protocols, and fitting of the obtained data. These models, represented in a few parameters and tabulated for a wide range of simulations, are new to the wireless network-on-chip community and allow architects to estimate, with very low computational effort, the performance and resource consumption that a wireless network will entail.

In the case of wireless channel modeling, we observed that channel losses below 40 dB are possible with interesting exponent values around $\gamma \approx 1$, and also coherence bandwidths beyond 20 GHz in some specific design points. At the physical layer of design, we assessed the dependence of the area and power of the entire transceiver chain, including the SerDes circuits, as functions of frequency, channel losses, and modulation rate. While OOK promises low area (100 Gb/s/mm^2) and low power (2 pJ/bit) compatible with the channel losses evaluated here, we estimate that the limited coherence bandwidth of the channel will force to either use less efficient yet higher-order modulation or to consider using multiple streams of OOK data. Finally, we demonstrated that our proposed MAC protocol gathers the best of CSMA-like and token-like protocols, and created a look-up table model for its performance across system sizes and workload characteristics in terms of spatiotemporal distribution.

In future work, we plan to iterate on current models and incorporate new data as they become available thanks to new simulations of experimental results. Examples include wireless channel measurements with RWTH and UoS, transceiver data from new tapeouts from RWTH, specific models for certain components of the transceiver, or the simulation of new multi-channel MAC protocols that will embody the tunable characteristic of the graphene antennas modeled in the project.

A. Background

Wireless chip-scale communications are among the different candidates for interconnecting processing elements and memory within complex computing packages. Specifically, the wireless paradigm has been proposed as a complement to the wired interconnects to (i) reduce the latency in communication between distant processors, possibly across chip boundaries, (ii) alleviate existing bandwidth bottlenecks caused by I/O pin limitations, and (iii) establish global and reconfigurable links. These are possible thanks to the inherent low latency, broadcast capability, and lack of path infrastructure of the wireless technology.

As illustrated in Figure 1.1, wireless chip-scale communications broadly refer to the implementation of intra-chip or inter-chip links with integrated antennas. In general terms, any of the components within a multiprocessor architecture (e.g. CPUs, GPUs, accelerators, memory) may be provided with a wireless transceiver that would serialize, modulate and radiate outgoing information. Before that, protocols at the network and link layer decide whether information needs to be wirelessly transmitted; if so, these protocols also determine when the transmission takes place and through which channel. EM waves propagate through the processor package until reaching the intended destinations, where they are demodulated and deserialized.

Signals radiated at the transmitting end suffer losses and dispersion, which affect the ability of the receiver to correctly demodulate the transmitted information. Moreover, two overlapping transmissions through the same channel would create a *collision* and be lost. In this context, the RF transceiver, together with the protocols at the physical layer and link layers of design, are developed to combat these detrimental effects and ensure that the information is delivered without errors. On the one hand, the transceiver and physical layer protocols are designed so that the data is transmitted at high speed and with enough power reach the receivers with enough Signal-to-Noise Ratio (SNR) allowing to meet the BER requirement of the communications scenario. This generally means applying a given modulation to the data and amplifying the modulated signals before radiating them, in a process that takes considerable power and silicon area. On the other hand, the link layer protocol manages access to the shared medium in an attempt to deliver data with the lowest latency and number of collisions. Such a process may cause delays due to the scheduling of transmissions or recovery from collisions that could not be avoided, thereby clearly affecting performance.

For all the reasons above, understanding the different aspects between the channel and the link layer of protocol design are crucial to model the performance and resource consumption of wireless chip-scale communications. In this chapter, we provide background on the chip-scale environment in an attempt to gain insight on the particularities of the scenario. First, in Section A.1, we review the salient characteristics of the wireless chip-scale communications context, from the physical landscape to the design drivers or the types of traffic to be served. Then, in Sections A.3 and A.4, we

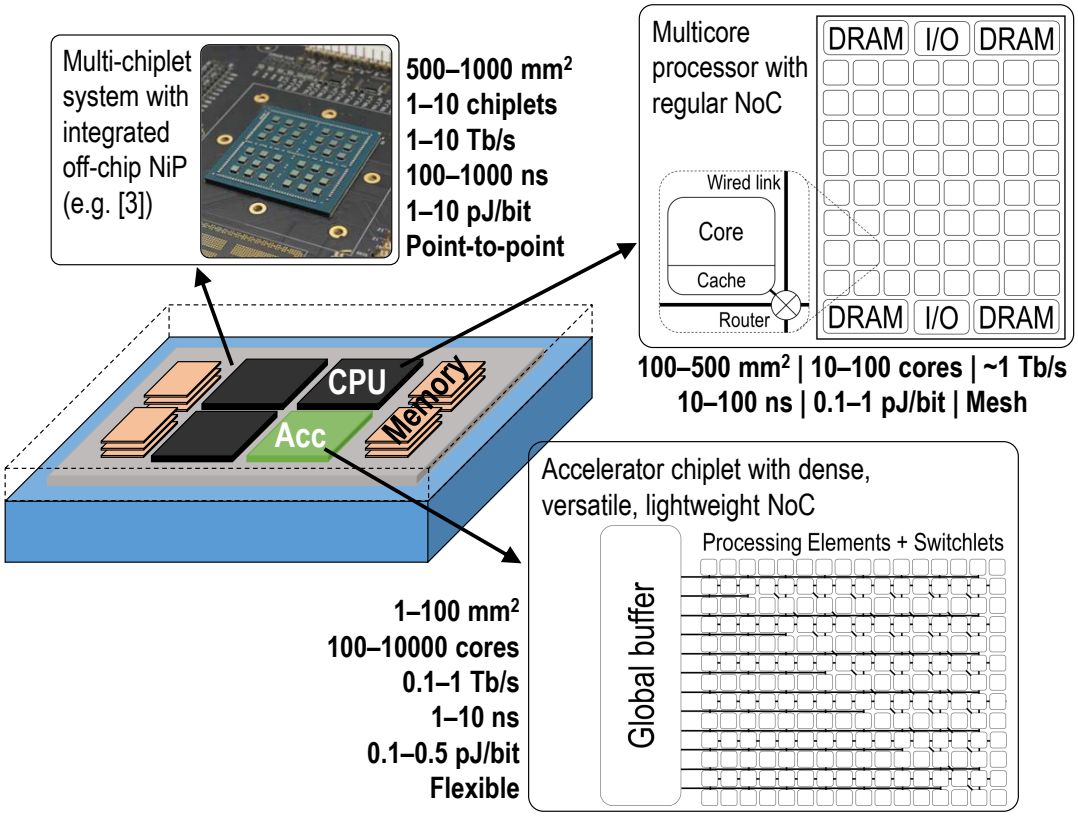


Figure A.1: The chip-scale communication landscape in the heterogeneous chiplet era: Network-in-Package (NiP) to interconnect chiplets, Network-on-Chip (NoC) for multicore processors, and dense fabrics for accelerators. For the three scenarios, we list popular system sizes, number of nodes, bisection bandwidth, latency, energy per transmitted bit, and topology.

discuss the fundamentals and related work in physical layer and link layer of design, respectively. We refer the reader to *Deliverable D3.1: Wireless Channel Models* [1] for a similar analysis of context and fundamentals for channel modeling for wireless in-package communications.

A.1 Context Analysis

The wireless chip-scale scenario has a unique blend of requirements and constraints that impact on the design of the protocol stack and the desirable performance and cost. We next summarize them in four main points, which relate to Figure A.1: high performance, resource awareness, monolithic system, and workload characteristics.

A.1.1 High Performance

Computing systems demand ultra-fast and reliable communications at the chip scale mainly because the communication latency slows down the computation and minor errors may corrupt an entire computation. Generally, architectures may try to be resilient to communication latency and errors; the former via latency hiding techniques through the computing and memory pipeline, which come at the expense of indirection

and complex hierarchies [62], and the latter via the use of approximate computing [63], which is far from widespread and cannot be applied to all types of data and application domains.

Most WNoC proposals consider wireless in the order of 10–100 Gb/s to achieve system-wide latencies around or below 10 ns, while it is generally accepted that the error rate should be comparable to that of wired interconnects, i.e. between 10^{-12} and 10^{-15} [64]. This has several implications in the design of the wireless communication. First, the channel needs to support such a high bandwidth, either via a single broadband channel or multiple sub-channels. Second, the physical layer of design needs to use a modulation that either has a very high spectral efficiency in bit/s/Hz, which generally requires a complex transceiver and challenging SNR levels to be decoded correctly; or a low order modulation with a very high modulation rate. Third, the MAC protocol needs to ensure high throughput with low delays, not only in average but also in the worst case. Note, however, that the aggregate bandwidth of the system is typically much larger to accommodate many simultaneous unicast transfers, which are responsibility of the wired interconnect fabric.

A.1.2 Resource Awareness

Nodes in wireless networks are typically mobile and hence have a limited battery, oftentimes leading to energy-constrained communication. In chip environments, the energy supply is generally guaranteed, yet limited by heat dissipation constraints which manifest through the maximum TDP. In the current era of computer architecture, power has actually become a driver of multiprocessor design, suggesting the use of power-gating techniques to increase the overall efficiency and meet TDP constraints [65]. Similarly, chip real state is a precious resource due to cost reasons relative to fabrication and yield, i.e. larger chips have a higher probability of fabrication faults.

From the perspective of physical layer of design, this generally implies that simple and low-power transceivers that support only low-order modulations are preferred because they do not require bulky or power-hungry components [36]. Similarly, MAC protocols need to be simple and minimize collisions to reduce the area and power overheads. From the perspective of the wireless channel, resource awareness forces architects to minimize path loss while increasing the frequency and looking for wide spectral bandwidths to accommodate the high requirements of data rate.

A.1.3 Monolithic System

A multicore processor is basically a monolithic system from the designers' point of view and often a proprietary solution. The design team has a certain control over the architecture and the physical landscape of the system. Hence, we argue that the propagation of EM waves takes place in a confined space, which moreover is fixed and known beforehand [66]. This represents one of the main uniquenesses of the WNoC scenario, since nodes in other wireless networks generally move within a propagation environment that can also be dynamic. Moreover, in traditional wireless systems, the network stack and the applications are designed and developed by different teams.

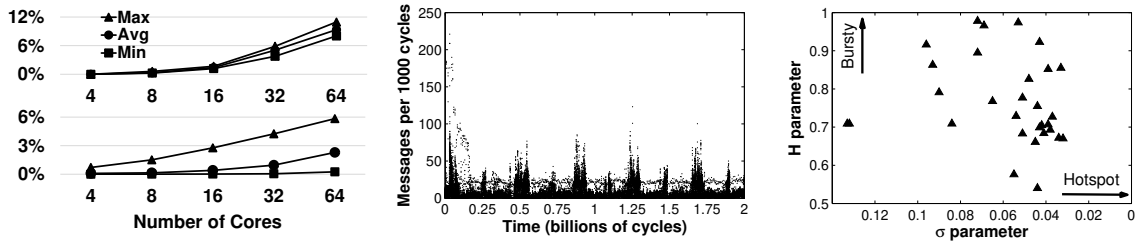
This aspect has multiple implications in the design of wireless in-package networks. For instance, the chip-scale channels become quasi-deterministic at the data-link layer as the physical landscape is static and known *a priori* [66]. At the physical layer,

such a static property can be exploited to streamline the performance of encoders and decoders, which now do not have to depend on signal statistics or, for that matter, need to combat a much narrower variance in the signal characteristics [56]. At the MAC layer, the monolithic aspect is positive in that protocols can be heavily optimized based on the prior knowledge on the applications, or even by entering the design loop of the architecture as discussed in depth in [29]. This helps combat the communication workloads of multiprocessors, which as we will see next, are challenging to serve due to their burstiness and variability.

A.1.4 Workload Characteristics

The communication workloads traditionally shown by single-chip multiprocessors are challenging to serve in most networks and especially in wireless ones. The main characteristics of such workloads are heterogeneity, variability, spatial hotspot behavior, and temporal bursty behavior. In more detail:

- **Heterogeneity:** Although architectures are generally designed trying to avoid expensive communication transactions, manycore processors face the challenge of having to support heterogeneous traffic profiles. Traditionally, local and unicast communications have dominated, but it has been shown that the global and multicast flows targeted by wireless communication can become significant in manycore processors [15, 28]. Figure A.2(a) illustrates this by plotting the percentage of long-range and multicast traffic as a function of the number of cores in multithreaded benchmark suites [67]. The rise of the chiplet paradigm might exacerbate this aspect, as specialization may lead to different chiplets generating completely different intra-/inter-chiplet traffic patterns.
- **Variability:** The existence of a wide range of programming models and application domains may cause large changes in terms of communication demands from one application to another. Moreover, the particular chiplet combination in a heterogeneous architecture influence in such a variability, as different applications may require to use a particular accelerator chiplet intensely while others may not. Within each particular application, phase behavior also leads to wild variations on the traffic characteristics over time [68]. Such a behavior is exemplified in Figure A.2(b), which clearly shows how the application *fluidanimate* alternates between communication-intensive and computation-intensive phases.
- **Spatial hotspot behavior:** Soteriou *et al.* revealed that most applications generate traffic unevenly across all the utilized processing elements [28]. Similarly, the characterization performed in [67] confirmed that multicast flows also follow a strong hotspot distribution, meaning that a few selected processing elements inject most of the traffic. Figure A.2(c) reproduces some of these results by plotting the standard deviation $\sigma \in [0, \infty)$ of the injection distribution, where small values represent hotspot traffic.
- **Temporal bursty behavior:** Soteriou *et al.* also demonstrated that, as it occurs in most networks, traffic in the NoC domain is self-similar. The consequence of this fact is that packets are injected in bursts followed by relatively long silences. The Hurst exponent $H \in [0.5, 1]$ evaluates this behavior, where $H = 0.5$ corresponds to memoryless traffic and large values indicate the presence of self-similarity. As shown in Figure A.2(c) for traffic in a single-chip multiprocessor,



(a) Percentage of long-range traffic (3 hops or more, top chart) and multicast traffic (bottom chart) for the SPLASH2 suite over MESI coherence [67].

(b) Phase behavior exhibited by the traffic generated by a 64-threaded instance of the *fluidanimate* application over MESI coherence [67].

(c) Spatiotemporal characteristics of the applications analyzed in [28]. Each mark represents a specific application.

Figure A.2: Workload characterization of different multiprocessor architectures and applications exhibiting (a) increasing heterogeneity, (b) intra-application variability, and (c) inter-application variability with bursty and hotspot traffic.

traffic tends to be busty ($H > 0.7$). This has been also confirmed for multicast traffic in single-chip architectures [67] and is likely to continue to hold in heterogeneous chiplet architectures.

The characteristics of traffic are important to understand the requirements placed on the link layer of design and, more specifically, on the MAC protocols. There are several conclusions to take away from this analysis. For instance, the heterogeneity and variability of traffic suggests that the MAC protocol should be reconfigurable to adapt to large-scale changes, i.e. between and within applications and taking into consideration the added heterogeneity of the chiplet paradigm, with a reasonable cost. Beyond that, the hotspot/bursty characteristics of traffic are generally detrimental to performance and call for flexible solutions that can provide fast and fine-grained adaptivity. This implies, then, that slow reconfigurability will not be enough to cater to the ever-changing communication needs of new architectures and systems.

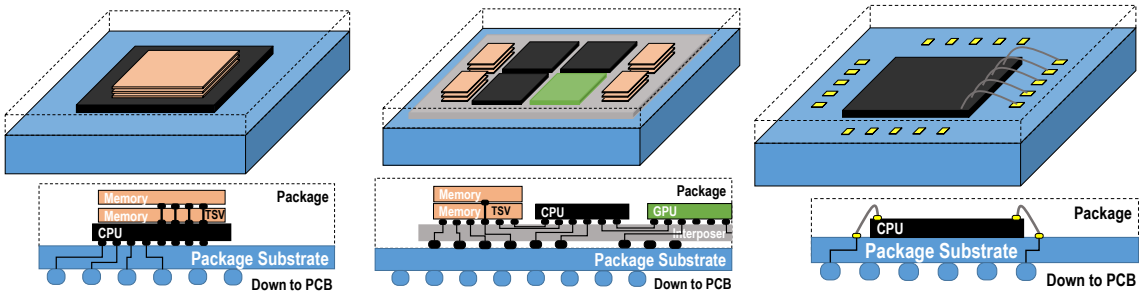
A.2 Wireless Channel

The first aspect to consider to evaluate the potential performance and efficiency of a wireless link is the wireless channel, which essentially determines the attenuation and dispersion that signals suffer during propagation. In conventional wireless scenarios where Line of Sight (LoS) propagation through free space can be generally assumed, losses are calculated as

$$FSPL = \left(\frac{4\pi d}{\lambda}\right)^2, \tag{A.1}$$

where d is the transmission distance and λ is the transmission wavelength. In realistic scenarios, where LoS and free space propagation do not hold anymore, models become more complex and require numerical simulations or experimental measurements to be developed [69–72]. This is the case of on-chip communication, as we discussed in depth in Deliverable D3.1 [1]. A typical fitting model for path loss PL in these scenarios is given by

$$PL = PL_0 + 10\gamma \log_{10} \frac{d}{d_0} + X_g \tag{A.2}$$



(a) Flip-chip (b) Silicon interposer (c) Wirebond

Figure A.3: Different flavours of computing packages capable of hosting multiple chips.

where PL_0 is the path loss at a distance d_0, d is the transmission distance, γ is the path loss exponent, and X_g is a random variable related to fading. Importantly, γ models how path loss scales with distance, which is 2 in freespace propagation where spreading losses dominate, lower than 2 in enclosed and waveguided low-loss environments, and higher than 2 in lossy and environments with many obstacles between transmitter and receiver.

Wireless chip-scale communication generally occurs within the confinements of a computing package. As discussed in depth in Deliverable D3.1 [1], computing packages can take different forms as shown in Figure A.3. Flip-chip packages, wherein the chip(s) are flipped over and connected to the Printed Circuit Board (PCB) board or organic substrate through solder bumps, are currently widespread. Then, the chip is surrounded by (i) a metallic heat sink contacted by a heat spreader and (ii) the package carrier, with several metal layers on top the PCB. An alternative to this are interposer-based packages, where a large silicon die is placed between the organic substrate and the chips, to provide finer-grained connections among the chips. Finally, wire bonding is the traditional packaging option whereby chips are placed directly on top of the package substrate and connected through bond wires between the PCB contacts and the chip pads, which are now on top.

Figure A.4 graphically depicts the process of wave propagation within computing packages. In the *intra-chip region*, waves through several layers of the chip, including the dielectric. In the *inter-chip region*, waves that have left the chip travel through the inter-chip space until they reach the boundaries of another chip or the package limits. Beyond the spreading losses, EM waves within package can suffer from reflections, refraction, diffraction and absorption. More specifically, *reflections* will appear both when a wave reaches the interface between two materials, which may happen often in this enclosed and highly integrated scenarios. Depending on the smoothness of the interface, measured relative to the wavelength, reflections can be specular or scattered. In addition, when transitioning from a medium to another, *refraction* of the EM wave will occur again depending on the change in the refraction index. *Diffraction* or bending of the wave around the (sharp) edges of chips can also occur due to the parallelogram-like form of different components. All these components may lead to relatively complex channel responses.

As discussed in Deliverable D3.1 [1], where we surveyed the state of the art on channel modeling for wireless chip-scale communications, the existing literature focuses on the 60–100 GHz band and does not model the computing package realistically. Without channel models that include the package at scale to higher frequencies,

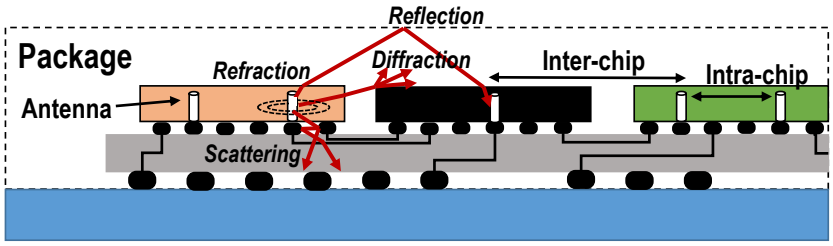


Figure A.4: Schematic representation of wave propagation in an interposer system with flip-chip package excited with vertical monopole antennas, distinguishing between intra- and inter-chip regions, and exemplifying different propagation phenomena.

accurate modeling of the performance and efficiency of wireless chip-scale communications is challenging. Aware of this gap, we have performed an extensive simulation campaign to characterize the wireless channel within realistic computing packages, pushing the carrier frequency upto 240 GHz in frequency domain analysis and to 1 THz band in the time domain analysis, and including wirebond, flip-chip, and interposer packages coherently with the WIPLASH vision.

In summary, our evaluations showed that (i) flip-chip and interposers are preferable over wirebond, that (ii) path loss of 30–40 dB and delay spreads below 0.1 ns can be achieved over distances of a few centimeters without cumbersome optimization processes, and (iii) that thinning down the silicon die is the most impactful design decision, which can be combined with other optimizations such as using epoxy resin instead of vacuum as filling material in the package. However, we also need to be aware that increasing the distance and the frequency of operation will increase both the path loss and delay spread. All these results form the base of our channel modeling later in Chapter 3.

A.3 Physical Layer

The Physical Layer (PHY) defines how bits are transmitted over the wireless links and, thus, plays a fundamental role in determining the requirements of the associated transceivers. These, in turn, largely determine the area and power consumption of the wireless network. In a WNoC, the PHY module will basically serialize processor messages, modulate the resulting bits at a given frequency much higher than the processor clock, and deliver the modulated signal to the antenna. The inverse operation is performed at reception.

One of the first decisions at PHY is the frequency of operation, which largely determines the dimensions of the on-chip antenna and the available bandwidth. On the one hand, a resonant antenna has a length and width commensurate to the wavelength λ within the medium where the antenna is placed. Note that, in the case of graphene antennas, the length is commensurate to the SPP wavelength $\lambda_{SPP} = \lambda/K$ where K is the compression factor of the SPP wave in the antenna. In any case, the size is inversely proportional to the operation frequency f_c [73]. On the other hand, in order to fulfill the bandwidth requirements B at such resonance frequency, a conventional resonant antenna must yield a quality factor of $Q \approx \frac{f_c}{B}$. A high quality factor implies a sharper resonance, which leads to a better efficiency but a lower bandwidth overall.

Also, maintaining a certain Q at higher frequencies leads to an equally higher bandwidth, reason for which it is generally held that it is easier to obtain high bandwidths at high frequencies.

Another crucial design decision at PHY is the modulation as it defines the *spectral efficiency* S_E of the system, this is, how many bps are transmitted for each Hz of frequency bandwidth. Thus, the transmission rate R is:

$$R = B \cdot S_E, \tag{A.3}$$

where B is the frequency bandwidth of the link. Hence, transmission rates of a system can be scaled by either increasing B or using a modulation with higher S_E. These have different consequences that are difficult to disentangle and that we will try to model through the deliverable. In principle, higher B needs to be supported by the antenna, channel, and transceiver; such a decision may require shifting to higher frequencies so that the components can accommodate the bandwidth easily. Technology evolution pushes the achievable frequencies, f_T and f_{MAX}, and their adoption may help fulfilling the B requirement. On the other hand, changing the modulation to increase S_E may not only lead to an increase of the SNR needed to achieve a certain BER, but also demand completely different transceiver architectures and the use of complex circuits, leading to non-trivial changes in the area and power. One can generally distinguish between coherent and non-coherent receivers depending on whether they are able to detect the phase of the signal. Coherent receivers are typically more complex and power-hungry due to the need for sophisticated components such as the Phase-Locked Loop (PLL), but admit a wider variety of modulations that are typically robust against noise. To exemplify this, consider that wireless signals are received with a given SNR. The BER at the receiver will be:

$$BER \propto \mathcal{F}\left(\frac{1}{SNR}\right), \tag{A.4}$$

where $\mathcal{F}$ is a function specific to each modulation and that does not always have a closed form. As shown in Figure A.5, coherent modulations like Phase-Shift Keying (PSK) or QAM will generally require a lower SNR to reach an objective BER than non-coherent modulations like OOK or Pulse Amplitude Modulation (PAM). Fixing the modulation family, one can increase the spectral efficiency at the expense of requiring a larger SNR to comply with a given error rate. In the end, however, the power consumption will be determined by the transceiver components.

Given that there is a fixed statistical relationship between the power received and the BER, one can evaluate the power that needs to be radiated at the transmitter to ensure a given signal strength at the receiver and, thus, to guarantee a given error rate. For that, we need to perform a *link budget* which takes as inputs the antenna gain as well as the losses between transmitter and receiver (possibly as a function of the distance) at a given transmission frequency to evaluate the total attenuation introduced by the link. In short, the SNR can be expressed as [74]

$$SNR = \frac{P_t \cdot G_t \cdot G_r \cdot B}{N_0 \cdot L_{RX} \cdot PL \cdot R} \tag{A.5}$$

where P_t is the power at the output of the transmitter, G_t and G_r are the gains of the transmitting and receiving antennas, L_{RX} is the loss of the receiver, and $N_0 = k \cdot T_0 \cdot B \cdot F$

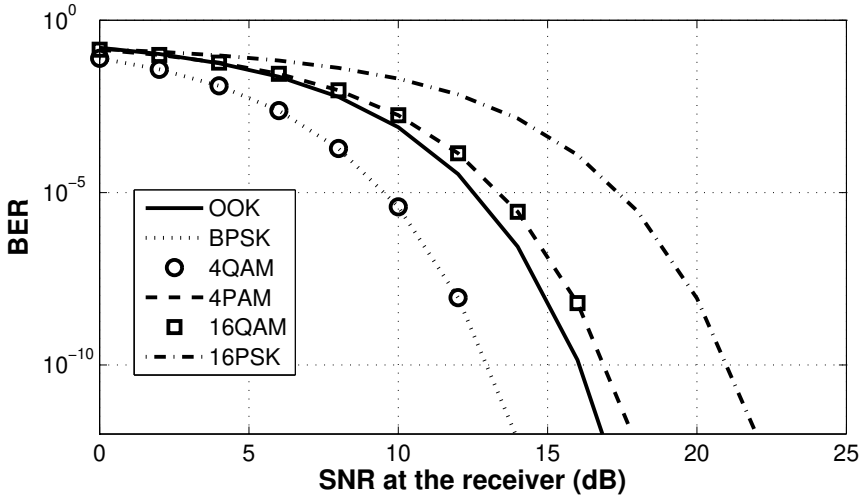


Figure A.5: Theoretical BER as a function of the SNR for different modulations.

is the input noise power that depends on the Boltzmann constant k, the receiver noise temperature T and the noise figure of the receiver F. Other known terms discussed above are B for bandwidth, R for transmission rate, and PL for path loss. Therefore, the link budget process requires a model of the channel and its associated losses, as discussed above. An alternative way of expressing this relation is also described in [74]

$$R = \frac{A}{SNR}\frac{P_t}{d^{\gamma}} \tag{A.6}$$

where R is the achievable rate, when P_t is transmitted over a channel of exponent γ as defined in Equation (A.2), with a modulation that requires SNR at the receiver, using a transceiver with $A = \frac{G_t \cdot G_r \cdot d_0^{\gamma}}{k \cdot T_0 \cdot F \cdot L_{RX} \cdot L_p(d_0)}$ being a constant related to the transceiver at hand and the reference path loss at distance d_0.

The existing work on PHY design for wireless chip-scale communications has generally taken an approach driven by high frequencies and simple modulations to satisfy the very high transmission speed and high efficiency requirements of the scenario [36, 60, 74]. Having 10–100 Gb/s as a reasonable target and silicon real estate as a precious resource, the use of frequencies beyond 60 GHz is proposed. Due to the relatively immature state of THz technology, high-order modulations or techniques requiring significant signal processing are discouraged. Instead, most proposals advocate for simple modulations such as OOK, which can be achieved by simply connecting the stream of bits to the circuit that generates the carrier wave, together with non-coherent (i.e. amplitude) detection. This combination avoids the use of bulky and power-hungry circuits such as PLLs and even eliminates the need for a Analog-to-Digital Converters (ADCs). It has been thus shown that OOK can be 1.5X and 2.5X more energy efficient than BPSK and QPSK in on-chip environments [7]. The downside of using low-order modulations is that, because the wireless channel may produce significant dispersion as shown in Deliverable D3.1 [1], a single carrier cannot be employed. Multi-carrier modulations may additionally relax the serialization/deserialization requirements of the wireless link.

The main issue concerning the related work in PHY design in wireless chip-scale communications is that a full working link has not been prototyped yet. This makes the

transceiver power and area models dependent on link budget formulations which are not necessarily accurate, as they in turn depend on the channel, for which adequate models have been largely missing as discussed in prior sections. Moreover, it is difficult to find area and power models that cover different situations in terms of frequency, channel losses, or technology, simply because transceivers are generally designed having a well-defined spectral mask and specifications to achieve. We will attempt to address these points in Chapter 4 using channel models from Chapter 3.

A.4 Link Layer

At the link layer of design, the MAC sub-layer implements mechanisms to ensure that all nodes can access to the medium in a reliable manner. As we will see, this plays a decisive role in determining the performance of any network as two simultaneous accesses to the same channel will fail and result into a waste of resources.

Two metrics are generally employed to evaluate the performance of a given MAC protocol. On the one hand, the *latency* of the protocol measures the time spent by a packet in the MAC queue, this is, from the instant the message is queued until the transmission is successful. Adding all factors, the transmission latency τ_L of a packet of length L through a PHY with data rate R, as calculated with Eq. (A.3), is:

$$\tau_L = \frac{d}{c_0} + \frac{L}{R} + \tau_{MAC}, \tag{A.7}$$

where the term $\frac{d}{c_0}$ is the propagation latency, $\frac{L}{R}$ is the transmission latency, and τ_{MAC} is the MAC delay, including retransmissions, timeouts and acknowledgments. In chip environments, it generally holds that $\frac{L}{R} + \tau_{MAC} \gg \frac{d}{c_0}$. A complete evaluation in terms of latency generally calculates τ_{MAC} at low loads (i.e. zero-load latency) and high loads, near saturation, both in average and standard deviation. A large variance may mean that the protocol is not fair, which is extremely harmful in this scenario.

On the other hand, an equally important metric is the MAC throughput M, which is calculated as average rate of correct transmissions, and its normalized version μ, which are evaluated as

$$\begin{aligned} M &= \frac{\sum L_i}{T}, \\ \mu &= \frac{\sum L_i}{T \cdot R}, \end{aligned} \tag{A.8}$$

this is, as the sum of the lengths of the successfully transmitted packets, divided by the total elapsed time T. M is calculated in bits per second, and its normalized version $\mu \leq 1$ is unitless when divided by the data rate of the channel R. Throughput is typically reported at saturation, this is, when latency surpasses a certain threshold.

Related works on MAC for WNoC can be divided into various groups, namely:

- **Multiplexing:** First MAC protocols proposed for the WNoC paradigm used time, frequency, or code orthogonal channels [23, 75]. These techniques are free of wasteful collisions and are capable of delivering high throughput, but do not work well under variable workloads since bandwidth is statically allocated. Moreover, it does not scale well beyond a few cores due to the hardware overhead of creating

more channels. Spatial multiplexing has also been proposed in [27] and could be of potential interest in WiPLASH due to the potential use of steerable antenna arrays. However, their applicability is unclear in packaged environments with lots of multipath and needs to be studied further.

- **Token passing:** Different variants of token passing have been examined as alternatives or even complements to channelization [23, 76, 77]. These solutions work well for distributed, high loads, but not for heterogeneous or hotspot traffic as the token has a fixed trajectory. Also, the protocol does not scale well due to the increasing token turnaround time. Mansoor *et al.* attempt to minimize these issues by means of a predictive scheme that estimates the optimal token occupancy of each node [76]. The work of [24] proposes token-like distributed arbitration protocol with single-bit concurrent probing. However, the probing introduces unfeasible bit-level synchronization among the involved nodes.
- **Random access protocols:** contention-based protocols provide flexible operation and low latency as nodes can attempt to gain access at any time instant and have been explored for WNoC [25, 26] due to their consequently low latency at low loads. BRS-MAC [26] minimizes latency also under moderate loads by using preamble transmission, collision detection, or collision notification via negative ACKnowledgments (ACKs). At high loads, however, the protocol saturates early due to the impact of collisions.
- **Hybrid approaches:** In [19, 25, 59], token/contention hybrid protocols are proposed that attempt to leverage the benefits of both approaches. In [59], utilization metrics are gathered and used to switch between token or random access modes, whereas [19] decides which protocol to use based on the load observed during the first thousands execution cycles of an application. This can have an negative impact in bursty and fast-changing traffic, as the protocol may converge to a non-optimal configuration.
- **Hybrids in other scenarios:** traditionally, research on LAN networks has also tried to combine fixed and random access. In [78], the number of nodes is divided into fixed-size groups connected via a virtual token-passing ring. Groups contend for the channel with CSMA/CD, but the token-holding group has priority. However, the fixed size of the groups and the requirement for ACKs after every transmission discourages its use on multiprocessors. Another hybrid protocol is given in [79], wherein the token holder has priority after the channel changes from busy to idle in any contention period. In the rest of cases, however, all nodes can contend for the channel, which leads to early saturation. In [80], a probabilistic TDMA scheme where nodes transmit to a preferred slot with a given probability a, or elsewhere within the frame otherwise, is proposed. The idea of the two extremes is laid out, but the decision on the parameter a is not discussed.

In summary, the on-chip scenario is driven by latency and reliability, yet with a strong emphasis on energy efficiency, which poses an important challenge at the MAC layer. Existing proposals are in between the high performance of multiplexing/token passing at high, distributed loads, and the promptness and adaptivity of random access protocols especially at low and hotspot loads. We believe that the sweet spot is somewhere in between these two extremes, in solutions that naturally and gradually adapt their characteristics to the load without the need of an external controller. This is what we aim to achieve with FUZZY TOKEN, which is presented in Chapter 5.

B. State of the Art in Multi-Chip Interconnects

Motivated by the rise of the architectural trends of disintegration and specialization, chiplet-based systems are becoming a hot topic in recent years [81, 82]. To enable this chiplet-based approach, fast and efficient chiplet-to-chiplet interconnect fabrics are required to perform data sharing and synchronization across the system. As we describe next, several alternative technologies can be employed to this end.

The WiPLASH project envisages wireless chip-scale communications as the enabler of a new breed of heterogeneous chiplet-based architectures within this new chiplet revolution. Within this context, the goal of this deliverable is to obtain performance and resource consumption models of such a wireless technology. In order to provide baseline models to which the wireless alternative can be compared, this section discusses the state of the art of multi-chip interconnects.

Table B.1 lists the different technological alternatives. Here, it is worth noting that a distinction is made between wireless technology in the mmWave band, and the WiPLASH approach with graphene-based antennas in the THz band. The table also shows that, besides the classical links transporting baseband signals through electrical wires in the package and the wireless links using the package as a transmission medium, two emerging technologies that can also offer inter-chiplet high-speed communication. They are integrated RF transmission lines [83–85] and silicon photonics packaged links [86–88], termed here as *RF/Optical*. As we discuss throughout the section, although these technologies are more efficient and provide more bandwidth than wireless communication, they are more complicated as they still require laying out an extra and overprovisioned network through the package, and less scalable as they are still limited by constraints related to pin scarcity, fanout, or laser power requirements.

The remainder of the chapter is organized as follows. In Section B.1, we discuss the main alternatives in terms of chiplet interconnection, placing emphasis in the packaging technology rather than on the physical media where communication takes place. Then, in Section B.2, we survey the state of the art on the physical layer of wired interconnects, making a distinction between baseband and optical wired/waveguided links. Finally, in Section B.3, we discuss a variety of works on transceivers for ultra-short range and high-rate wireless communications, including chip-scale wireless communications and RF interconnects.

B.1 Chiplet Interconnection Alternatives

The traditional approach to interconnecting multiple chips within the same package is the Multi-Chip Module [MCM, Figure B.1(a)] [89, 90], in a solution that has been recently adopted by AMD in several of its processors [91, 92]. MCMs rely on the integration and interconnection of chiplets directly on top of organic package substrate.

Table B.1: Comparison of different interconnect technologies for NiP. Capacity refers to bisection bandwidth.

Metric	**Electrical**	**RF/Optical**	**Wireless**	***WiPLASH***
Medium	Wires	Waveguides	Package	Package
Frequency	Baseband	mmWave/Optical	mmWave	Terahertz
Capacity (Tb/s)	0.1–1	1–100	0.01–0.1	0.1–1
Latency (ns)	10–100	10–100	1–10	1–10
Energy (pJ/b)	1–10	0.1–10	1–10	1–10
Multiplexing	No	Time*	Time	Total**
Broadcast	Poor	Expensive	Native	Native

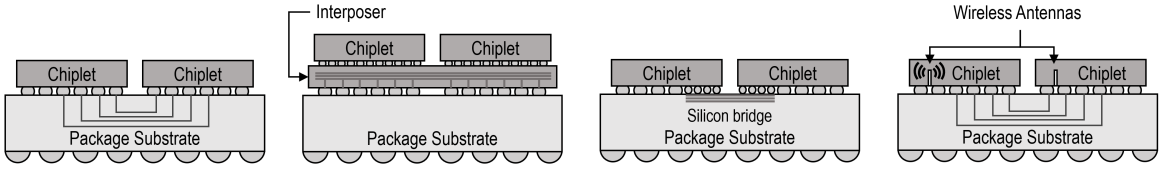
*only if global waveguides are used. **space, time, and frequency.

This option offers large room for accommodating chiplets, e.g. beyond 70×70 mm^2 [89] yet at a relatively coarse I/O bump pitch over 100 μm [90]. To make up for the large pin sizes, Ground-Referenced Signaling (GRS) serial links at multiple tens of Gb/s are implemented instead of more traditional parallel links. This, however, increases the hop latency up to a few tens of nanoseconds [93] and discourages the use of long links across the package. The scalability of the solution is thus limited by latency problems; in fact, most multi-chip architectures count on either a fully connected NiP, e.g. Infinity Fabric [92], or a mesh NiP [89, 93, 94] but up to a few tens of chiplets.

An alternative technology is the silicon interposer, which is effectively a *large chip upon which other smaller dies can be stacked,* as shown in Figure B.1(b) [95]. This allows to interconnect chiplets at a much greater density (i.e. around an order of magnitude greater) than in the classical MCM [95, 96], which enables the implementation of low-latency parallel links. Interposers can be passive, i.e. containing only wires through the metallization layers of the interposer, or active, i.e. containing transistors and other active elements. Therefore, active interposers can even host routers within the interposer, which unleash a set of new choiches in terms of NiPs topology [95,97–99]. However, all these advantages come at the expense of area limitations and a high manufacturing cost. In general, interposers are bound to the reticle limit, which makes interposers larger than 800-mm^2 a challenge. Still, TSMC has recently demonstrated that mask stitching and other techniques can be used to build 1700-mm^2 and even larger interposers [100]. Still, the cost of this solutions increases significantly, rendering them appropriate for very high-end systems. In any case, the connectivity of interposers is still limited by the amount of available pins, which discourages the implementation of high-radix topologies, especially in passive interposers. Instead, the pins are typically used in this context to increase the bandwidth of parallel links in a mesh topology. Therefore, cost and latency are limiting the scalability of this approach.

A third alternative recently promoted by Intel is the Embedded Multi-die Interconnect Bridge (EMIB), shown in Figure B.1(c) [101–103]. The solution consists in integrating very small silicon dies or *bridges* within the package substrate, to which chiplets can be connected at a fine granularity. These bridges are strategically located at the edge of chiplets, allowing two adjacent chiplets to be interconnected with high bandwidth and low latency. As a result, EMIB offers the speed of an interposer without its size constraints. However, the connectivity is clearly limited to neighboring chiplets: in systems with large chiplet counts, certain data communication patterns will require many hops to complete.

In all these cases where only contiguous chiplets are interconnected for various reasons, the high latency associated with inter-chiplet communication renders multi-hop coherence transactions across chiplets extremely costly, decreasing performance and jeopardizing the scaling of the system. Hence, in this context, wireless technology represents an opportunity to greatly alleviate these issues and trigger intresting architectural innovations. As shown in Figure B.1(d), communication happens through electromagnetic waves are radiated by antennas integrated within the chiplets. The waves propagate within the package at the speed of light. That leads to system-wide low latency (comparable to that of a single MCM hop) and inherently broadcast communication without any pin or cost-related size constraints [5, 104]. However, since the medium is shared, wireless technology can only offer a moderate aggregate bandwidth of a few hundred Gb/s. Hence, this technology is expected to complement a wired interconnect alternative. In next sections, we review the state of the art of the underlying technologies for wired and wireless inter-chiplet links.



(a) Multi-Chip Package (b) Silicon Interposer (c) Silicon Bridge (d) Wireless MCM

Figure B.1: Chiplet-to-chiplet interconnection technologies, according to Intel [4] (a, b, c) and Guirado *et al.* [5] (d).

B.2 Physical Layer of Wired Interconnects

The packages described above provide an environment for the integration of wires and waveguides that can transport signals across chiplets. In the next subsections, we review the state of the art in the physical layer of design of such wired interconnects, making a distinction between electrical and optical technologies.

B.2.1 Electrical Links

Baseband signaling through metallic wires embedded within the package substrate or the interposer are the traditional and most widespread alternative. There are essentially two alternatives in this regard, which we describe next: parallel and serial links.

B.2.1.1 Parallel Links

Parallel links are the conventional approach for baseband signaling, where the modulation consists in the charging and discharging of the wires at the pace marked by the system clock. The latency of this kind of links is determined by the amount of synchronous repeaters that may be placed along the wire to avoid timing violations; whereas the throughput of a single wire is equal to one bit per clock cycle T_{clk}. Then, links are built by placing W wires in parallel, leading to a rate of $R = T_{clk}W = \frac{W}{f_{clk}}$. The power consumption of such links essentially depends on the wire length as well

as the capacitance per unit of length, which in turn depends on the wire width and the materials surrounding it. The area occupied by the wire drivers and repeaters is negligible as compared to other types of links.

For its simplicity, the parallel approach is used pervasively. In the multi-chip scenario, however, parallel links are only affordable in interposer-based systems where the I/O pitch is much finer than in MCMs. For recent implementations of such systems, one can find proposals with a diversity of specifications: 1.21 Gb/s/pin at 0.59 pJ/bit [105], 12.8 Gb/s at 0.38 pJ/bit [106], 13 Gb/s and 0.34 pJ/bit at a range of 11mm [107], or 20 Gb/s with 0.47 pJ/bit for below 1-mm lnks [108]. Some specifications may differ as some works report only the chiplet-interposer connection without adding the length of the link within the interposer.

B.2.1.2 Serial Links

Due to the insatiable appetite of modern workloads for bandwidth, the multi-chip paradigm faces the challenge of satisfying such increasing needs. The issue of MCMs and, to a lesser extent, interposers is that the amount of available pins is limited, which means that either connectivity or bandwidth may be reduced. A way of alleviating this bottleneck is to implement serial links in each of the inter-chiplet wires. In serial links, data coming from the processor or memory is serialized and modulated at a high speed using simple low-order modulations such as OOK or 4-PAM, and deserialized on the other end. This way, one can accommodate one order of magnitude higher bandwidth that with conventional signaling. However, this comes at the cost of significant latency related to the need to serialize and deserialize the data. The power consumption will depend on multiple factors such as the desired BER, the link length, or whether the modulation clock is transmitted in parallel for synchronization.

Although serial links have been researched for many years to satisfy the needs of computing systems, the proliferation of chiplet-based systems has triggered new developments in the area. Table B.2 summarizes recent works on serial links at different scales within computing systems, showing a variety of per-lane speeds (up to 64 Gb/s) and efficiencies in the range of 1–10 pJ/bit, essentially depending on the assumed loss at the link.

B.2.2 Optical Links

An alternative to the traditional baseband signaling is found in the optical domain, triggered by the huge advances made in the field of integrated and silicon photonics [128, 129]. On the one hand, optical waveguides can be integrated within silicon chips, as a modified Si/SiO_2 interface can be used to implement the core and cladding of such a waveguide. As a result, these can be also built within interposers, which serve as natural media for inter-chiplet optical interconnection [130]. Waveguides integrated in PCB follow a different process but are equally possible. On the other hand, components such as miniaturized optical modulators, splitters, photodetectors, and even laser sources can be integrated within the chiplets [129]. Optical signals are therefore modulated on-chip, coupled within the waveguides off-chip and back to the chiplets, where they can be filtered and detected. Although the technology is less mature than the classical electrical signaling, this option provides the advantage of speed-of-light propagation, leading to ultra-low latency across the system, and ultra-

Table B.2: Comparison of recent electrical links from the literature

Ref	Tech	Speed/lane	Energy	Loss	BER
[109]	16nm	25 Gb/s	1.17 pJ/bit	8.5 dB	10^{-15}
[110]	16nm	56 Gb/s	2.25 pJ/bit	11 dB	10^{-15}
[111]	16nm	64.375 Gb/s	2.95 pJ/bit	8.6 dB	10^{-6}
[112]	16nm	56 Gb/s	5.8 pJ/bit	32 dB	10^{-12}
[113]	16nm	56 Gb/s	9.1 pJ/bit	31 dB	10^{-8}
[114]	28nm	20.83 Gb/s	5,64 pJ/bit	3 dB	10^{-15}
[115]	28nm	20 Gb/s	0.54 pJ/bit	1 dB	10^{-12}
[116]	28nm	56.2 Gb/s	4.4 pJ/bit	18.4 dB	10^{-12}
[117]	32nm	12 Gb/s	3.15 pJ/bit	10 dB	10^{-12}
[118]	32nm	12 Gb/s	1.4 pJ/bit	3 dB	10^{-9}
[119]	40nm	16 Gb/s	5.3 pJ/bit	12dB	10^{-12}
[117]	45nm	10 Gb/s	1.4 pJ/bit	8 dB	10^{-16}
[120]	45nm	10 Gb/s	5.3 pJ/bit	11.1 dB	10^{-9}
[121]	45nm	12 Gb/s	2.5 pJ/bit	15 dB	10^{-14}
[122]	45nm	8.9 Gb/s	1.9 pJ/bit	20 dB	10^{-9}
[123]	65nm	10 Gb/s	4.18 pJ/bit	14.5 dB	10^{-12}
[124]	65nm	16 Gb/s	5.62 pJ/bit	15 dB	10^{-12}
[114]	65nm	5 Gb/s	2.7 pJ/bit	7 dB	10^{-12}
[125]	65nm	12.5 Gb/s	0.98 pJ/bit	12.1 dB	10^{-12}
[126]	90nm	6.25 Gb/s	2.2 pJ/bit	15 dB	10^{-15}
[127]	90nm	5 Gb/s	11.8 pJ/bit	16 dB	10^{-12}

high bandwidth as a single waveguide can carry multiple wavelengths, each of which can easily transport 10 Gb/s. The power consumption is intrinsically much lower than baseband signaling, yet the reality is that the insertion losses introduced by all the elements, including the waveguides, couplers, splitters, modulators, typically lead to high laser power requirements and dilute the energy advantage of these interconnects.

Work on silicon photonics has been performed for over a decade [131], yet the explosive growth of data center networks and the advent of the chiplet paradigm has given a new light to this area with the promise of achieving a bandwidth density of Tb/s/mm^2 and efficiencies of a few pJ/bit or even lower [86].

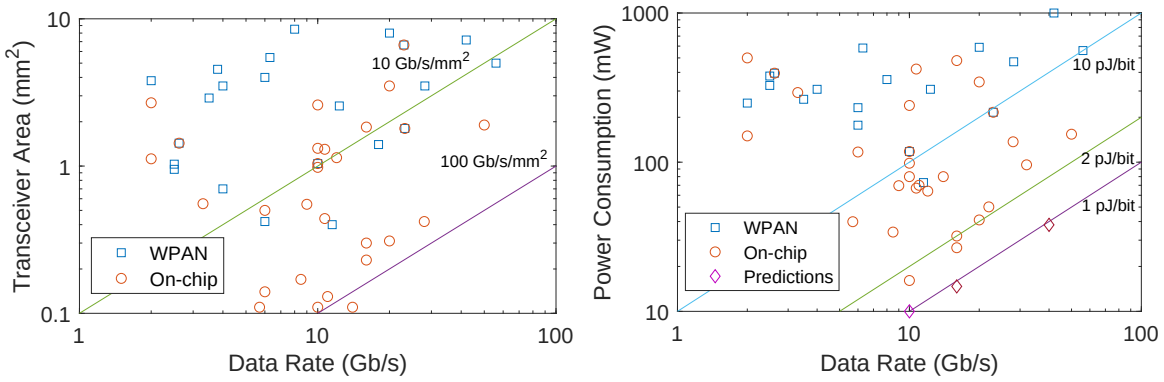
B.3 Physical Layer of RF/Wireless Interconnects

The physical layer of RF interconnects is similar to that of serial links in that data is serialized and modulated. However, RF interconnects imply the modulation of signals using a carrier frequency much higher than the system clock or modulation rate, in the mmWave or THz bands. This requires a transceiver capable of performing this up-/down-conversion and, if needed, of amplifying the RF signals.

RF interconnects can both consist of *wired* links whereby EM waves are guided within integrated transmission lines, or *wireless* links where the waves are radiated through the by chip package by means of an antenna. In the former case, several works have explored multiple alternatives in the form of plastic or polymer waveguides to be interfaced with the chips [85, 132, 133]. For the latter case, several works have

Table B.3: Selection of transceiver proposals for chip-scale communications.

Reference	Year	Technology	Modulation	Frequency	Speed	Power	Area
[135]	2012	65nm CMOS	OOK	260 GHz	10 Gb/s	1.17 W	2.6 mm^2
[36]	2014	65nm CMOS	OOK	60 GHz	16 Gb/s	32 mW	0.23 mm^2
[136]	2015	65nm CMOS	QPSK	240 GHz	16 Gb/s	480 mW	1.84 mm^2
[61]	2017	130nm SiGe	BPSK	190 GHz	50 Gb/s	154 mW	1.9 mm^2
[137]	2020	65nm CMOS	OOK	60 GHz	12.5 Gb/s	33 mW	0.15 mm^2
[74]	2021	250nm InP	OOK	290 GHz	20 Gb/s	N/A	0.3 mm^2



(a) Area (b) Power

Figure B.2: Area and power consumption of sub-THz and THz transceivers (from 0.06 to 0.43 THz) for short-range high-rate wireless applications. Each data point indicates the area/power and data rate of a single transceiver prototype or theoretical predictions made in the literature. Only the analog part is considered. Data extracted from [6] and references therein.

investigated on-chip antennas capable of providing integrability and wide bandwidths for this scenario. For instance, Hou *et al.* present a Vivaldi antenna with a peak gain of 5.5 dBi and 30% bandwidth around 150 GHz [134]. Approaching the THz band, microstrip leaky-wave antennas achieving 4.9 dBi with more than 26% bandwidth around 245 GHz or skirt-shaped designs with peak gain of 7.1 dBi and a huge 65% bandwidth at 1 THz have been reported [52]. A complete analysis of different alternatives is given in Deliverable D3.1 [1].

With regards to the transceiver implementation, Table B.3 lists a selection of proposals that have been conceived for chip-to-chip or on-chip communication applications. We observe how the performance and resource consumption of these proposals has been improving over the years and has provided designs that are comparable in performance and efficiency to wired chip-to-chip interconnects.

For a broader analysis covering applications with moderately longer distances and higher rates, Figure B.2 shows a snapshot of the state of the art of high-speed and short-range wireless communication transceivers, which uses data from our own survey of designs for both on-chip and Wireless Personal Area Network (WPAN) applications. Table B.4 lists the specification ranges of the analyzed transceivers.

On the on-chip side of things, current transceiver implementations modulate data on a high frequency carrier in the V band range (40-75 GHz) or beyond, using simple schemes that reduce the area and power footprint. A representative example is that proposed by Yu *et al*, a 30/60/90-GHz transceiver implemented with 65nm CMOS that performs very close to the targets set by the WNoC paradigm as it delivers up to 48

Table B.4: Summary of the specifications of the analyzed transceivers.

Applications	Wireless Personal Area Networks (WPAN) Wireless Chip-scale Networks
Technology	28–130 nm CMOS 28–45 mn FDSOI CMOS 55–250 nm SiGe BiCMOS/HBT 80 nm InP HEMT, 250 nm InP DHBT
Transceiver Architecture	Impulse Radio (IR), Continuous Wave (CW)
Modulation	On-Off Keying (OOK), Amplitude Shift Keying (ASK), Phase Shift Keying (BPSK/QPSK/), Frequency Shift Keying (FSK), Quadrature Amplitude Modulation (16/64/128/256 QAM)
Operation Frequency (f_c)	8–820 GHz
Transmission Range (d_{max})	0.06–222 cm
Data Rate (R)	2–120 Gbps

Gbps with a BER of 10^{-15} at a few centimeters, while occupying 0.8mm^2 and consuming 95mW ($\sim$2pJ/bit) [7]. Another example is the transceiver presented in [138], which is implemented in 65nm CMOS and operates in the 85-90 GHz band, achieving an energy efficiency of 1.5 pJ/bit when transmitting at 6 Gbps and occupying 0.09mm^2. However, this design does not integrate an amplifier, which would render it insufficient for wireless links. Beyond this, several proposals have pushed the frequency up to the 135-140 GHz band, yet with modest performance due to the early stages of development of THz technology. A design close to the WNoC requirements is presented in [139] in 40nm CMOS, delivering 10 Gbps at around 10 pJ/bit with a BER of 10^{-11} for a transmission range of 10 cm. Following these advancements, recent years have seen a surge in THz circuits for wireless communications and imaging [140–142]. First complete integrated transceiver designs have been appearing, with two good examples being the SiGe BiCMOS implementation at 190 GHz promising 40 Gbps and 3.9 pJ/bit at 2cm range [61] or the 240-GHz design achieving 16 Gb/s at 30 pJ/bit for a few centimeters as well [136, 143]. Further, a transceiver fabricated with an InP process has been demonstrated to work beyond 10 Gb/s at around 300 GHz with an area of around 0.25 mm^2 [74], yet without power consumption results. Such trend is expected to continue in the THz band, where significant efforts are devoted to filling the so-called *THz gap* [144–148].

B.3.1 Models for Wireless Interconnects

Given the amount of published data on complete transceivers, one can build models based on the published specifications. Yet surveys of wireless transceivers are scarce, limited to a specific application, or have not had continuity to be updated across the years. Examples relevant to WNoC could be the surveys from Gorisse *et al.* [43] or Blanckenstein *et al.* [149]. The former was published in 2012 and focused on multi-gigabit transceivers. The amount of data points was rather small at the time and the work has not been continued. The latter was published in 2015 and focused on ultra-low power transceivers ($<$1 mW) for Wireless Sensor Network (WSN). Such an anal-

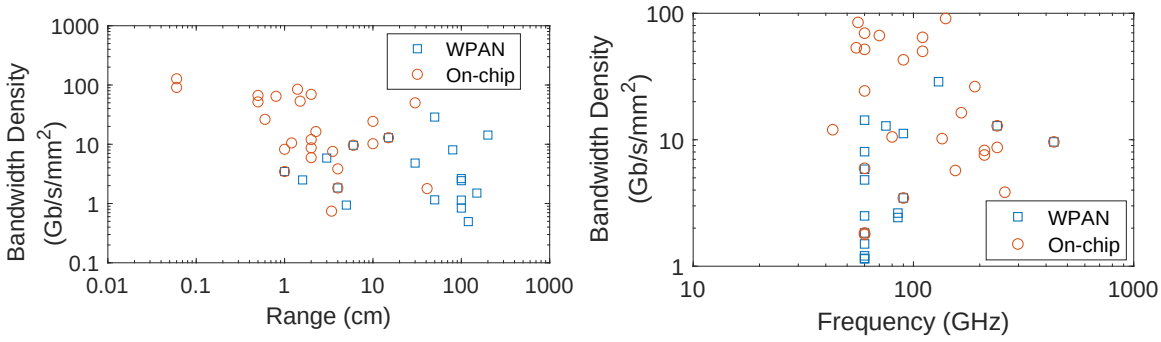


Figure B.3: Transceiver bandwidth density $\delta_{TRX} = \frac{R}{A_{trx}}$ as a function of the transmission range or central frequency for [6–8] and references therein.

ysis, however, is hardly applicable to WNoC due to the low transmission rates, which are in the Kb/s and Mb/s range.

In light of the lack of a survey covering the transceivers developed in the last decade for high-speed and short-range wireless communications, we performed our own survey that considers a heterogeneous set of transceiver proposals as summarized in Table B.4 and summarized above. Here, we evaluate the area and power of all the designs and attempt to see their dependencies.

B.3.1.1 Area Models

Figure B.3 plots the relation of the bandwidth density with respect to the transmission range and to the frequency. On the one hand, as expected, on-chip designs have been demonstrated in ranges lower than those for WPAN applications. In either case, the lower the range, the larger the bandwidth density. This can be explained in two ways: (i) the transceiver components are simpler as the performance requirements are relaxed, especially in the case of the power amplifier, which could even be eliminated. Therefore the transceiver is smaller. (ii) For a fixed error rate, and thus a fixed SNR, reducing the transmission range implies that a larger bitrate is supported. Therefore, the bandwidth density that can be accommodate is theoretically larger. From this figure, one could extract a relation between the two terms through linear fitting, despite each analyzed work considering a different path loss model or measurement setup.

On the other hand, the plot describing the relation between bandwidth density and frequency does not provide a clear trend. One reason is that many works focus on the unlicensed band at 60 GHz and therefore there is not enough representative frequencies for a model. Yet, one can infer certain trends from the plot, namely, (i) increasing frequency towards the THz band seems to reduce the bandwidth density for on-chip environments and improve moderately that of WPAN transceivers. In the former case, higher path loss lead to the need of more powerful amplifiers which take significantly more area, in bands where the PAE decreases. In the case of WPAN, instead, going to higher bands provides larger bandwidth and relaxes the spectral efficiency requirements, which simplifies the transceivers.

Based on the trends identified above, we have repeated the plot on the area as a function of the datarate, yet now adjusting the area according to the transmission range using the fitting model provided above. As shown Figure B.4, we plot all points plus two lines corresponding to a conservative fit that considers all designs equally important

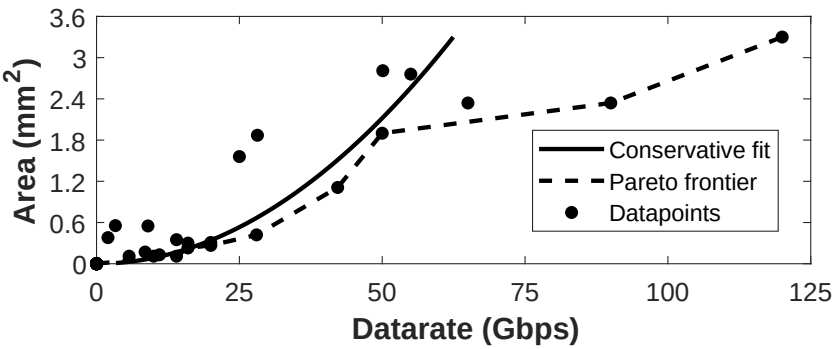


Figure B.4: Transceiver area as a function of the datarate for [6–8] and references therein.

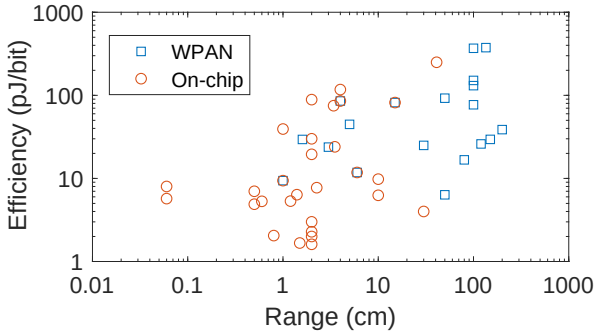


Figure B.5: Energy per bit as a function of the transmission range for [6–8] and references therein.

and to a Pareto frontier for the most area-efficient implementations. Interestingly, areas below 0.5 mm^2 are currently achievable even beyond 25 Gb/s at the chip scale.

B.3.1.2 Power Models

In a similar approach than that taken for the area modeling, here we plot the energy per bit of a single transceiver as a function of the transmission range. From the results of Figure B.5, it is very interesting to observe that the efficiency improves as the range of operation is reduced. This is because at lower transmission distances, the channel introduces less losses and the transceiver requirements in terms of gain are relaxed.

Due to the apparent relation between efficiency and transmission range, there have been works that have proposed efficiency figures of merit that include the transmission range. For instance, Gorisse *et al.* considers a normalization by the square root of the distance. These considerations are accounted for in Figure B.6, which illustrates how the power consumption scales with the datarate, based on out analysis and a normalization of the transmission range and error rate for a fair comparison between designs. Again, a conservative fit and the Pareto frontier are extracted from this the data in this plot.

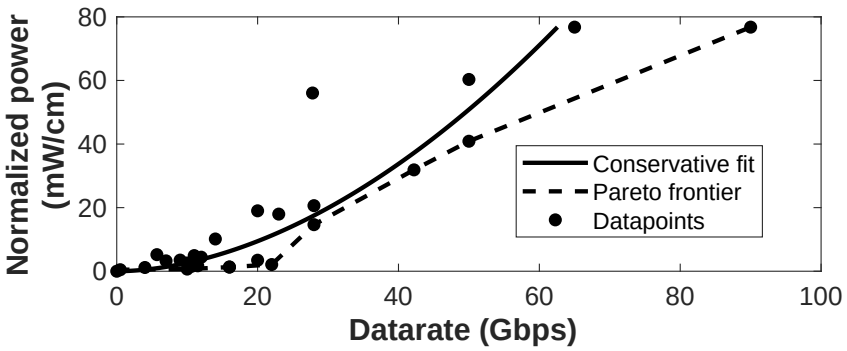


Figure B.6: Transceiver power as a function of the datarate for [6–8] and references therein. Power is normalized to transmission range and to 10^{-9} error rate. Energy per bit can be obtained dividing the power by the datarate.

C. Channel Simulations and Models

In this section the results of the simulations made to each configuration are collected. The variations of parameters and materials performed to each scenario allowed evaluating the behavior of each configuration in time and frequency domain. The appendix is divided into 3 sections that contemplate each scenario, each section is divided into 2 more to show the results on path losses and delay spread gather from frequency and time domain analysis to provide path loss and dispersion models. For the time domain simulations, the delay spread at distance of 2mm, τ_{rms}(2mm), the exponent γ_t, the worst case of delay spread τ_{rms}, and the coherence bandwidth B_c in GHz are gathered. For all the analysis, unless otherwise stated, the parameters are the ones stated in the default configuration. The path losses and delay spread considered will be at a distance of 2mm.

C.1 Flip-chip Models

The default package dimensions and materials for the simulations, as well as the variations can be seen in tables 3.1 and 3.2 respectively.

The first stage of simulations was done at mmWave frequencies to later scale towards THz frequencies. Approximately 62 variations and simulation scenarios were made to characterize the behavior of the flip chip in frequency and time. Here is provided the impact of varying the dimensions of some parameters such as the silicon and the heat spreader thicknesses or the lateral filling material. The idea is to get a grasp of how this changes in the scenario affect the path losses and delay spread. Unless otherwise specified the values for the parameters will be the same as in Section 3.2

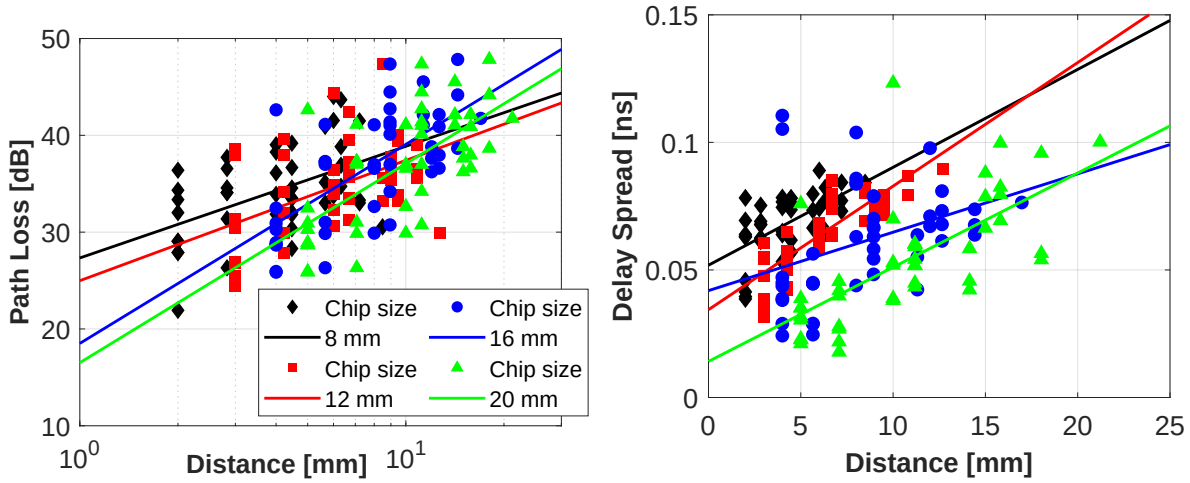
Die Size: Varying the size of the die while maintaining the others parameters constant has an influence on path losses and delay spread as shown in Fig. C.1(a). Some observations to draw from this are that larger chips allow to implement longer wireless links that lead to larger losses and dispersion. The range of losses doesn't vary much with the size of the chips. In terms of the delay spread, larger chips lead to higher delay spread in general. Also is worth noticing that larger chips apparently improve the mid range links in performance. The longer distance between antennas and the lower impact of parameters such as margin dimensions could be the cause of this improvement. This result might not be extensible to all combinations of Si and AIN thicknesses.

Package dimensions and filler: The package was first simulated on the original conditions and then the lateral interface material was changed to epoxy. In Fig. C.1 for 60 GHz, the use of epoxy seems to improve the path losses.The effect of the filler was simulated for 240GHz, not seen here for the sake of brevity. For 240 GHz, the effect of the material change is less noticeable, although it shows points of improvement (at

8.5mm the path loss is improved by 12dB). Is also worth mentioning, that the impact of using epoxy instead of vacuum is always present in the simulations, although the amount of influence varies with the silicon and AIN thickness. The reason for this behaviour can be the larger refractive index of the epoxy, which becomes closer to that of SI/AIN.

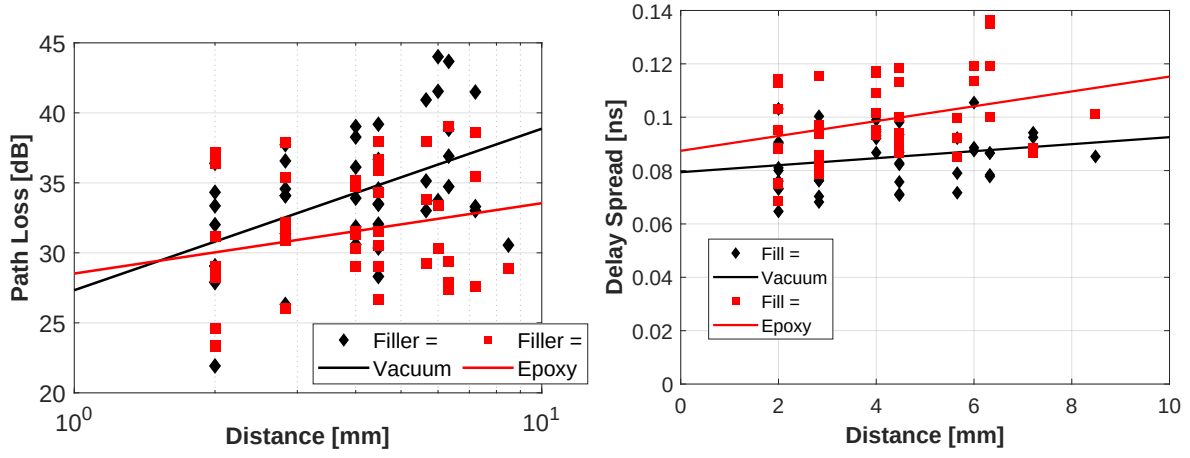
From Fig. C.1 we observe that different dimensions of the lateral margin have a relatively small impact o the scaling of path losses over distance. The trend is similar for 60GHz and 240GHz. The effect is more noticeable in long links, whose main components comes from this lateral space. When reducing the margin, the waves travelling through the sides of the packages arrives stronger and faster, reducing the path loss. In the delay spread the effect is more noticeable, for a package margin of 1mm the delay spread is below 0.05 ns.

Below, we list the tables concerning all simulation variations from the perspective of frequency and time analysis. In the tables, *Margin* refers to lateral space between the end of the chip and the package limits.



(a) (b)

Figure C.1: Path losses and Delay Spread for different die sizes



(a) (b)

Figure C.2: Path losses and Delay Spread for different die sizes at 60GHz

C.2 Interposer

The default package dimensions and materials for the simulations, as well as the variations can be seen in Tables 3.6 and 3.7 respectively.

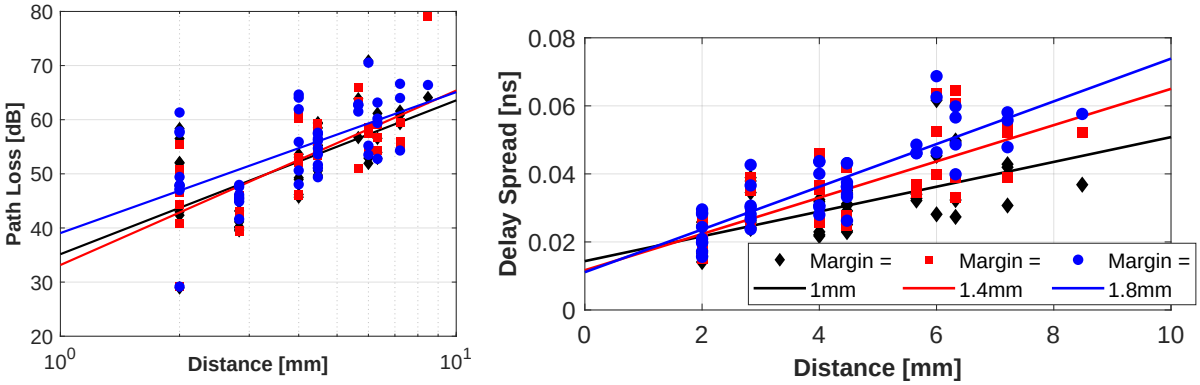
C.2.1 Channel Models

In this section is evaluated the influence of different variations in interposer results. For this, around 50 simulations were made to get a good charaterization of the package.

Number of chiplets:The interposer was evaluated with 4 and 16 chiplets, always leaving a separation of 2mm between chiplets and between the edge chiplets and package limits. Figure C.4(a) shows that more chiplets lead to an improvement of the path loss, upto 10dB. Reasons for this behavior may be (i) that the waves leave the chiplet sooner and, instead of propagating through lossy silicon, they propagate through the lossless filler and/or couple onto the interposer to reach the rest of chiplets more efficiently. We also note that such small chiplets may become a resonant structure and lead to distorted or more directive radiation patterns at certain frequencies. We can also see that more chiplets lead to a rather constant increment of the delay spread of around 0.02 ns in average.The worst-case delay spread increases from 0.2 to 0.25 ns, reducing the coherence bandwidth from 5 to 4 GHz. One possible reason is the more frequent change of propagation medium, which may be generating more reflections at the interface between the chiplets and the package. These reflections may accumulate at the tail of the received signal.

Inter-chiplet separation In Fig. C.2.1 the path loss and delay spread for different chip separations is plotted. seem to imply that larger separations lead to better path loss, especially at longer distances. The improvement can be larger than 20 dB. The reasons are compatible with the discussion made above for varying number of chiplets. On the other hand, the delay spread analysis seems to imply that the improvement in path loss comes at the cost of a degradation of the delay spread. However, the effect is clearly focused on mid-range links.

Interposer resistivity An interposer package was simulated for low-resistivity and high-resistivity silicon interposers. It appears that the high-resistivity silicon may be well supporting the propagation of waves within the package. For a silicon thickness of



(a) (b)

Figure C.3: Path losses and Delay Spread at 240GHz for different lateral space or margin.

Table C.1: Channel models of flip-chip in the frequency domain.

Frequency	Die-side	Si	AIN	Margin	Filler	PL_0	γ
60GHz	8	0.1	0.1	1	Epoxy	28.03	1.8461
60GHz	8	0.1	0.1	1.4	Epoxy	29.56	1.5203
60GHz	8	0.1	0.1	1.8	Epoxy	26.66	2.0128
60GHz	8	0.1	0.5	1	Epoxy	30.03	0.5024
60GHz	8	0.5	0.1	1	Epoxy	34.38	3.4063
60GHz	8	0.5	0.5	1	Epoxy	30.1	3.3954

Frequency	Die-side	Si	AIN	Margin	PL_0	γ
60GHz	12	0.1	0.1	1	29.38	3.1777
60GHz	16	0.1	0.1	1	14.32	4.0046
60GHz	20	0.1	0.1	1	21.44	3.4457
60GHz	12	0.1	0.5	1	28.83	1.2427
60GHz	16	0.1	0.5	1	24.69	2.0561
60GHz	20	0.1	0.5	1	27.06	1.3886
60GHz	12	0.5	0.1	1	27.53	5.3195
60GHz	16	0.5	0.1	1	27.07	5.5588
60GHz	20	0.5	0.1	1	19.72	5.9863
60GHz	12	0.5	0.5	1	20.39	5.0687
60GHz	16	0.5	0.5	1	21.13	5.0764
60GHz	20	0.5	0.5	1	18.09	4.7711

Frequency	Die-side	Si	AIN	Margin	PL_0	γ
120GHz	8	0.1	0.1	1	22.03	3.6080
120GHz	8	0.1	0.5	1	18.79	0.7214

Frequency	Die-side	Si	AIN	Margin	PL_0	γ
180GHz	8	0.1	0.1	1	42.41	1.6648
180GHz	8	0.1	0.5	1	31.69	2.7625

Frequency	Die-side	Si	AIN	Margin	PL_0	γ
240GHz	8	0.1	0.1	1	43.71	2.8421
240GHz	8	0.1	0.1	1.4	47.48	3.0795
240GHz	8	0.1	0.1	1.8	51.02	2.4627
240GHz	8	0.1	0.5	1	44.49	1.3667
240GHz	8	0.5	0.1	1	54.83	4.1602
240GHz	8	0.5	0.5	1	51.27	2.8549

Frequency	Die-side	Si	AIN	Margin	PL_0	γ
240GHz	8	0.1	0.5	1	43.58	0.4257

0.1 mm and AIN of 0.5 mm, the impact of using bulk silicon instead of a high-resistivity material on path loss is marginal.A potential reason may be that the bump arrays are sort of a barrier hindering the coupling of waves to the interposer. The case of the delay spread is interestingly different. In this case, the use of bulk silicon reduces

Table C.2: Channel models of flip-chip in the time domain.

Die-side	Si	AIN	Margin	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
8	0.1	0.1	1	0.0202	0.0057	0.0687	14.55
8	0.1	0.1	1.4	0.02138	0.0076	0.0695	14.38
8	0.1	0.1	1.8	0.0212	0.0125	0.1015	9.85
8	0.1	0.5	1	0.08201	0.0013	0.1055	9.48
8	0.5	0.1	1	0.01173	0.0041	0.0483	20.72
8	0.5	0.5	1	0.03043	0.0090	0.1217	8.2144

Die-side	Si	AIN	Margin	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
12	0.1	0.5	1	0.08931	0.0037	0.1473	6.7874
16	0.1	0.5	1	0.07602	0.0055	0.1531	6.5332
20	0.1	0.5	1	0.05679	0.005	0.1684	5.9376

Die-side	Si	AIN	Margin	Filler	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
8	0.1	0.1	1	Epoxy	0.03267	0.0103	0.1062	9.4201
8	0.1	0.1	1.4	Epoxy	0.03427	0.0123	0.1126	8.8802
8	0.1	0.1	1.8	Epoxy	0.03804	0.0166	0.1588	6.4202
8	0.1	0.5	1	Epoxy	0.09297	0.0028	0.1363	7.3384
8	0.5	0.1	1	Epoxy	0.01068	0.018	0.1454	6.8785
8	0.5	0.5	1	Epoxy	0.03407	0.0111	0.1185	8.4392

the delay spread in half. One possible reason may be that the lossy interposer is attenuating multipath rays that would otherwise lead to higher dispersion.

Filling material The filling material was switched from vacuum to epoxy. The results that, similarly to in flip-chip packages, using epoxy instead of vacuum could improve path loss. However, in the interposer case, the introduction of epoxy resin as package filling material seems to be also helping reduce the delay spread, which did not happen in simple flip-chips. The reason for this behavior is that the change of

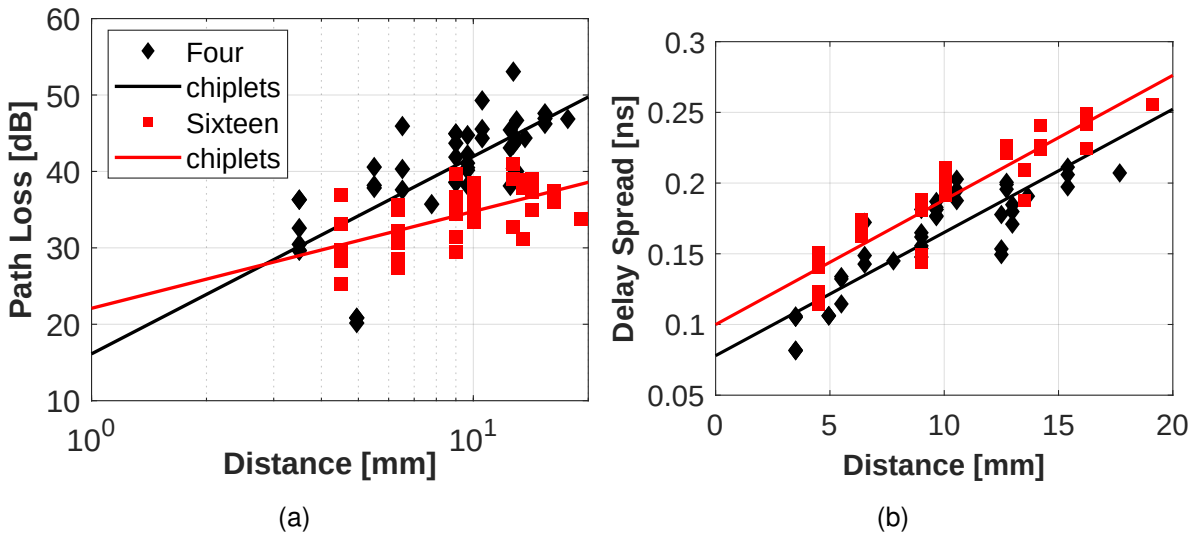


Figure C.4: Path loss and delay spread in an interposer divided into 4 or 16 chiplets

refractive index between the chiplet and the package is less abrupt, reducing intra-chiplet reflections and improving the chiplet–package transition. We speculate that, since chiplets are smaller and the filling material is also present in the space between chiplets (and not only in the package margins), the impact is more profound and positive.

C.3 WireBond

The default package dimensions and materials for the simulations, as well as the variations can be seen in tables 3.11 and 3.12 respectively. We note that, due to the high attenuation of some ports at very close positions, the fitting lines may lead negative exponents.

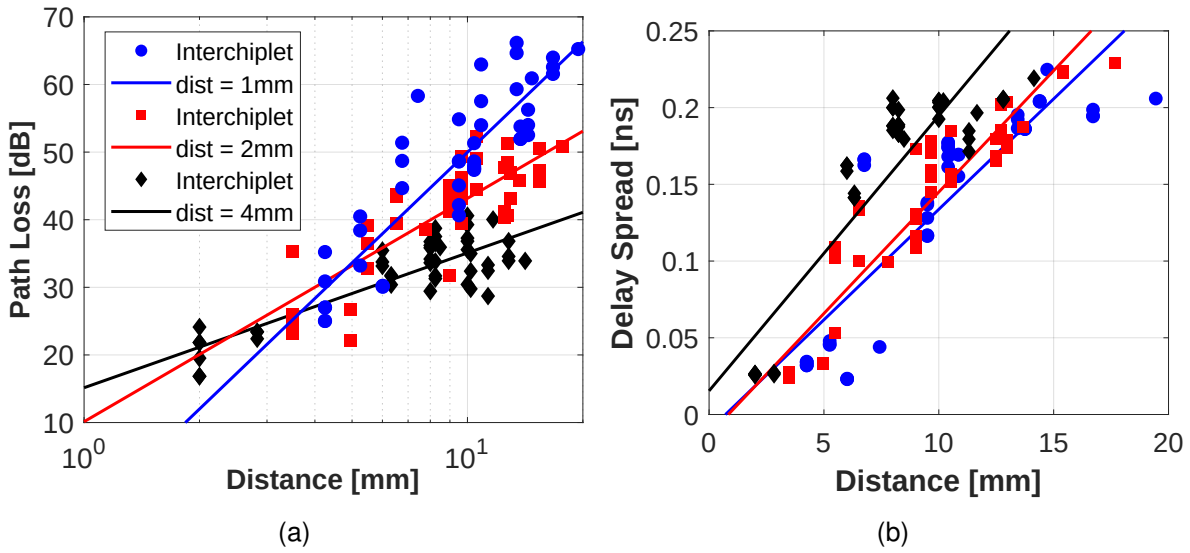


Figure C.5: Path loss and delay spread for different inter-chiplet spacings.

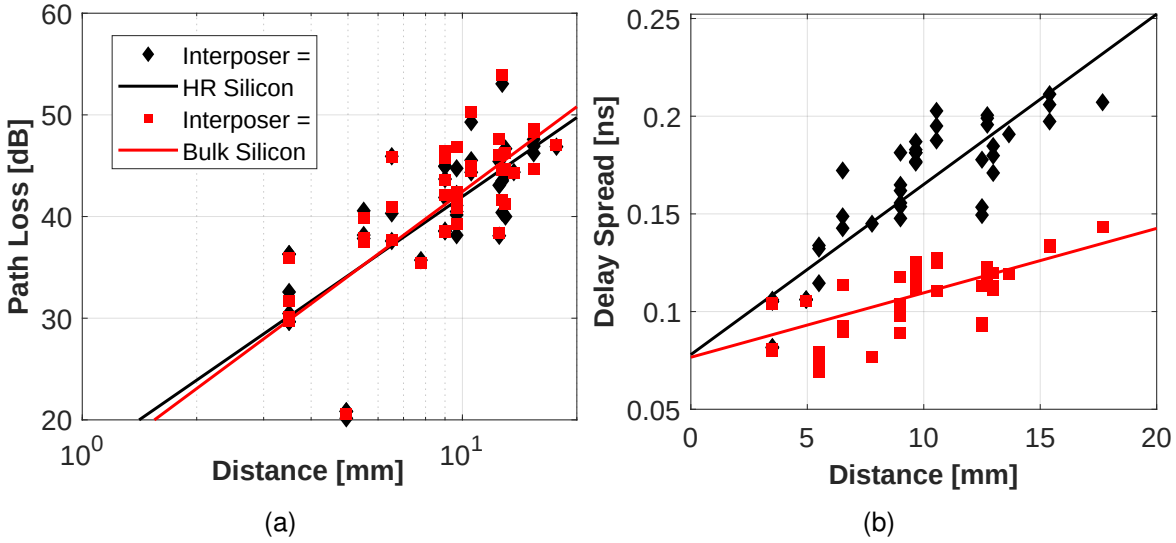
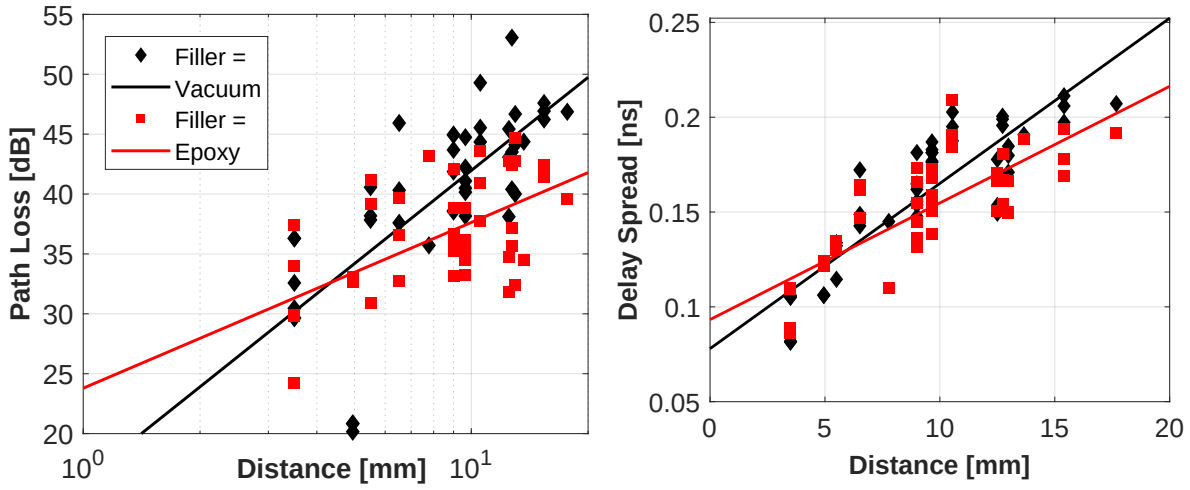


Figure C.6: Path loss and delay spread in an interposer for low-resistivity and high-resistivity silicon interposers.

Die size: Fig C.8 increasing the die size improves the attenuation suffered by the electromagnetic waves by 5–10 dB in average between 8-mm and 20-mm dies. In terms of delay spread, the size of the chip does not change the main linearly increasing trend. When scaling to 12 or 16 mm, the same average trend is observed, but simply extrapolated to longer distances.

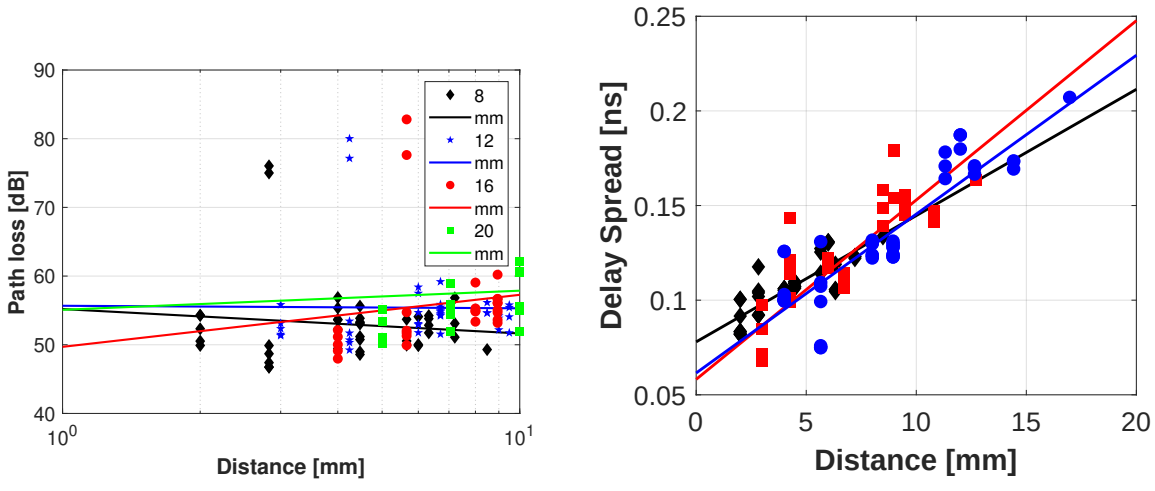
I/O pitch: Figure C.9 shows the path loss and delay spread as a function of the number of bond wires. The values for path loss may increase by more than 10 dB when increasing the number of bond wires in the periphery of the chip from 32 to 128. From the second plot, we see that the presence of a dense array of bond wires seems to have a positive effect on the delay spread. Worst-case spreads as low as 0.1 ns (10 GHz) are shown in this plot for 128 wires in contrast to the value of 0.13ns (7.7 GHz) for 32 wires

Top of the package: Finally is assessed how the decisions relative to the vertical dimensions of the package and the material of the lid affect the channel. The use of a metallic can improve the path loss by variable amounts between a few dB upto more



(a) (b)

Figure C.7: Path loss and delay spread for different filling materials.



(a) (b)

Figure C.8: Path loss and delay spread for different die sizes.

Table C.3: Channel models of the interposer package in the frequency domain. By default, number of chiplets is 4 and filler is vacuum.

Frequency	Die-side	Si	AIN	Separation	PL_0	γ
60GHz	20	0.1	0.1	1	27.98	2.2173
60GHz	20	0.1	0.1	2	31.42	3.3606
60GHz	20	0.1	0.1	4	25.21	5.3179
60GHz	20	0.1	0.5	2	31.15	2.5834
60GHz	20	0.5	0.1	2	39.77	4.5022
60GHz	20	0.5	0.5	2	31.06	5.0369

Frequency	Die-side	Si	AIN	Separation	Filler	PL_0	γ
60GHz	20	0.1	0.1	1	Epoxy	18.21	1.6875
60GHz	20	0.1	0.1	2	Epoxy	19.73	2.8751
60GHz	20	0.1	0.1	4	Epoxy	19.78	3.5958
60GHz	20	0.1	0.5	2	Epoxy	35.96	1.3719
60GHz	20	0.5	0.1	2	Epoxy	38.25	4.1678
60GHz	20	0.5	0.5	2	Epoxy	29.8	4.5977

Frequency	Die-side	Si	AIN	Separation	PL_0	γ
120GHz	20	0.1	0.5	2	15.35	5.3794
180GHz	20	0.1	0.5	2	23	4.6952
240GHz	20	0.1	0.5	2	35.48	3.6252

Frequency	Die-side	Si	AIN	Separation	Chiplets	PL_0	γ
60GHz	20	0.1	0.1	2	16	29.37	1.8956
60GHz	20	0.1	0.5	2	16	34.44	1.2671
60GHz	20	0.5	0.1	2	16	40.42	2.0063
60GHz	20	0.5	0.5	2	16	41.1	2.0063

Frequency	Die-side	Si	AIN	Separation	Interposer material	PL_0	γ
60GHz	20	0.1	0.1	2	Bulk silicon	31.17	3.7019
60GHz	20	0.1	0.5	2	Bulk silicon	31.62	2.772
60GHz	20	0.5	0.1	2	Bulk silicon	38.96	4.9392
60GHz	20	0.5	0.5	2	Bulk silicon	30.7	5.2484

than 10 dB. In case of delay spread, since the package starts becoming an attenuated reverberation chamber, the delay spread is expected to increase. As shown in the figure, the delay spread with the metallic cover is around 2x larger than the delay spread obtained with the ceramic cover.

The dimensions of the molding compound that fills the cavity containing the bond wires does not have a noticeable impact in path loss as is seen from Fig. C.11(a).This suggests that the main propagation mechanism is either surface waves at the interface between the insulator and other materials, or space waves within the silicon/heat spreading material. In terms of delay spread, the effect of the molding compound thickness is more noticeable –it leads to an increase of around 40 ps.

Table C.4: Channel models of the interposer package in the time domain. By default, number of chiplets is 4 and filler is vacuum.

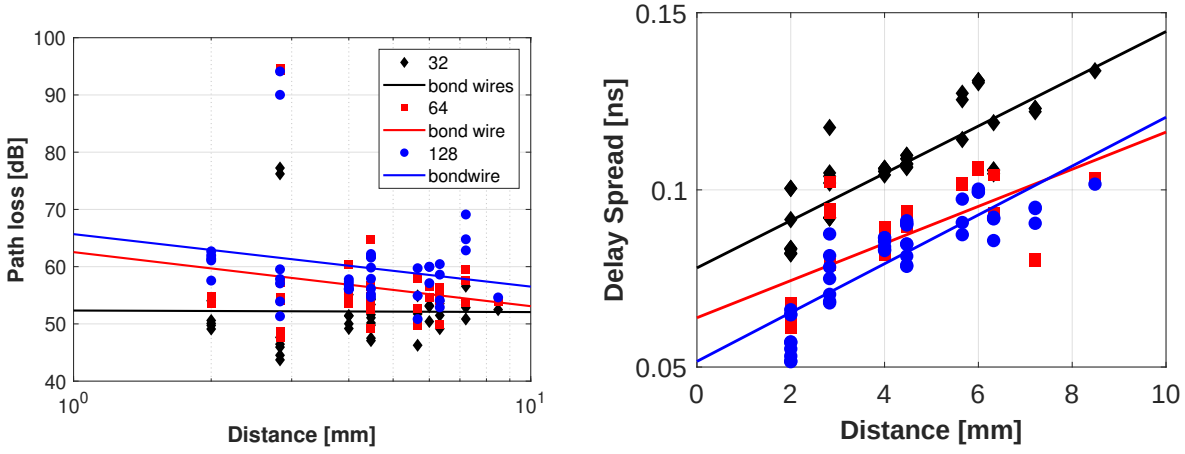
Die-side	Si	AIN	Separation	Chiplets	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
20	0.1	0.5	2	16	0.1176	0.0088	0.2553	3.91

Die-side	Si	AIN	Separation	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
20	0.1	0.1	1	0.05191	0.0144	0.2191	4.56
20	0.1	0.1	2	0.01838	0.0158	0.2288	4.37
20	0.1	0.1	4	0.01977	0.0157	0.2247	4.44
20	0.1	0.5	2	0.09539	0.0087	0.2113	4.73
20	0.5	0.1	2	0.003031	0.0167	0.2275	4.39
20	0.5	0.5	2	0.02422	0.0119	0.1994	5.01

Die-side	Si	AIN	Separation	Filler	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
20	0.1	0.5	2	epoxy	0.1056	0.0062	0.2088	4.78

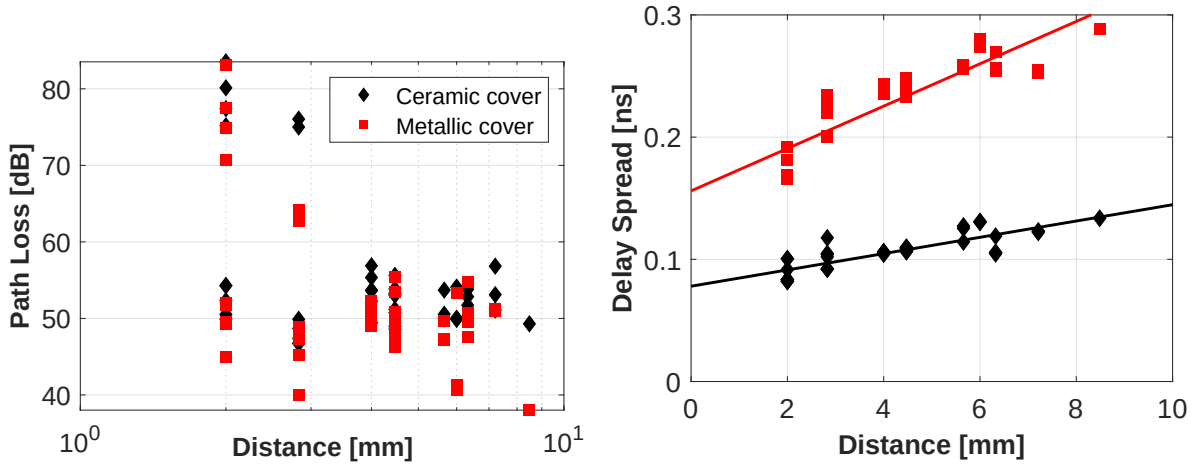
Die-side	Si	AIN	Separation	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
20	0.1	0.5	2	0.08321	0.0033	0.1433	6.97

Die-side	Si	AIN	Separation	Material	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
20	0.1	0.1	2	Bulk Si	0.01498	0.0044	0.0888	11.25
20	0.1	0.5	2	Bulk Si	0.1362	0.0048	0.2363	4.23
20	0.5	0.1	2	Bulk Si	0.01035	0.0034	0.0697	14.34
20	0.5	0.5	2	Bulk Si	0.08032	0.0014	0.1789	5.58



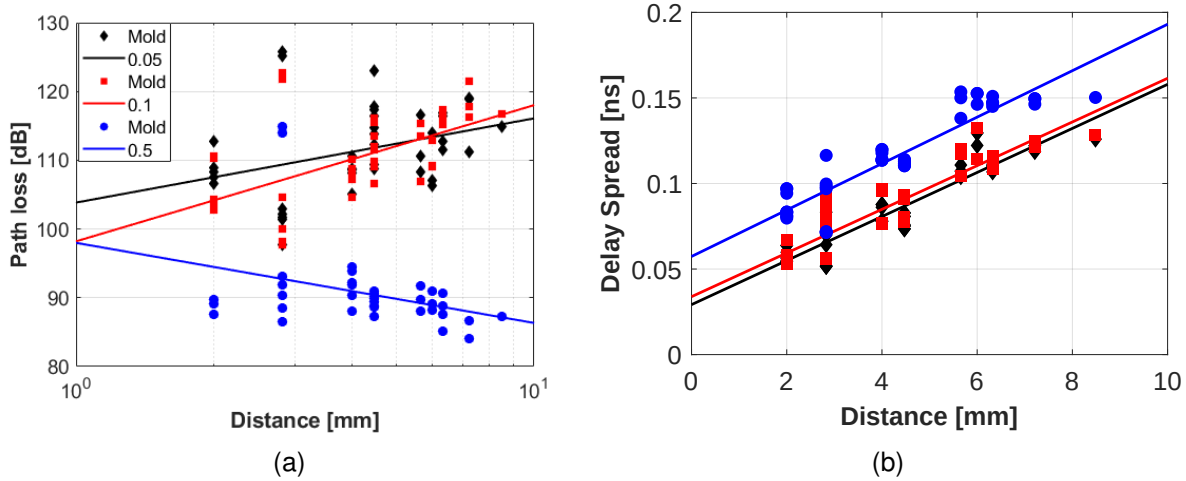
(a) (b)

Figure C.9: Path loss and delay spread for different number of wires.



(a) (b)

Figure C.10: Path loss and delay spread for different enclosure materials.



(a) (b)

Figure C.11: Path loss and delay spread for different mold margins.

Table C.5: Channel models of wirebond in the frequency domain.

Frequency	Die-side	Si	AIN	Mold	PL_0	γ
60GHz	8	0.1	0.1	0.1	100.8	-2.1261
60GHz	8	0.1	0.1	0.05	99.4	-1.6974
60GHz	8	0.1	0.5	0.1	107.7	-2.2400
60GHz	8	0.1	0.5	0.05	109.3	-2.5698
60GHz	8	0.5	0.1	0.1	112.9	-0.2212
60GHz	8	0.5	0.1	0.05	114.9	-0.5495
60GHz	8	0.5	0.1	0.5	104.4	-2.8302
60GHz	8	0.5	0.5	0.1	108.3	-1.9770
60GHz	8	0.5	0.5	0.05	108.5	-1.7985

Frequency	Die-side	Si	AIN	Mold	Enclosure	PL_0	γ
60GHz	8	0.1	0.1	0.1	PEC	100.2	-2.279
60GHz	8	0.1	0.1	0.05	PEC	98.64	-1.6084
60GHz	8	0.1	0.1	0.5	PEC	102.1	-2.5356
60GHz	8	0.1	0.5	0.1	PEC	111.3	-2.7069
60GHz	8	0.1	0.5	0.05	PEC	102.5	-1.2904
60GHz	8	0.5	0.1	0.1	PEC	112.1	-2.4114
60GHz	8	0.5	0.1	0.05	PEC	110.3	-1.6998
60GHz	8	0.5	0.5	0.1	PEC	108.8	-2.0726
60GHz	8	0.5	0.5	0.05	PEC	108.6	-2.1842

Frequency	Die-side	Si	AIN	Mold	PL_0	γ
60GHz	12	0.1	0.1	0.1	109.2	-2.3959
60GHz	16	0.1	0.1	0.1	109	-1.8517
60GHz	20	0.1	0.1	0.1	102.7	-0.3921
60GHz	12	0.1	0.5	0.1	113.6	-2.0932
60GHz	16	0.1	0.5	0.1	115.2	-1.6186
60GHz	20	0.1	0.5	0.1	116.7	-1.2912
60GHz	12	0.5	0.1	0.1	112.9	-0.0194
60GHz	16	0.5	0.1	0.1	111.1	0.6454
60GHz	20	0.5	0.1	0.1	115.2	0.4642
60GHz	12	0.5	0.5	0.1	109.1	-0.8993
60GHz	16	0.5	0.5	0.1	107.8	0.0582
60GHz	20	0.5	0.5	0.1	103.4	0.6274

Frequency	Die-side	Si	AIN	Mold	BondWires	PL_0	γ
60GHz	8	0.1	0.1	0.1	64	118.8	-4.3933
60GHz	8	0.1	0.5	0.1	64	121.11	-3.5367
60GHz	8	0.1	0.5	0.1	128	127.3	-3.2772

Frequency	Die-side	Si	AIN	Mold	PL_0	γ
120GHz	8	0.1	0.5	0.1	114.1	n/a
180GHz	8	0.1	0.5	0.1	128.2	-2.9172
240GHz	8	0.1	0.5	0.1	129.9	-3.7184

Table C.6: Channel models of wirebond in the time domain.

Die-side	Si	AIN	Mold	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
8	0.1	0.1	0.1	0.05938	0.0128	0.1322	7.56
8	0.1	0.1	0.05	0.05495	0.0129	0.1296	7.71
8	0.1	0.1	0.5	0.08449	0.0136	0.1535	6.51
8	0.1	0.5	0.1	0.09132	0.0067	0.1336	7.48
8	0.5	0.1	0.1	0.06472	0.0207	0.1962	5.09
8	0.5	0.5	0.1	0.08703	0.0220	0.1943	5.14

Die-side	Si	AIN	Mold	Enclosure	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
8	0.1	0.5	0.1	PEC	0.1907	0.0173	0.2882	3.47

Die-side	Si	AIN	Mold	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
12	0.1	0.5	0.1	0.07713	0.0095	0.1791	5.58
16	0.1	0.5	0.1	0.07832	0.0084	0.2071	4.82

Die-side	Si	AIN	Mold	Bondwires	τ_{rms}(2mm)	γ_t	τ_{rms}	B_C
8	0.1	0.5	0.1	64	0.07439	0.0052	0.1065	9.39
8	0.1	0.5	0.1	128	0.06537	0.0069	0.1016	9.84

D. MAC Protocol Simulations and Models

In this appendix, we gather all the results from simulations of BRS, Token, and FUZZY TOKEN for the multiple traffic patterns and system sizes from $N = 16$ to $N = 1024$. Tables can be found in the subsequent pages.

Table D.1: Latency-throughput characteristic of the evaluated MAC protocols for 16 nodes and different workloads.

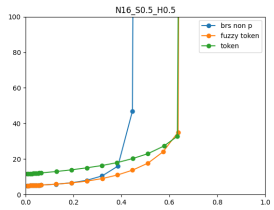
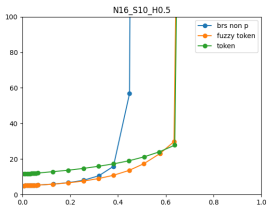
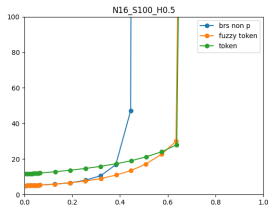
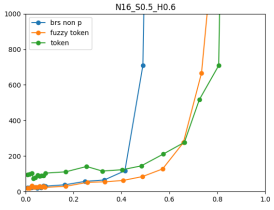
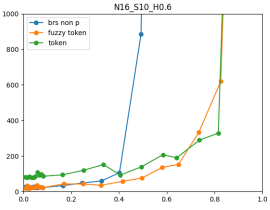
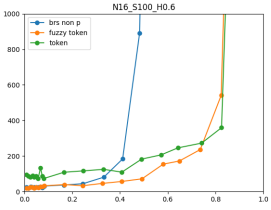
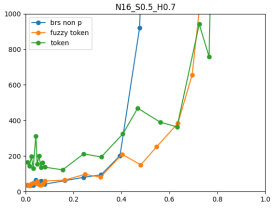
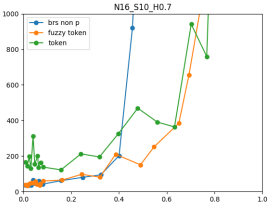
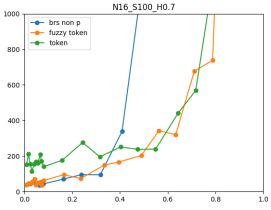
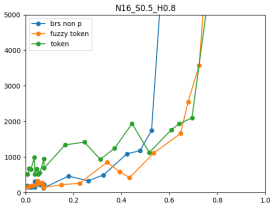
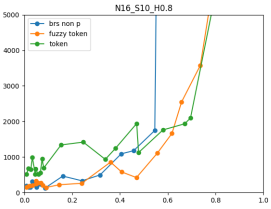
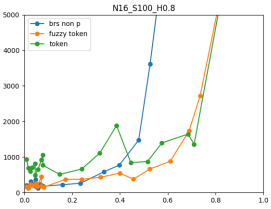
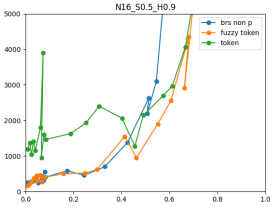
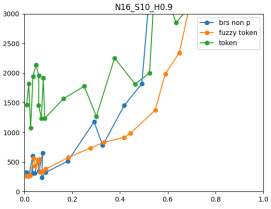
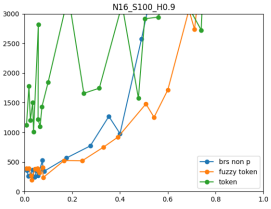
N	H	σ = 0.5	σ = 10	σ = 100
16	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.2: Latency-throughput characteristic of the evaluated MAC protocols for 32 nodes and different workloads.

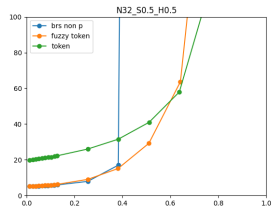
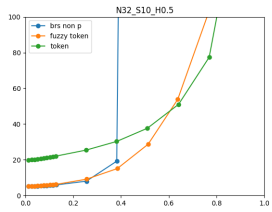
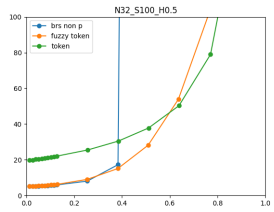
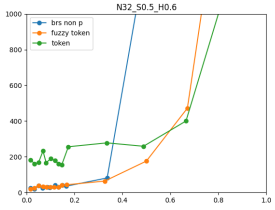
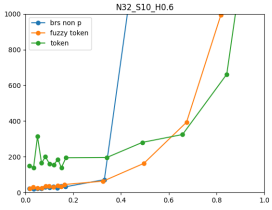
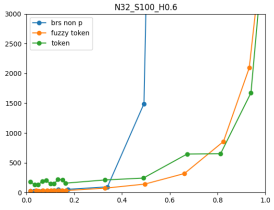
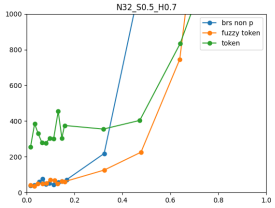
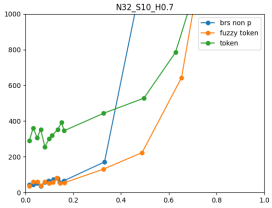
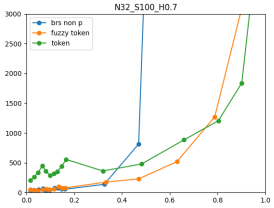
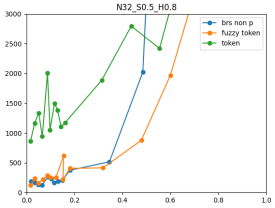
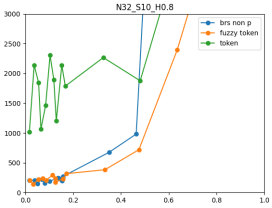
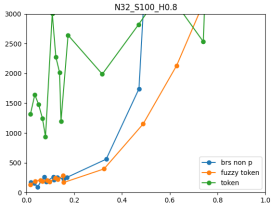
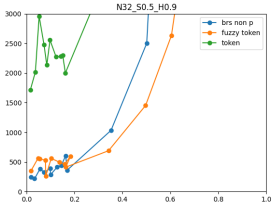
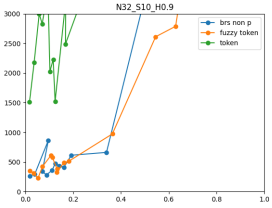
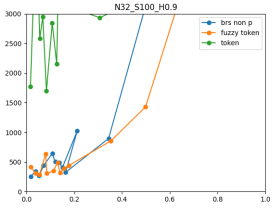
N	H	σ = 0.5	σ = 10	σ = 100
32	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.3: Latency-throughput characteristic of the evaluated MAC protocols for 64 nodes and different workloads.

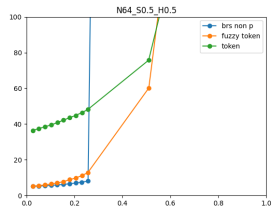
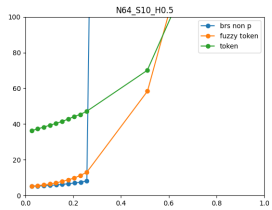
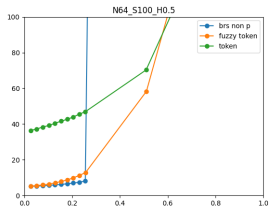
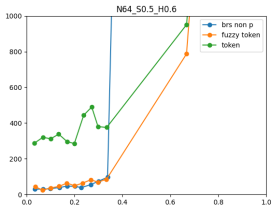
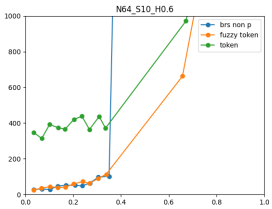
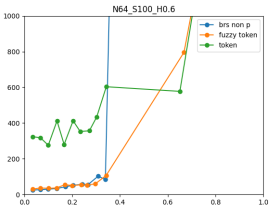
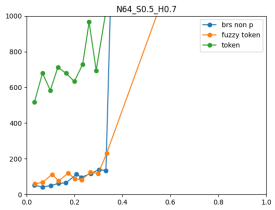
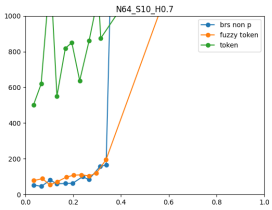
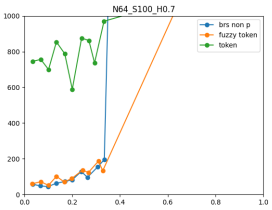
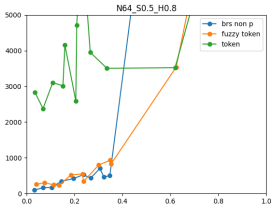
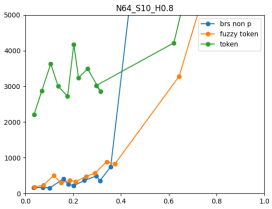
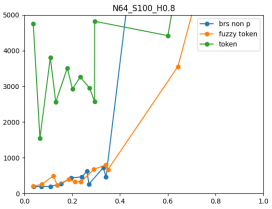
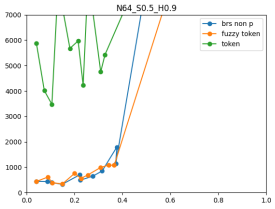
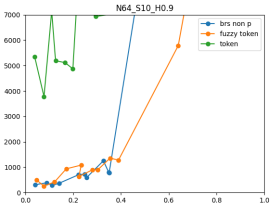
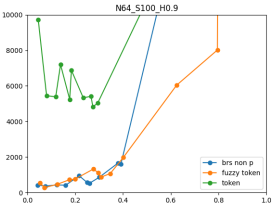
N	H	σ = 0.5	σ = 10	σ = 100
64	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.4: Latency-throughput characteristic of the evaluated MAC protocols for 128 nodes and different workloads.

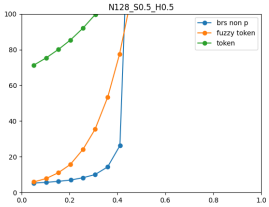
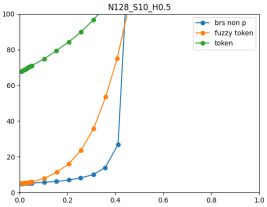
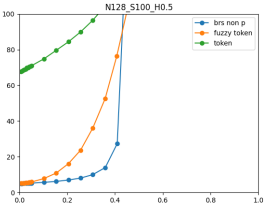
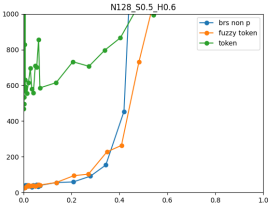
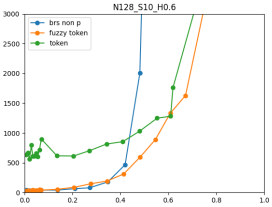
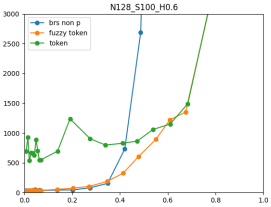
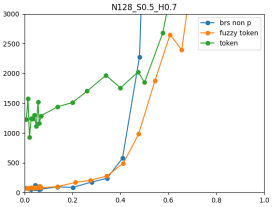
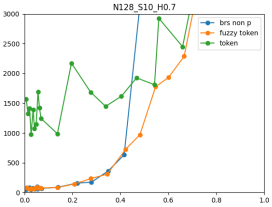
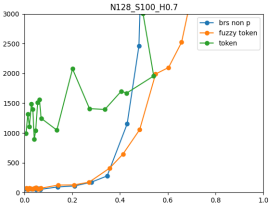
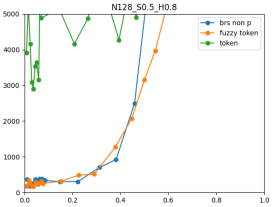
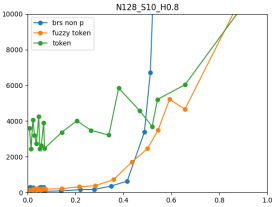
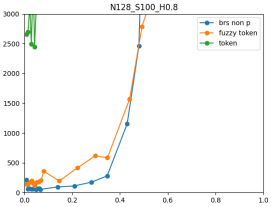
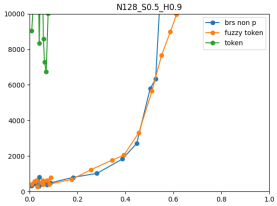
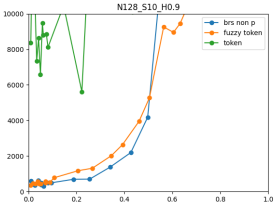
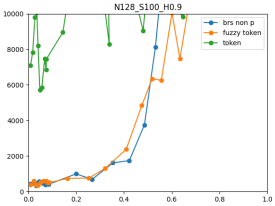
N	H	σ = 0.5	σ = 10	σ = 100
128	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.5: Latency-throughput characteristic of the evaluated MAC protocols for 256 nodes and different workloads.

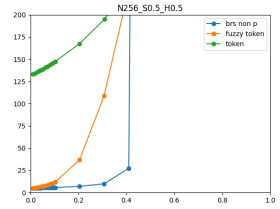
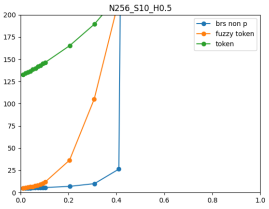
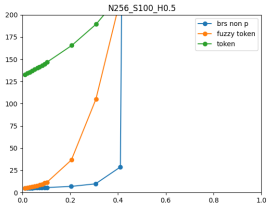
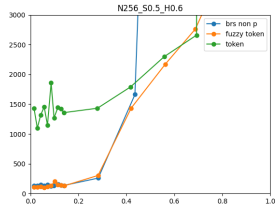
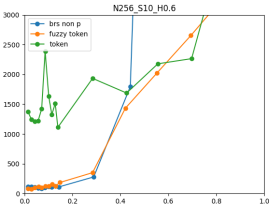
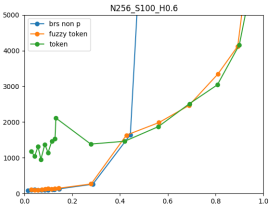
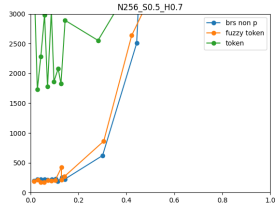
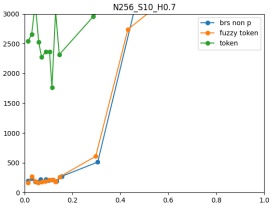
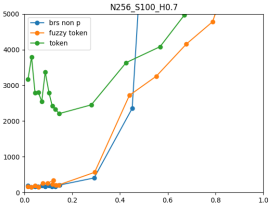
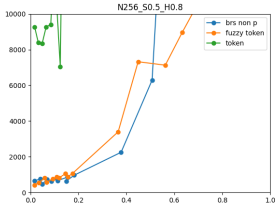
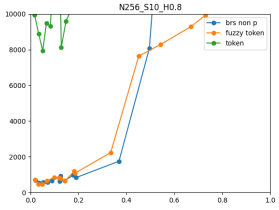
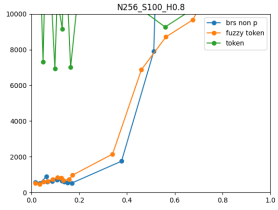
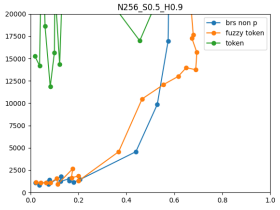
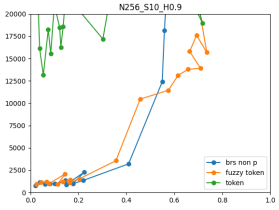
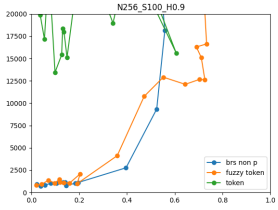
N	H	σ = 0.5	σ = 10	σ = 100
256	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.6: Latency-throughput characteristic of the evaluated MAC protocols for 512 nodes and different workloads.

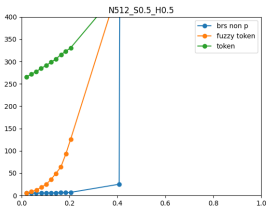
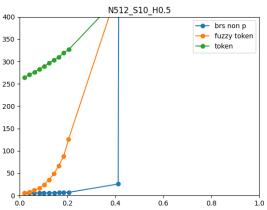
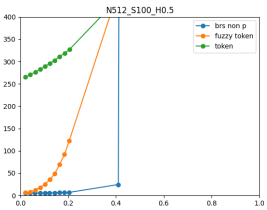
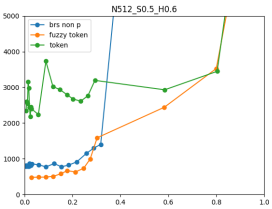
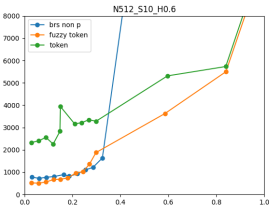
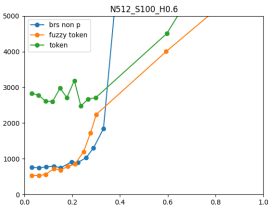
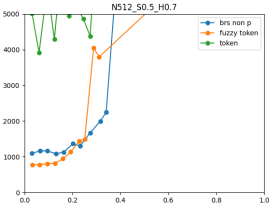
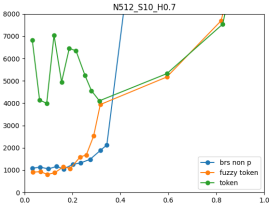
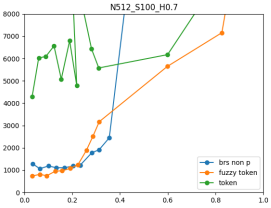
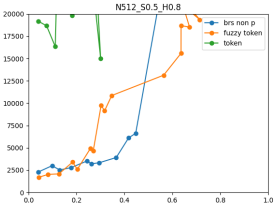
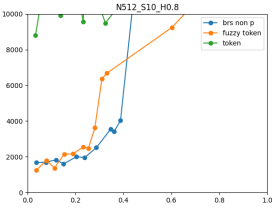
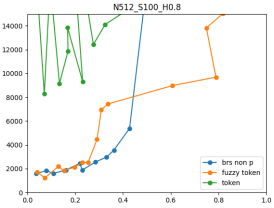
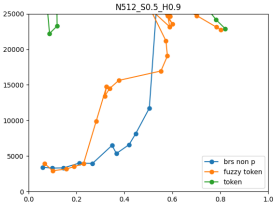
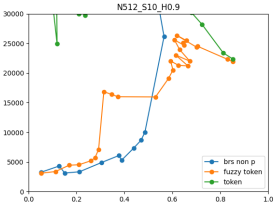
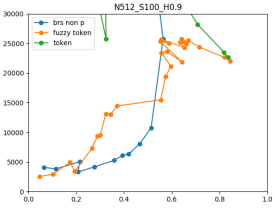
N	H	σ = 0.5	σ = 10	σ = 100
512	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.7: Latency-throughput characteristic of the evaluated MAC protocols for 1024 nodes and different workloads.

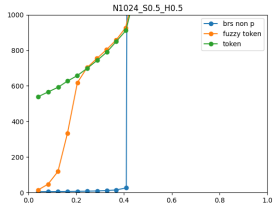
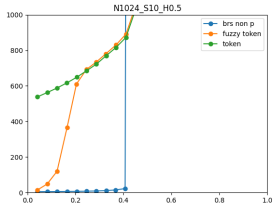
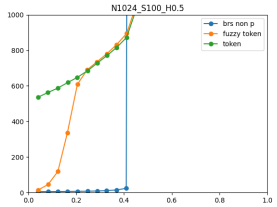
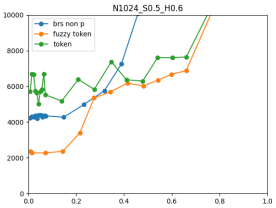
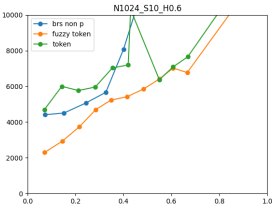
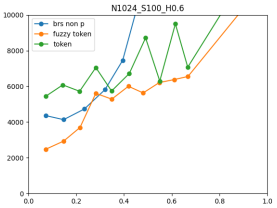
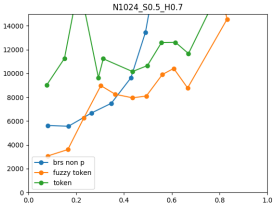
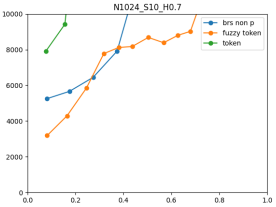
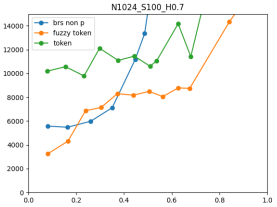
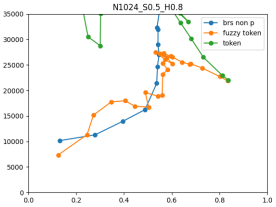
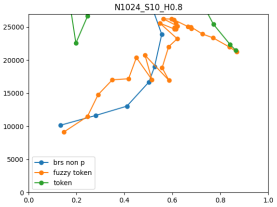
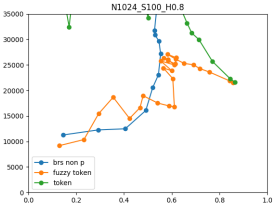
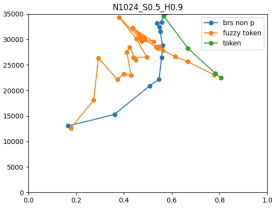
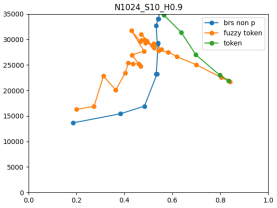
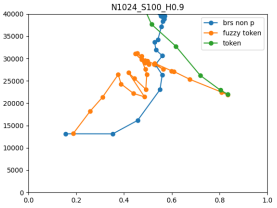
N	H	σ = 0.5	σ = 10	σ = 100
1024	0.5			
	0.6			
	0.7			
	0.8			
	0.9			

Table D.8: BRS model parameters for different system sizes and workloads.

BRS, 16 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.39	0.38	0.38
	α	-3.3	-3.2	-3.8
	β	60.6	60.3	63
	τ_{ZL}	5.2	5.2	5.2
0.6	λ_{sat}	0.41	0.4	0.39
	α	-10.9	-86.5	-385.3
	β	532.3	676.4	1513.8
	τ_{ZL}	24	26.5	34.2
0.7	λ_{sat}	0.36	0.3	0.3
	α	-181	-180.1	-593.4
	α	1327	1279	2888
	τ_{ZL}	49.2	49.5	63.4
0.8	λ_{sat}	0.15	0.15	0.18
	α	-777	-895.6	-1213.3
	β	6185	6893.3	6420
	τ_{ZL}	226	232.5	246.7
0.9	λ_{sat}	0.1	0.1	0.05
	α	-209	402.4	1837.6
	β	5656	5116.3	965.5
	τ_{ZL}	324.2	353	248.5

BRS, 32 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.39	0.39	0.39
	α	-16	2	-16
	β	115.3	37.6	117.1
	τ_{ZL}	5.8	5.1	5.6
0.6	λ_{sat}	0.37	0.37	0.33
	α	19.9	14.9	-430.7
	β	434.5	415.3	4045.3
	τ_{ZL}	25.2	21.2	30.6
0.7	λ_{sat}	0.37	0.2	0.3
	α	-310.8	38.8	-3.1
	β	2475	1039.4	1853.2
	τ_{ZL}	57	43.1	41.2
0.8	λ_{sat}	0.15	0.15	0.18
	α	389.4	-663.5	427.7
	β	2073	5725.6	3873.6
	τ_{ZL}	142	206	137.8
0.9	λ_{sat}	0.1	0.1	0.01
	α	812	2256.3	5982.8
	β	4067.2	-3056.4	-21993
	τ_{ZL}	237.9	239	123.9

BRS, 64 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.35	0.28	0.28
	α	1.1	0.5	0.27
	β	40	42.6	43.5
	τ_{ZL}	5.1	5.1	5.1
0.6	λ_{sat}	0.35	0.2	0.2
	α	-153.3	-39.1	-6.4
	β	476.7	711.6	622
	τ_{ZL}	34	30	24.3
0.7	λ_{sat}	0.35	0.2	0.2
	α	176.2	-360	-93.6
	β	476.7	1892.8	1274.6
	τ_{ZL}	34	70.4	51.8
0.8	λ_{sat}	0.3	0.15	0.18
	α	-1025.7	-584.7	-332.4
	β	17174	5158.7	6773.8
	τ_{ZL}	120.5	196	181.6
0.9	λ_{sat}	0.2	0.1	0.05
	α	1060.5	1660	172.5
	β	-11620.6	1477.1	5516.2
	τ_{ZL}	421.8	182.5	356.2

BRS, 128 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.4	0.4	0.4
	α	-4.7	4.8	6.3
	β	16.2	12.9	-17
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.45	0.45	0.45
	α	650	-102.6	-117.1
	β	-10739	1168.8	892
	τ_{ZL}	29.7	33.6	39.2
0.7	λ_{sat}	0.4	0.4	0.4
	α	188.1	-77.9	-72.1
	β	-369.8	2187	1662
	τ_{ZL}	65.8	66.4	60.6
0.8	λ_{sat}	0.2	0.2	0.2
	α	942	-702.8	-847.8
	β	-3804.2	-7483.7	3943.3
	τ_{ZL}	278.8	272.7	104
0.9	λ_{sat}	0.2	0.2	0.2
	α	937.7	-93119	84
	β	4799.6	219054	7602.8
	τ_{ZL}	400	4773	427.9

Table D.9: BRS model parameters for different system sizes and workloads (cont.).

BRS, 256 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.42	0.42	0.41
	α	4	4.5	4.6
	β	25	21.1	19
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.4	0.3	0.25
	α	124.7	-1234	-155.1
	β	-447	9165	2055
	τ_{ZL}	131	136	100
0.7	λ_{sat}	0.2	0.15	0.17
	α	232.5	-145	-365.9
	β	-286	1727	3055
	τ_{ZL}	202	205	182
0.8	λ_{sat}	0.1	0.01	0.1
	α	-580.7	-1333	6347
	β	8543	16276	-37356
	τ_{ZL}	656	630	435
0.9	λ_{sat}	0.01	0.01	0.05
	α	7511.3	-7552	4483
	β	-13210.6	54802	-16849
	τ_{ZL}	761	1146	724

BRS, 512 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.4	0.4	0.4
	α	2.7	2	3
	β	32	33.7	28
	τ_{ZL}	5	5	5.1
0.6	λ_{sat}	0.28	0.25	0.23
	α	1833	1832	927
	β	-120325	-120325	-4878
	τ_{ZL}	792	792	729
0.7	λ_{sat}	0.1	0.01	0.01
	α	-2740	1665	1665
	β	16125	-9033.8	-9033.9
	τ_{ZL}	1215	1041	1041
0.8	λ_{sat}	0.01	0.1	0.01
	α	3517	-3265	-3265.9
	β	2675	19855	19855
	τ_{ZL}	2246	1805	1805
0.9	λ_{sat}	0.01	0.01	0.01
	α	-1411	-1411.3	-7764
	β	15015	15015	35320
	τ_{ZL}	3374	3374	3843

BRS, 1024 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.4	0.4	0.4
	α	3.44379	-5944	-245388
	β	28	440835	504967
	τ_{ZL}	5	12864	26350
0.6	λ_{sat}	0.01	0.01	0.01
	α	927	-245388	-238865
	β	-4878	504967	487224
	τ_{ZL}	729	26350	27201
0.7	λ_{sat}	0.01	0.01	0.01
	α	-3426	-238865	-203322
	β	12753	487224	393471
	τ_{ZL}	1337	27201	32088
0.8	λ_{sat}	0.01	0.01	0.01
	α	-189383	-203322	-314655
	β	368984	393471	517459
	τ_{ZL}	30584	32088	54911
0.9	λ_{sat}	0.01	0.01	0.01
	α	-105281	-314655	-181995
	β	234206	517459	444812
	τ_{ZL}	23731	54911	13100

Table D.10: Token model parameters for different system sizes and workloads.

Token, 16 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.65	0.68	0.68
	α	6	-3	6
	β	28	60	21
	τ_{ZL}	11	5	11
0.6	λ_{sat}	0.5	0.65	0.65
	α	-1	18	-34
	β	308	259	401
	τ_{ZL}	91	86	90
0.7	λ_{sat}	0.35	0.4	0.4
	α	-169	-200	386
	β	1202	1259	-398
	τ_{ZL}	175	176	148
0.8	λ_{sat}	0.01	0.1	0.15
	α	4361	3364	2014
	β	-5307	-3196	-2327
	τ_{ZL}	512	554	648
0.9	λ_{sat}	0.01	n/a	n/a
	α	6170	-529	6124
	β	-10795	2453	-7813
	τ_{ZL}	1276	1638	1261

Token, 32 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.6	0.6	0.23
	α	-88	16	0.2
	β	252	25	70
	τ_{ZL}	25	19	20
0.6	λ_{sat}	0.01	0.01	0.015
	α	-52	-173	-602
	β	1074	528	6049
	τ_{ZL}	177	190	169
0.7	λ_{sat}	n/a	0.01	0.01
	α	n/a	122	5581
	β	n/a	942	-32791
	τ_{ZL}	n/a	306	133
0.8	λ_{sat}	n/a	n/a	0.01
	α	n/a	n/a	-1987
	β	n/a	n/a	71157
	τ_{ZL}	n/a	n/a	1352
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Token, 64 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.4	0.4	0.4
	α	32	31	30
	β	62	52	53
	τ_{ZL}	35	35	35
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Token, 128 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.2	0.01	0.01
	α	70	80	-85
	β	-1	-135	-239
	τ_{ZL}	67	67	64
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Table D.11: Token model parameters for different system sizes and workloads (cont.)

Token, 256 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.01	0.01	0.01
	α	132	130	126
	β	200	177	196
	τ_{ZL}	131	130	131
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

512 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.01	0.01	0.01
	α	266	245	236
	β	359	517	429
	τ_{ZL}	259	259	260
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Token, 1024 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.01	0.01	0.01
	α	394	403	427
	β	1311	1065	1018
	τ_{ZL}	523	521	519
0.6	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.7	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.8	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a
0.9	λ_{sat}	n/a	n/a	n/a
	α	n/a	n/a	n/a
	β	n/a	n/a	n/a
	τ_{ZL}	n/a	n/a	n/a

Table D.12: Fuzzy token parameters for different system sizes and workloads.

Fuzzy Token, 16 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.65	0.68	0.68
	α	-2	-1	-1
	β	46	44	43
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.62	0.65	0.65
	α	-16	-83	-115
	β	315	420	511
	τ_{ZL}	25	28	29
0.7	λ_{sat}	0.35	0.4	0.4
	α	68	109	-68
	β	530	442	946
	τ_{ZL}	40	38	51
0.8	λ_{sat}	0.3	0.23	0.18
	α	-36	-14	965
	β	2829	2495	-382
	τ_{ZL}	40	209	170
0.9	λ_{sat}	0.1	0.1	0.05
	α	70	879	321
	β	4737	1604	3128
	τ_{ZL}	310	341	314

Fuzzy Token, 32 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.65	0.68	0.68
	α	291	2.9	-43
	β	658	50	177
	τ_{ZL}	22	5	7
0.6	λ_{sat}	0.62	0.65	0.65
	α	78	108	286
	β	132	64	-1509
	τ_{ZL}	23	21	21
0.7	λ_{sat}	0.35	0.4	0.4
	α	102	-41	-748
	β	482	866	7284
	τ_{ZL}	39	50	60
0.8	λ_{sat}	0.3	0.23	0.18
	α	2672	511	1325
	β	-4965	379	-4301
	τ_{ZL}	77	171	126
0.9	λ_{sat}	0.1	0.1	0.05
	α	79	775	-544
	β	1312	2812	9254
	τ_{ZL}	446	309	375

Fuzzy Token, 64 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.5	0.5	0.5
	α	-1	-2	-2
	β	120	123	123
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.35	0.2	0.2
	α	92	-56	-128
	β	227	819	836
	τ_{ZL}	29	32	39
0.7	λ_{sat}	0.35	0.2	0.2
	α	-348	-388	-63
	β	1925	1870	1205
	τ_{ZL}	89	93	62
0.8	λ_{sat}	0.3	0.15	0.18
	α	3268	-188	3916
	β	-21099	5028	-12427
	τ_{ZL}	161	240	68
0.9	λ_{sat}	0.2	0.1	0.05
	α	6597	2184	-3506
	β	-4100	1374	22912
	τ_{ZL}	252	263	561

Fuzzy Token, 128 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.4	0.4	0.4
	α	6	4	6
	β	221	12	-16
	τ_{ZL}	5	4	4
0.6	λ_{sat}	0.45	0.45	0.45
	α	252	151	73
	β	-410	589	689
	τ_{ZL}	28	28	29
0.7	λ_{sat}	0.4	0.4	0.4
	α	56	-90	171
	β	1715	2009	923
	τ_{ZL}	74	70	59
0.8	λ_{sat}	0.2	0.2	0.2
	α	97	2165	972
	β	4299	-18043	1303
	τ_{ZL}	233	135	132
0.9	λ_{sat}	0.2	0.2	0.2
	α	106	1744	-2261
	β	10241	8255	15687
	τ_{ZL}	485	387	543

Table D.13: Fuzzy token parameters for different system sizes and workloads (cont.)

Fuzzy Token, 256 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.25	0.25	0.25
	α	6	3.77	10
	β	571	593	520
	τ_{ZL}	5	5	5
0.6	λ_{sat}	0.2	0.2	0.25
	α	1006	233	-26
	β	-3936	3162	2823
	τ_{ZL}	81	79	98
0.7	λ_{sat}	0.2	0.15	0.17
	α	-1468	-821	382
	β	1513	6810	7217
	τ_{ZL}	218	215	148
0.8	λ_{sat}	0.1	0.1	0.1
	α	6184	3142	6006
	β	-14870	-8602	-21973
	τ_{ZL}	349	490	360
0.9	λ_{sat}	0.01	0.01	0.05
	α	-11806	-1620	13381
	β	92880	31607	-65840
	τ_{ZL}	1419	1026	567

Fuzzy Token, 512 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.2	0.2	0.2
	α	-357	-202	-179
	β	4258	3365	3252
	τ_{ZL}	15	10	9
0.6	λ_{sat}	0.1	0.1	0.1
	α	-1816	36	568
	β	14370	8956	5934
	τ_{ZL}	529	496	495
0.7	λ_{sat}	0.01	0.01	0.01
	α	-3160	-7700	-137
	β	23175	46786	10515
	τ_{ZL}	870	1150	739
0.8	λ_{sat}	0.01	0.01	0.01
	α	-10357	4654	1745
	β	79312	5178	11065
	τ_{ZL}	2138	1133	1429
0.9	λ_{sat}	0.01	0.01	0.01
	α	-36158	8558	-21068
	β	127565	6086	138691
	τ_{ZL}	5610	2565	3408

Fuzzy Token, 1024 Nodes				
H	Par.	σ		
		0.5	10	100
0.5	λ_{sat}	0.2	0.2	0.2
	α	2825	3225	-779
	β	2086	635	12472
	τ_{ZL}	-162	-180	25
0.6	λ_{sat}	0.15	0.15	0.15
	α	-2743	15043	-1626
	β	40395	-10946	35072
	τ_{ZL}	2273	1152	2412
0.7	λ_{sat}	0.01	0.01	0.01
	α	43693	2752	23158
	β	-51954	41439	-11256
	τ_{ZL}	-812	2695	1269
0.8	λ_{sat}	0.01	0.01	0.01
	α	96417	-5349	-47810
	β	-111019	92486	190557
	τ_{ZL}	-3470	7826	12009
0.9	λ_{sat}	0.01	0.01	0.01
	α	215759	119395	55623
	β	-281913	-157824	24399
	τ_{ZL}	-16785	-1690	1982

Bibliography

[1] WiPLASH consortium, "Wireless Channel Modeling," in *European Commission, H2020-FETOPEN, Project WiPLASH: Accepted Public Deliverable D3.1, 17-Jan-2021*, 2021.

[2] B. Murmann, "ADC Performance Survey 1997-2021," 2021.

[3] H. Wang, T.-Y. Huang, N. Sasikanth, *et al.*, "Power Amplifiers Performance Survey 2000-2021," 2021.

[4] *Thin & Light & High Performance Graphics*, Hot Chips 2018 (accessed August 12, 2021).

[5] R. Guirado, H. Kwon, S. Abadal, E. Alarcón, and T. Krishna, "Dataflow-architecture co-design for 2.5D DNN accelerators using wireless network-on-package," in *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 806–812, IEEE, 2021.

[6] F. Lemic, S. Abadal, W. Tavernier, P. Stroobant, D. Colle, E. Alarcón, J. Marquez-Barja, and J. Famaey, "Survey on terahertz nanocommunication and networking: A top-down perspective," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1506–1543, 2021.

[7] X. Yu, J. Baylon, P. Wettin, D. Heo, P. Pratim Pande, and S. Mirabbasi, "Architecture and Design of Multi-Channel Millimeter-Wave Wireless Network-on-Chip," *IEEE Design & Test*, vol. 31, no. 6, pp. 19–28, 2014.

[8] K. K. Tokgoz, S. Maki, J. Pang, N. Nagashima, I. Abdo, S. Kawai, T. Fujimura, Y. Kawano, T. Suzuki, T. Iwai, K. Okada, and A. Matsuzawa, "A 120Gb/s 16QAM CMOS millimeter-wave wireless transceiver," *Proceedings of the ISSCC '18*, pp. 168–170, 2018.

[9] R. Marculescu, U. Ogras, L.-S. Peh, N. Enright Jerger, and Y. Hoskote, "Outstanding research problems in NoC design: system, microarchitecture, and circuit perspectives," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 1, pp. 3–21, 2009.

[10] D. Bertozzi, G. Dimitrakopoulos, J. Flich, and S. Sonntag, "The fast evolving landscape of on-chip communication," *Design Automation for Embedded Systems*, vol. 19, no. 1, pp. 59–76, 2015.

[11] S. Abadal, R. Guirado, H. Taghvaee, A. Jain, E. P. de Santana, P. Haring Bolívar, M. Saeed, R. Negra, Z. Wang, K.-T. Wang, *et al.*, "Graphene-based wireless agile interconnects for massive heterogeneous multi-chip processors," *arXiv preprint arXiv:2011.04107*, 2020.

[12] J. Kim, K. Choi, and G. Loh, "Exploiting new interconnect technologies in on-chip communication," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 2, pp. 124–136, 2012.

[13] D. Matolak, A. Kodi, S. Kaya, D. DiTomaso, S. Laha, and W. Rayess, "Wireless networks-on-chips: architecture, wireless channel, and devices," *IEEE Wireless Communications*, vol. 19, no. 5, 2012.

[14] R. S. Narde, J. Venkataraman, A. Ganguly, and I. Puchades, "Intra-and Inter-Chip Transmission of Millimeter-Wave Interconnects in NoC-based Multi-Chip Systems," *IEEE Access*, vol. 7, pp. 112200–15, 2019.

[15] S. Abadal, B. Sheinman, O. Katz, O. Markish, D. Elad, Y. Fournier, D. Roca, M. Hanzich, G. Houzeaux, M. Nemirovsky, E. Alarcón, and A. Cabellos-Aparicio, "Broadcast-Enabled Massive Multicore Architectures: A Wireless RF Approach," *IEEE MICRO*, vol. 35, no. 5, pp. 52–61, 2015.

[16] R. G. Kim, W. Choi, Z. Chen, P. P. Pande, D. Marculescu, and R. Marculescu, "Wireless NoC and Dynamic VFI Codesign: Energy Efficiency Without Performance Penalty," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 7, pp. 2488–2501, 2016.

[17] M. A. I. Sikder, A. Kodi, W. Rayess, D. Ditomaso, D. Matolak, and S. Kaya, "Exploring wireless technology for off-chip memory access," in *Proceedings of the HOTI '16*, pp. 92–99, 2016.

[18] S. Abadal, E. Alarcón, A. Cabellos-Aparicio, and J. Torrellas, "WiSync: An Architecture for Fast Synchronization through On-Chip Wireless Communication," in *Proceedings of the ASPLOS '16*, pp. 3–17, 2016.

[19] V. Fernando, A. Franques, S. Abadal, S. Misailovic, and J. Torrellas, "Replica: A Wireless Manycore for Communication-Intensive and Approximate Data," in *Proceedings of the ASPLOS '19*, pp. 849–863, 2019.

[20] A. Franques, A. Kokolis, S. Abadal, V. Fernando, S. Misailovic, and J. Torrellas, "WiDir: A Wireless-Enabled Directory Cache Coherence Protocol," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 304–317, IEEE, 2021.

[21] S. Abadal, M. Iannazzo, M. Nemirovsky, A. Cabellos-Aparicio, H. Lee, and E. Alarcón, "On the area and energy scalability of wireless network-on-chip: A model-based benchmarked design space exploration," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1501–1513, 2015.

[22] D. DiTomaso, A. Kodi, S. Kaya, and D. Matolak, "iWISE: Inter-router wireless scalable express channels for network-on-chips (NoCs) architecture," in *Proc. of IEEE 19th Annu. Symp. High Perform. Interconnects*, pp. 11–18, 2011.

[23] D. DiTomaso, A. Kodi, D. Matolak, S. Kaya, S. Laha, and W. Rayess, "A-WiNoC: Adaptive Wireless Network-on-Chip Architecture for Chip Multiprocessors," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 12, pp. 3289–3302, 2015.

[24] K. Duraisamy, R. G. Kim, and P. P. Pande, "Enhancing Performance of Wireless NoCs with Distributed MAC Protocols," in *Proceedings of the ISQED '15*, pp. 406–11, 2015.

[25] N. Mansoor and A. Ganguly, "Reconfigurable Wireless Network-on-Chip with a Dynamic Medium Access Mechanism," in *Proceedings of the NoCS '15*, p. Article 13, 2015.

[26] A. Mestres, S. Abadal, J. Torrellas, E. Alarcón, and A. Cabellos-Aparicio, "A MAC protocol for Reliable Broadcast Communications in Wireless Network-on-Chip," in *Proceedings of the NoCArc '16*, pp. 21–26, 2016.

[27] H. Mondal, S. Gade, M. Shamim, S. Deb, and A. Ganguly, "Interference-Aware Wireless Network-on-Chip Architecture using Directional Antennas," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 3, no. 3, pp. 193–205, 2017.

[28] V. Soteriou, H. Wang, and L. Peh, "A Statistical Traffic Model for On-Chip Interconnection Networks," in *Proceedings of MASCOTS '06*, pp. 104–116, 2006.

[29] S. Abadal, A. Mestres, J. Torrellas, E. Alarcón, and A. Cabellos-Aparicio, "Medium access control in wireless network-on-chip: A context analysis," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 172–178, 2018.

[30] "CST Microwave Studio."

[31] X. Timoneda, S. Abadal, A. Cabellos-Aparicio, D. Manessis, J. Zhou, A. Franques, J. Torrellas, and E. Alarcón, "Millimeter-Wave Propagation within a Computer Chip Package," in *Proceedings of the ISCAS '18*, 2018.

[32] X. Timoneda, A. Cabellos-Aparicio, D. Manessis, E. Alarcón, and S. Abadal, "Channel Characterization for Chip-scale Wireless Communications within Computing Packages," in *Proceedings of the NOCS '18*, 2018.

[33] F. T. Chen, J. M. Wu, and M. C. F. Chang, "40-Gb/s 0.7-V 2:1 MUX and 1:2 DEMUX with Transformer-Coupled Technique for SerDes Interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 4, pp. 1042–1051, 2015.

[34] S. Saxena, G. Shu, R. K. Nandwana, M. Talegaonkar, A. Elkholy, T. Anand, W.-S. Choi, and P. K. Hanumolu, "A 2.8 mw/gb/s, 14 gb/s serial link transceiver," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 5, pp. 1399–1411, 2017.

[35] A. A. S. SH, K. S. Reddy, *et al.*, "A 20 gb/s latency optimized serdes transmitter for data centre applications," in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–4, IEEE, 2020.

[36] X. Yu, J. Baylon, P. Wettin, D. Heo, P. P. Pande, and S. Mirabbasi, "Architecture and design of multichannel millimeter-wave wireless NoC," *IEEE Design & Test*, vol. 31, no. 6, pp. 19–28, 2014.

[37] W. Bae, "State-of-the-art circuit techniques for low-jitter phase-locked loops: Advanced performance benchmark fom based on an extensive survey," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, 2021.

[38] M. Abdulaziz, T. Forsberg, M. Törmänen, and H. Sjöland, "A 10-mw mm-wave phase-locked loop with improved lock time in 28-nm fd-soi cmos," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 4, pp. 1588–1600, 2019.

[39] WiPLASH consortium, "SiGe BiCMOS test devices," in *European Commission, H2020-FETOPEN, Project WiPLASH: Submitted Public Deliverable D1.1, 30-Sep-2021*, 2021.

[40] A. S. Abd El-Hameed, A. Barakat, A. B. Abdel-Rahman, A. Allam, and R. K. Pokharel, "Ultracompact 60-ghz cmos bpf employing broadside-coupled open-loop resonators," *IEEE Microwave and Wireless Components Letters*, vol. 27, no. 9, pp. 818–820, 2017.

[41] M. G. Bautista, H. Zhu, X. Zhu, Y. Yang, Y. Sun, and E. Dutkiewicz, "Compact millimeter-wave bandpass filters using quasi-lumped elements in 0.13-um (bi)-cmos technology for 5g wireless systems," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 7, pp. 3064–3073, 2019.

[42] M. S. Khan, F. A. Tahir, and H. M. Cheema, "Design of bowtie-slot on-chip antenna backed with e-shaped fss at 94 ghz," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*, pp. 1–3, IEEE, 2016.

[43] J. Gorisse, D. Morche, and J. Jantunen, "Wireless transceivers for gigabit-per-second communications," in *10th IEEE International NEWCAS Conference*, pp. 545–548, IEEE, 2012.

[44] P.-Y. Chang, S.-H. Su, S. S. Hsu, W.-H. Cho, and J.-D. Jin, "An ultra-low-power transformer-feedback 60 ghz low-noise amplifier in 90 nm cmos," *IEEE Microwave and Wireless Components Letters*, vol. 22, no. 4, pp. 197–199, 2012.

[45] H. Li, J. Chen, D. Hou, Z. Li, P. Zhou, and W. Hong, "A 230-ghz sige amplifier with 21.8-db gain and 3-dbm output power for sub-thz receivers," *IEEE Microwave and Wireless Components Letters*, 2021.

[46] M. Najmussadat, R. Ahamed, M. Varonen, D. Parveg, Y. Tawfik, and K. A. Halonen, "Design of a 240-ghz lna in 0.13 μm sige bicmos technology," in *2020 15th European Microwave Integrated Circuits Conference (EuMIC)*, pp. 17–20, IEEE, 2021.

[47] J. Pang, S. Maki, S. Kawai, N. Nagashima, Y. Seo, M. Dome, H. Kato, M. Katsuragi, K. Kimura, S. Kondo, Y. Terashima, H. Liu, T. Siriburanon, A. T. Narayanan, N. Fajri, T. Kaneko, T. Yoshioka, B. Liu, Y. Wang, R. Wu, K. K. Tokgoz, M. Miyahara, A. Shirane, and K. Okada, "A 50.1-Gb/s 60-GHz CMOS Transceiver for IEEE 802.11ay With Calibration of LO Feedthrough and I/Q Imbalance," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 5, pp. 1375–1390, 2019.

[48] L. Kleinrock and F. Tobagi, "Packet Switching in Radio Channels: Part I–Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics," *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1400–1416, 1975.

[49] R. Ubal, P. Mistry, D. Schaa, H. Ave, and D. Kaeli, "Multi2Sim: A Simulation Framework for CPU-GPU Computing," in *Proceedings of the PACT'12*, 2012.

[50] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[51] K. Kimoto, N. Sasaki, S. Kubota, W. Moriyama, and T. Kikkawa, "High-Gain On-Chip Antennas for LSI Intra- / Inter-Chip Wireless Interconnection," *Proceedings of the EuCAP '09*, pp. 278–282, 2009.

[52] O. Markish, B. Sheinman, O. Katz, D. Corcos, and D. Elad, "On-chip mmWave Antennas and Transceivers," in *Proceedings of the NoCS '15*, p. Art. 11, 2015.

[53] N. Kim, D. Wu, D. Kim, A. Rahman, and P. Wu, "Interposer design optimization for high frequency signal transmission in passive and active interposer using through silicon via (tsv)," in *2011 IEEE 61st electronic components and technology conference (ECTC)*, pp. 1160–1167, IEEE, 2011.

[54] H. M. Cheema and A. Shamim, "The last barrier: On-chip antennas," *IEEE Microwave Magazine*, vol. 14, no. 1, pp. 79–91, 2013.

[55] T. O. Dickson, H. A. Ainspan, and M. Meghelli, "6.5 a 1.8 pj/b 56gb/s pam-4 transmitter with fractionally spaced ffe in 14nm cmos," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 118–119, IEEE, 2017.

[56] X. Timoneda, S. Abadal, A. Franques, D. Manessis, J. Zhou, J. Torrellas, E. Alarcón, and A. Cabellos-Aparicio, "Engineer the Channel and Adapt to it: Enabling Wireless Intra-Chip Communication," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 3247–3258, 2020.

[57] D. Clark, K. Pogran, and D. Reed, "An introduction to local area networks," *Proceedings of the IEEE*, vol. 66, no. 11, pp. 1497–1517, 1978.

[58] A. Franques, S. Abadal, H. Hassanieh, and J. Torrellas, "Fuzzy-Token: An adaptive mac protocol for wireless-enabled manycores," in *Design, Automation and Test in Europe Conference (DATE)*, 2021.

[59] N. Mansoor, A. Vashist, M. M. Ahmed, M. S. Shamim, S. A. Mamun, and A. Ganguly, "A traffic-aware medium access control mechanism for energy-efficient wireless network-on-chip architectures," *arXiv preprint arXiv:1809.07862*, 2018.

[60] S. Laha, S. Kaya, D. W. Matolak, W. Rayess, D. DiTomaso, and A. Kodi, "A New Frontier in Ultralow Power Wireless Links: Network-on-Chip and Chip-to-Chip Interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 2, pp. 186–198, 2015.

[61] D. Fritsche, P. Stärke, C. Carta, and F. Ellinger, "A Low-Power SiGe BiCMOS 190-GHz Transceiver Chipset With Demonstrated Data Rates up to 50 Gbit/s Using On-Chip Antennas," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 9, pp. 3312–3323, 2017.

[62] D. Culler, J. P. Singh, and A. Gupta, *Parallel computer architecture: a hardware/software approach*. Gulf Professional Publishing, 1999.

[63] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–33, 2016.

[64] S. Deb, A. Ganguly, P. P. Pande, B. Belzer, and D. Heo, "Wireless NoC as Interconnection Backbone for Multicore Chips: Promises and Challenges," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 2, pp. 228–239, 2012.

[65] V. Catania, A. Mineo, S. Monteleone, M. Palesi, and D. Patti, "Improving Energy Efficiency in Wireless Network-on-Chip Architectures," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 14, no. 1, p. Art. 9, 2018.

[66] D. Matolak, S. Kaya, and A. Kodi, "Channel modeling for wireless networks-on-chips," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 180–186, 2013.

[67] S. Abadal, R. Martínez, J. Solé-Pareta, E. Alarcón, and A. Cabellos-Aparicio, "Characterization and Modeling of Multicast Communication in Cache-Coherent Manycore Processors," *Computers and Electrical Engineering (Elsevier)*, vol. 51, no. April, pp. 168–183, 2016.

[68] T. Sherwood, E. Perelman, G. Hamerly, S. Sair, and B. Calder, "Discovering and Exploiting Program Phases," *IEEE Micro*, vol. 23, no. 6, pp. 84–93, 2003.

[69] S. Redfield, S. Woracheewan, H. Liu, P. Chiang, J. Nejedlo, and R. Khanna, "Understanding the ultrawideband channel characteristics within a computer chassis," *IEEE Antennas and Wireless Propagation Letters*, vol. 10, pp. 191–194, 2011.

[70] C.-X. Wang, J. Bian, J. Sun, W. Zhang, and M. Zhang, "A survey of 5g channel measurements and models," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3142–3168, 2018.

[71] J. Fu, P. Juyal, and A. Zajić, "Thz channel characterization of chip-to-chip communication in desktop size metal enclosure," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 12, pp. 7550–7560, 2019.

[72] S. Abadal, C. Han, and J. M. Jornet, "Wave propagation and channel modeling in chip-scale wireless communications: A survey from millimeter-wave to terahertz and optics," *IEEE access*, vol. 8, pp. 278–293, 2019.

[73] C. A. Balanis, *Antenna theory: analysis and design*. John wiley & sons, 2015.

[74] C. Yi, D. Kim, S. Solanki, J. H. Kwon, M. Kim, S. Jeon, Y. C. Ko, and I. Lee, "Design and Performance Analysis of THz Wireless Communication Systems for Chip-to-Chip and Personal Area Networks Applications," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1785–1796, 2021.

[75] S.-B. Lee, S.-W. Tam, I. Pefkianakis, S. Lu, M.-C. F. Chang, C. Guo, G. Reinman, C. Peng, M. Naik, L. Zhang, and J. Cong, "A scalable micro wireless interconnect structure for CMPs," in *Proceedings of the MOBICOM '09*, p. 217, 2009.

[76] N. Mansoor, S. Shamim, and A. Ganguly, "A Demand-Aware Predictive Dynamic Bandwidth Allocation Mechanism for Wireless Network-on-Chip," in *Proceedings of the SLIP '16*, 2016.

[77] S. Abadal, J. Torrellas, E. Alarcón, and A. Cabellos-Aparicio, "OrthoNoC: A Broadcast-Oriented Dual-Plane Wireless Network-on-Chip Architecture," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 3, pp. 628–641, 2018.

[78] V. Kapsalis, S. Koubias, and H. Haralabidis, "New hybrid mac-layer protocol for real-time bus networks," *IEE Proceedings-Communications*, vol. 141, no. 5, pp. 325–333, 1994.

[79] Tsern-Huei Lee, "Performance analysis of a class of hybrid token-csma/cd protocols," in *Proceedings of the TENCON '91*, Aug 1991.

[80] A. Ephremides and O. Mowafi, "Analysis of a Hybrid Access Scheme for Buffered Users-Probabilistic Time Division," *IEEE Transactions on Software Engineering*, vol. SE-8, no. 1, pp. 52–61, 1982.

[81] "Waiting For Chiplet Interfaces."

[82] S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, and S. White, "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 2021.

[83] B. Yu, Y. Ye, X. Ding, Y. Liu, Z. Xu, X. Liu, and Q. J. Gu, "Ortho-mode sub-thz interconnect channel for planar chip-to-chip communications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 66, no. 4, pp. 1864–1873, 2017.

[84] J. W. Holloway, L. Boglione, T. M. Hancock, and R. Han, "A fully integrated broadband sub-mmwave chip-to-chip interconnect," *IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 7, pp. 2373–2386, 2017.

[85] J. W. Holloway, G. C. Dogiamis, and R. Han, "Innovations in terahertz interconnects: High-speed data transport over fully electrical terahertz waveguide links," *IEEE Microwave Magazine*, vol. 21, no. 1, pp. 35–50, 2020.

[86] M. Wade, E. Anderson, S. Ardalan, P. Bhargava, S. Buchbinder, M. L. Davenport, J. Fini, H. Lu, C. Li, R. Meade, M. Ramamurthy, Chandruand Rust, F. Sedgwick, V. Stojanovic, D. Van Orden, C. Zhang, C. Sun, S. Shumarayev, C. O'Keeffe, T. Hoang, D. Kehlet, R. Mahajan, M. Guzy, C. Allen, and T. Tran, "TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical I/O," *IEEE Micro*, vol. 40, no. 2, pp. 63–71, 2020.

[87] P. Fotouhi, S. Werner, J. Lowe-Power, and S. B. Yoo, "Enabling scalable chiplet-based uniform memory architectures with silicon photonics," in *Proceedings of the International Symposium on Memory Systems*, pp. 222–334, 2019.

[88] N. C. Abrams, Q. Cheng, M. Glick, M. Jezzini, P. Morrissey, P. O'Brien, and K. Bergman, "Silicon photonic 2.5 d multi-chip module transceiver for high-performance data centers," *Journal of Lightwave Technology*, vol. 38, no. 13, pp. 3346–3357, 2020.

[89] A. Arunkumar, E. Bolotin, B. Cho, U. Milic, E. Ebrahimi, O. Villa, A. Jaleel, C.-J. Wu, and D. Nellans, "MCM-GPU: Multi-chip-module GPUs for continued performance scalability," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 2, pp. 320–332, 2017.

[90] B. Zimmer, R. Venkatesan, Y. S. Shao, J. Clemons, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina, S. Tell, Y. Zhang, W. Dally, J. Emer, C. Gray, S. Keckler, and B. Khailany, "A 0.32–128 TOPS, scalable multi-chip-module-based deep neural network inference accelerator with ground-referenced signaling in 16 nm," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 920–932, 2020.

[91] *AMD Epyc 7742 Processor*, 2019 (accessed April 16, 2020).

[92] *AMD Infinity Architecture Technology*, 2019 (accessed July 12, 2021).

[93] Y. S. Shao, J. Clemons, R. Venkatesan, B. Zimmer, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina, S. Tell, Y. Zhang, W. Dally, J. Emer, C. Gray, B. Khailany, and S. Keckler, "Simba: Scaling deep-learning inference with multi-chip-module-based architecture," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 14–27, 2019.

[94] J. Yin, Z. Lin, O. Kayiran, M. Poremba, M. S. B. Altaf, N. E. Jerger, and G. H. Loh, "Modular routing design for chiplet-based systems," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pp. 726–738, IEEE, 2018.

[95] A. Kannan, N. E. Jerger, and G. H. Loh, "Enabling interposer-based disintegration of multi-core processors," in *2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 546–558, IEEE, 2015.

[96] S. S. Iyer, "Heterogeneous integration for performance and scaling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 7, pp. 973–982, 2016.

[97] A. Kannan, N. E. Jerger, and G. H. Loh, "Exploiting interposer technologies to disintegrate and reintegrate multicore processors," *Ieee Micro*, vol. 36, no. 3, pp. 84–93, 2016.

[98] S. Bharadwaj, J. Yin, B. Beckmann, and T. Krishna, "Kite: a family of heterogeneous interposer topologies enabled via accurate interconnect modeling," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2020.

[99] D. Stow, I. Akgun, and Y. Xie, "Investigation of cost-optimal network-on-chip for passive and active interposer systems," in *2019 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, pp. 1–8, IEEE, 2019.

[100] J. Kim, G. Murali, H. Park, E. Qin, H. Kwon, V. C. K. Chekuri, N. M. Rahman, N. Dasari, A. Singh, M. Lee, H. Torun, K. Roy, M. Swaminathan, S. Mukhopadhyay, T. Krishna, and S. Lim, "Architecture, chip, and package codesign flow for interposer-based 2.5-D chiplet integration enabling heterogeneous IP reuse," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 11, pp. 2424–2437, 2020.

[101] C. Liu, J. Botimer, and Z. Zhang, "A 256Gb/s/mm-shoreline AIB-Compatible 16nm FinFET CMOS Chiplet for 2.5 D Integration with Stratix 10 FPGA on EMIB and Tiling on Silicon Interposer," in *2021 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–2, IEEE, 2021.

[102] *Intel Ponte Vecchio Xe-HPC GPGPU*, 2019 (accessed July 12, 2021).

[103] R. Mahajan, R. Sankman, N. Patel, D.-W. Kim, K. Aygun, Z. Qian, Y. Mekonnen, I. Salama, S. Sharan, D. Iyengar, and D. Mallik, "Embedded multi-die interconnect bridge (EMIB)–a high density, high bandwidth packaging interconnect," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pp. 557–565, IEEE, 2016.

[104] V. Pano, I. Tekin, I. Yilmaz, Y. Liu, K. R. Dandekar, and B. Taskin, "TSV antennas for multi-band wireless communication," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 10, no. 1, pp. 100–113, 2020.

[105] P. Vivet, E. Guthmuller, Y. Thonnart, G. Pillonnet, C. Fuguet, I. Miro-Panades, G. Moritz, J. Durupt, C. Bernard, D. Varreau, *et al.*, "Intact: A 96-core processor with six chiplets 3d-stacked on an active interposer with distributed interconnects and integrated power management," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 79–97, 2020.

[106] P.-H. Chang, C.-L. Chung, Y.-Y. Hsu, and C.-F. Chiang, "Signal and power integrity analysis of a 0.38 pj/bit 12.8 gb/s parallel interface for die-to-die link applications," in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, pp. 1264–1269, IEEE, 2021.

[107] M. W. Chaudhary, A. Heinig, and B. Choubey, "13-gb/s transmitter for bunch of wires chip-to-chip interface standard," in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 333–336, IEEE, 2020.

[108] Y.-Y. Hsu, P.-C. Kuo, C.-L. Chuang, P.-H. Chang, H.-H. Shen, and C.-F. Chiang, "A 7nm 0.46 pj/bit 20gbps with ber 1e-25 die-to-die link using minimum intrinsic auto alignment and noise-immunity encode," in *2021 Symposium on VLSI Technology*, pp. 1–2, IEEE, 2021.

[109] J. W. Poulton, J. M. Wilson, W. J. Turner, B. Zimmer, X. Chen, S. S. Kudva, S. Song, S. G. Tell, N. Nedovic, W. Zhao, *et al.*, "A 1.17-pj/b, 25-gb/s/pin ground-referenced single-ended serial link for off-and on-package communication using a process-and temperature-adaptive voltage regulator," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 43–54, 2018.

[110] M. Erett, D. Carey, J. Hudner, R. Casey, K. Geary, P. Neto, M. Raj, S. McLeod, H. Zhang, A. Roldan, *et al.*, "A 126mw 56gb/s nrz wireline transceiver for synchronous short-reach applications in 16nm finfet," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 274–276, IEEE, 2018.

[111] L. Wang, Y. Fu, M.-A. LaCroix, E. Chong, and A. C. Carusone, "A 64-gb/s 4-pam transceiver utilizing an adaptive threshold adc in 16-nm finfet," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 2, pp. 452–462, 2018.

[112] P. Upadhyaya, C. F. Poon, S. W. Lim, J. Cho, A. Roldan, W. Zhang, J. Namkoong, T. Pham, B. Xu, W. Lin, *et al.*, "A fully adaptive 19-to-56gb/s pam-4 wireline transceiver with a configurable adc in 16nm finfet," in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*, pp. 108–110, IEEE, 2018.

[113] Y. Frans, J. Shin, L. Zhou, P. Upadhyaya, J. Im, V. Kireev, M. Elzeftawi, H. Hedayati, T. Pham, S. Asuncion, *et al.*, "A 56-gb/s pam4 wireline transceiver using a 32-way time-interleaved sar adc in 16-nm finfet," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 1101–1110, 2017.

[114] G. Balamurugan, J. Kennedy, G. Banerjee, J. E. Jaussi, M. Mansuri, F. O'Mahony, B. Casper, and R. Mooney, "A scalable 5–15 gbps, 14–75 mw low-power i/o transceiver in 65 nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 1010–1019, 2008.

[115] J. W. Poulton, W. J. Dally, X. Chen, J. G. Eyles, T. H. Greer, S. G. Tell, J. M. Wilson, and C. T. Gray, "A 0.54 pj/b 20 gb/s ground-referenced single-ended short-reach serial link in 28 nm cmos for advanced packaging applications," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 12, pp. 3206–3218, 2013.

[116] T. Shibasaki, T. Danjo, Y. Ogata, Y. Sakai, H. Miyaoka, F. Terasawa, M. Kudo, H. Kano, A. Matsuda, S. Kawai, *et al.*, "3.5 a 56gb/s nrz-electrical 247mw/lane serial-link transceiver in 28nm cmos," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 64–65, IEEE, 2016.

[117] F. O'Mahony, J. E. Jaussi, J. Kennedy, G. Balamurugan, M. Mansuri, C. Roberts, S. Shekhar, R. Mooney, and B. Casper, "A 47$times$10 gb/s 1.4 mw/gb/s parallel interface in 45 nm cmos," *IEEE journal of solid-state circuits*, vol. 45, no. 12, pp. 2828–2837, 2010.

[118] T. O. Dickson, Y. Liu, S. V. Rylov, A. Agrawal, S. Kim, P.-H. Hsieh, J. F. Bulzacchelli, M. Ferriss, H. A. Ainspan, A. Rylyakov, *et al.*, "A 1.4 pj/bit, power-scalable 16$\times$ 12 gb/s source-synchronous i/o with dfe receiver in 32 nm soi cmos technology," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 8, pp. 1917–1931, 2015.

[119] M. Mansuri, J. E. Jaussi, J. T. Kennedy, T.-C. Hsueh, S. Shekhar, G. Balamurugan, F. O'Mahony, C. Roberts, R. Mooney, and B. Casper, "A scalable 0.128–1 tb/s, 0.8–2.6 pj/bit, 64-lane parallel i/o in 32-nm cmos," *IEEE Journal of solid-state circuits*, vol. 48, no. 12, pp. 3229–3242, 2013.

[120] T. O. Dickson, Y. Liu, S. V. Rylov, B. Dang, C. K. Tsang, P. S. Andry, J. F. Bulzacchelli, H. A. Ainspan, X. Gu, L. Turlapati, *et al.*, "An 8x 10-gb/s source-synchronous i/o system based on high-density silicon carrier interconnects," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 4, pp. 884–896, 2012.

[121] T. O. Dickson, J. F. Bulzacchelli, and D. J. Friedman, "A 12-gb/s 11-mw half-rate sampled 5-tap decision feedback equalizer with current-integrating summers in 45-nm soi cmos technology," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1298–1305, 2009.

[122] A. Shokrollahi, D. Carnelli, J. Fox, K. Hofstra, B. Holden, A. Hormati, P. Hunt, M. Johnston, J. Keay, S. Pesenti, *et al.*, "10.1 a pin-efficient 20.83 gb/s/wire 0.94 pj/bit forwarded clock cnrz-5-coded serdes up to 12mm for mcm packages in 28nm cmos," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 182–183, IEEE, 2016.

[123] J. Song, S. Hwang, H.-W. Lee, and C. Kim, "A 1-v 10-gb/s/pin single-ended transceiver with controllable active-inductor-based driver and adaptively calibrated cascaded-equalizer for post-lpddr4 interfaces," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 1, pp. 331–342, 2017.

[124] H. Lee, K.-Y. K. Chang, J.-H. Chun, T. Wu, Y. Frans, B. Leibowitz, N. Nguyen, T. Chin, K. Kaviani, J. Shen, *et al.*, "A 16 gb/s/link, 64 gb/s bidirectional asymmetric memory interface," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1235–1247, 2009.

[125] K. Fukuda, H. Yamashita, G. Ono, R. Nemoto, E. Suzuki, N. Masuda, T. Takemoto, F. Yuki, and T. Saito, "A 12.3-mw 12.5-gb/s complete transceiver in 65-nm cmos process," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 12, pp. 2838–2849, 2010.

[126] J. Poulton, R. Palmer, A. M. Fuller, T. Greer, J. Eyles, W. J. Dally, and M. Horowitz, "A 14-mw 6.25-gb/s transceiver in 90-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 12, pp. 2745–2757, 2007.

[127] B. Casper, J. Jaussi, F. O'Mahony, M. Mansuri, K. Canagasaby, J. Kennedy, E. Yeung, and R. Mooney, "A 20gb/s forwarded clock transceiver in 90nm cmos b.," in *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pp. 263–272, IEEE, 2006.

[128] C. Thraskias, E. Lallas, N. Neumann, L. Schares, B. Offrein, R. Henker, D. Plettemeier, F. Ellinger, J. Leuthold, and I. Tomkos, "Survey of Photonic and Plasmonic Interconnect Technologies for Intra-Datacenter and High-Performance Computing Communications," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 4, pp. 2758–2783, 2018.

[129] S. Pasricha and M. Nikdast, "A survey of silicon photonics for energy-efficient manycore computing," *IEEE Design & Test*, vol. 37, no. 4, pp. 60–81, 2020.

[130] D.-W. Kim, K. Au, H. Y. L. X. Luo, Y. L. Ye, S. Bhattacharya, and G. Q. Lo, "2.5 d silicon optical interposer for 400 gbps electronic-photonic integrated circuit platform packaging," in *2017 IEEE 19th Electronics Packaging Technology Conference (EPTC)*, pp. 1–4, IEEE, 2017.

[131] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn, "Corona: System implications of emerging nanophotonic technology," *ACM SIGARCH Computer Architecture News*, vol. 36, no. 3, pp. 153–164, 2008.

[132] Q. J. Gu, "THz interconnect: The last centimeter communication," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 206–215, 2015.

[133] N. Van Thienen, Y. Zhang, and P. Reynaert, "Bidirectional communication circuits for a 120-ghz pmf data link in 40-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 7, pp. 2023–2031, 2018.

[134] D. Hou, Y.-z. Xiong, W. Hong, W. L. Goh, and J. Chen, "Silicon-based On-chip Antenna Design for Millimeter-wave / THz Applications," in *Proceedings of the EDAPS '11*, pp. 130–133, 2011.

[135] J.-D. Park, S. Kang, S. Thyagarajan, E. Alon, and A. Niknejad, "A 260 GHz fully integrated CMOS transceiver for wireless chip-to-chip communication," in *Proceedings of the VLSIC '12*, pp. 48–49, 2012.

[136] S. V. Thyagarajan, S. Kang, and A. M. Niknejad, "A 240 GHz Fully Integrated Wideband QPSK Receiver in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 10, pp. 2268–2280, 2015.

[137] C. W. Byeon, K. C. Eun, and C. S. Park, "A 2.65-pJ/Bit 12.5-Gb/s 60-GHz OOK CMOS Transmitter and Receiver for Proximity Communications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 68, no. 7, pp. 2902–2910, 2020.

[138] N. Weissman and E. Socher, "9mW 6Gbps Bi-directional 85-90GHz Transceiver in 65nm CMOS," in *Proceedings of the EuMIC '14*, pp. 25–28, 2014.

[139] N. Ono, M. Motoyoshi, K. Takano, K. Katayama, R. Fujimoto, and M. Fujishima, "135 GHz 98 mW 10 Gbps ASK transmitter and receiver chipset in 40 nm CMOS," in *Proceedings of the VLSIC '12*, pp. 50–51, 2012.

[140] E. Seok, D. Shim, C. Mao, R. Han, S. Sankaran, C. Cao, W. Knap, and K. K. O, "Progress and challenges towards terahertz CMOS integrated circuits," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 8, pp. 1554–1564, 2010.

[141] B. Khamaisi, S. Jameson, and E. Socher, "A 0.58-0.61THz single on-chip antenna transceiver based on active X30 LO chain on 65nm CMOS," in *Proceedings of the EuMIC '16*, pp. 97–100, 2016.

[142] R. Han and E. Afshari, "A High-Power Broadband Passive Terahertz Frequency Doubler in CMOS," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 3, pp. 1150–1160, 2013.

[143] S. Kang, S. V. Thyagarajan, and A. M. Niknejad, "A 240 GHz Fully Integrated Wideband QPSK Transmitter in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 10, pp. 2256–2267, 2015.

[144] X. Wu and K. Sengupta, "Dynamic Waveform Shaping With Picosecond Time Widths," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 2, pp. 389–405, 2017.

[145] S. Lee, R. Dong, T. Yoshida, S. Amakawa, S. Hara, A. Kasamatsu, J. Sato, and M. Fujishima, "An 80Gb/s 300GHz-Band Single-Chip CMOS Transceiver," in *Proceedings of the ISSCC '19*, pp. 170–172, IEEE, 2019.

[146] H. Aggrawal, P. Chen, and M. M. Assefzadeh, "Gone in a Picosecond: Techniques for the Generation and Detection of Picosecond Pulses and Their Applications," *IEEE Microwave Magazine*, vol. 17, no. 12, pp. 24–38, 2016.

[147] R. Han, Z. Hu, C. Wang, J. Holloway, X. Yi, M. Kim, and J. Mawdsley, "Filling the gap: Silicon terahertz integrated circuits offer our best bet," *IEEE Microwave Magazine*, vol. 20, no. 4, pp. 80–93, 2019.

[148] D. Correas-Serrano and J. S. Gomez-Diaz, "Graphene-based Antennas for Terahertz Systems: A Review," *FERMAT*, 2017.

[149] J. Blanckenstein, J. Klaue, and H. Karl, "A survey of low-power transceivers and their applications," *IEEE Circuits and Systems Magazine*, vol. 15, no. 3, pp. 6–17, 2015.