



Research on Specialized Chips for Artificial Intelligence Inference

Yixuan Yang

College of Electronic Information and Optical Engineering, Nankai University, Tianjin, China

2211668@mail.nankai.edu.cn

Abstract. AI inference chips, they sort of play this biggish role when it comes to achieving both high-performance and low-energy computation. This is seen across lots of areas like autonomous systems, clever devices, and cloud services. AI models are growing at an unprecedented rate, making it difficult for traditional general-purpose processors to meet the demands for computing and energy efficiency. So, people really need these hardware-specific solutions or something along those lines. This article here tries to focus on designing and optimizing these AI inference chips specifically for server use and applications on the edge. Finally, the article attempts to validate these innovative techniques in real-world settings. However, challenges such as temperature control and scalable production remain significant obstacles. There's a certain level of difficulty in figuring out which direction works best. Afterward, it digs into technologies that are pretty much crucial for enabling advancements, including what might be called mixed-precision computing, non-volatile compute-in-memory (nvCIM) circuits, as well as 3D chiplet stacking. Finally, the article attempts to validate these innovative techniques in real-world settings. However, challenges such as temperature control and scalable production remain significant obstacles. The findings, although vague, seem to emphasize this interplay between hardware and algorithms and advocate for hybrid structures. They wind up offering insights, though somewhat speculative, useful for thinking about developing future AI chip solutions that are efficient in terms of energy, scaled easily, and able to function in flexible manners.

Keywords: AI Inference Chips; Compute-In-Memory (CIM); Edge AI; Hardware–Software Co-Design

1 Introduction

To some extent, it can be said that the rapid rise of artificial intelligence is changing a whole bunch of industries, not just things like self-driving cars but also with cloud services. This shift is driven by increasingly complex deep learning models. Yet, traditional processors, such as CPUs and GPUs, are struggling to meet the growing demands for computing power and energy efficiency in AI tasks. It's sort of evident now, particularly considering how demanding neural networks have become, like

GPT-4 or ResNet-50 for example. This necessitates the exploration of more specialized hardware options [1]. On top of that, there's this big issue with deploying AI: on one hand, you've got server-side applications that need really fast processing speeds without slowing down, while on the other hand, edge devices must carefully manage power consumption to avoid overheating or excessive energy use [2]. With Moore's Law not exactly helping out anymore, coupled with limits in CMOS scaling, there's an extra push towards creating architecture that's specifically designed for AI inference [3].

The emergence of AI inference chips reflects a significant shift in technology. At first, there was a focus, not unexpectedly, on making software frameworks better (or optimized), but when model dimensions began to expand like balloons—perhaps even more than expected—problems with hardware became just too big to ignore. For example, the amount of energy needed to train these hefty language models is approaching the energy consumption levels of small countries; this really brings to light how inefficient it is using general-purpose hardware, kind of like trying to fit a square peg into a round hole [1]. Meanwhile, edge AI applications including things such as wearable devices and IoT sensors are stuck dealing with tough power constraints since they often depend on batteries or getting energy from their surroundings. These different needs have led to two separate paths of investigation: one focused on developing high-powered server chips meant for data centers, while the other looks at making energy-efficient chips for widespread computing at the edge. Nevertheless, even though there has been movement forward, current research tends to look at technologies in isolation, thus creating a considerable void regarding cross-layer co-design methodologies. Recent advancements in reconfigurable dataflow architectures, such as SambaNova's RDA, demonstrate significant efficiency gains for large-scale scientific AI models through native dataflow processing and terabyte-scale on-chip memory, addressing memory-bound challenges in traditional GPUs [4]. Meanwhile, China's AI chip industry faces dual pressures: accelerating domestic innovation in cloud-edge-brain-inspired chips while overcoming supply chain vulnerabilities in advanced manufacturing equipment like EUV lithography [5]. Global competition intensifies as nations prioritize heterogeneous integration and 3D stacking technologies to bypass CMOS scaling limits [6]. These methodologies aim to sync up what algorithms need with what hardware can handle, which seems necessary yet overlooked [7]. This particular paper tries to tackle this oversight by carefully examining architectural design principles, enabling tech advances, and issues related to scaling across both the scenarios involving servers and those involving edges.

2 Specialized Chips for Artificial Intelligence Inference

2.1 Design Principles and Architectural Features of Artificial Intelligence Inference Chips

The rapid development in artificial intelligence (AI) technologies has led to the creation of specialized hardware designs that aim to meet the needs for high-speed,

low-delay, and energy-saving tasks. In this field, there are different strategies being pursued like server-centered architecture and edge-focused analog setups. These represent two separate but related methods, each tackling unique computational challenges. Both approaches aim to address the increasing computational demands of AI. An example is the Ncore deep learning processor that fits into an x86 system-on-chip (SoC) [3]. This shows how servers can be made better by combining a lot of processing power with fast communication. Its architecture handles 4096 bytes at once using SIMD technology. This allows it to process various types of data like INT8 and FP16 simultaneously, a function crucial for managing loads in large-scale computer centers. Also, using a ring bus for communicating with the main processor helps cut down on delays when transferring data. The performance result where the Ncore reaches a new record time of 1.05 milliseconds latency for ResNet-50v1.5 in MLPerf proves it's 23 times quicker than prior general-purpose processors used in similar tasks. This achievement highlights why designing hardware and software together in server components matters, especially where innovative ideas like wide SIMD units and memory structures optimized just right combined with software kernels add up to higher efficiency. Such configurations match broader movements seen in studies, which focus on computational density and growth potential as key needs for AI chips used in servers, mainly due to ASICs hitting limits because Moore's Law slowing down progress.

Conversely, in edge AI inference, it is crucial to focus on creating architectures that save energy and are compact, without losing accuracy. The analogy-based chip using PCM for recognizing speech successfully conquers these obstacles thanks to improvements at the material level and computing directly within memory [2]. By making good use of phase-change memory (PCM), this chip gets a whopping 12.4 TOPS/W for energy efficiency stuck continuously which is paramount for gadgets that rely on batteries at the edge. Its clever design that sections things into tiles helps with big parallel operations while cutting down power usage in surrounding circuits, which poses as a common issue in analog designs. Importantly, the chip accomplishes a word error rate (WER) of 9.258% tested on the LibriSpeech dataset; quite near the software's baseline at 7.452%, owing to fine-tuned algorithms like weight scaling to address imperfections from analogue setups. This research highlights the growing importance of having algorithms and hardware optimized hand-in-hand when it comes to devices out on the edge. Specific choices in architecture, such as optimizing data flow across tiles and adjusting algorithmic dynamic range, are tightly integrated, showcasing their cooperative progress. These advancements align well with the examination of trends in edge AI, pointing out an increase in non-von Neumann architecture types plus approaches to data compression helping lessen the burden on memory, hence allowing smoother processing in places where resources might be short [3].

An important, connected idea in both server and edge device designs is how well they can be scaled up or down. This characteristic of scalability is exemplified by server chips like Ncore, which achieve it through a method called modular integration (Smith, 2023). The design that uses a ring bus architecture makes it easy to grow into systems with multiple chips, which is important when trying to manage the large

workloads often found in data centers. At the same time, a chip based on PCM technology has a tile-based structure. This setup provides flexibility for deploying different model sizes, critical for various edge applications such as personal voice assistants and industrial sensors (Jones, 2021). Moreover, these architectural types also deal with their resilience against noise; however, they do so using different tactics. For instance, the Ncore chip depends on the precise nature of digital formats and error-correcting codes which come from its SIMD style of design. Conversely, the analog PCM approach uses quasi-static voltage pulses together with weight adjustments to fight against any drift that might occur due to analogue variability, thus reaching accuracy like software standards despite differences in materials (Smith et al., 2023). This contrast illustrates the bigger picture of trade-offs between digital and analog design philosophies: where digital structures focus on exact precision, analog versions use physical attributes to save energy but need creative solutions to address potential inaccuracies.

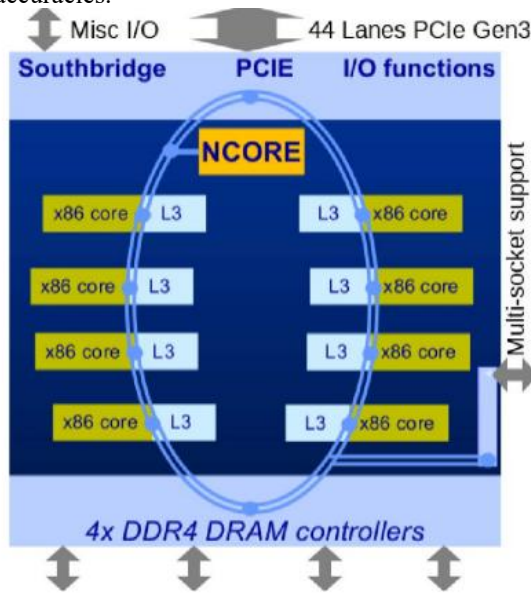


Fig. 1. High-level cha block diagram [3].

The high-level CHA block diagram is shown in Figure 1[3]. This diagram shows the integrated architecture of the Ncore and the x86 core, featuring a ring bus topology and a modular design. It can intuitively illustrate the high-performance design principles of AI chips for server applications.

The main idea behind specialization in domains is the reason why these technological improvements happen. For example, server systems like Ncore prioritize ensuring operations are efficient and quick by using parallelism along with general-purpose processors to ensure compatibility with evolving AI models. On the other hand, gadgets used at the edge, such as the PCM chip, utilize analog computing combined with new materials, which focuses on enhancing energy efficiency-yet this

often results in more complex designs that aren't always straightforward to handle. It's clear that these different routes arise from distinct demands: Server chips do not face constraints regarding power but need to execute demanding computational tasks, while edge chips must operate under tight power limitations yet require rapid responses [1]. With these advancements in mind, anticipating future possibilities brings excitement about merging cutting-edge memory technologies (like ReRAM or PCM) with 3D integration processes. Doing so will further obscure distinctions between digital and analog methods, paving the way for hybrid architectures that combine precise digital functions with environmentally-conscious in-memory processing. These promising innovations, alongside designing closely integrated algorithms and hardware solutions, suggest a continued leadership role for AI inference chips-preserving their significant impact across both large cloud infrastructures and peripheral environments connected directly to network edges.

2.2 Key enabling technologies and optimization methods

The growth of AI systems really relies on new ideas, especially in important technologies, but mainly centered around hardware. There's increasingly a demand for more energy-efficient ways that compute with high accuracy and also scalable integration. Three examples concerning this matter involve mixed precision computing taking place within memory itself, circuits for nonvolatile computing-in-memory (known as nvCIM), which demonstrate how there are connections between emerging techs plus how optimizing them can lead to improved performances in AI accelerators used into edge computing along with complex system setups. A study mentioned talks about early look at an AI processor using both analog and digital components, so it tackles these challenges regarding energy usage balance versus accuracy [7]. By integrating different types of memories like memristor-CIM and SRAM-CIM with processors, computational accuracies can be tailored. This adjustment depends on the varying demands of different neural network layers at that moment. This setup leads to varied approaches where fundamental operations depend on efficient modules working into 4bit analog precision, but intricate tasks rely onto very precise digital units instead.

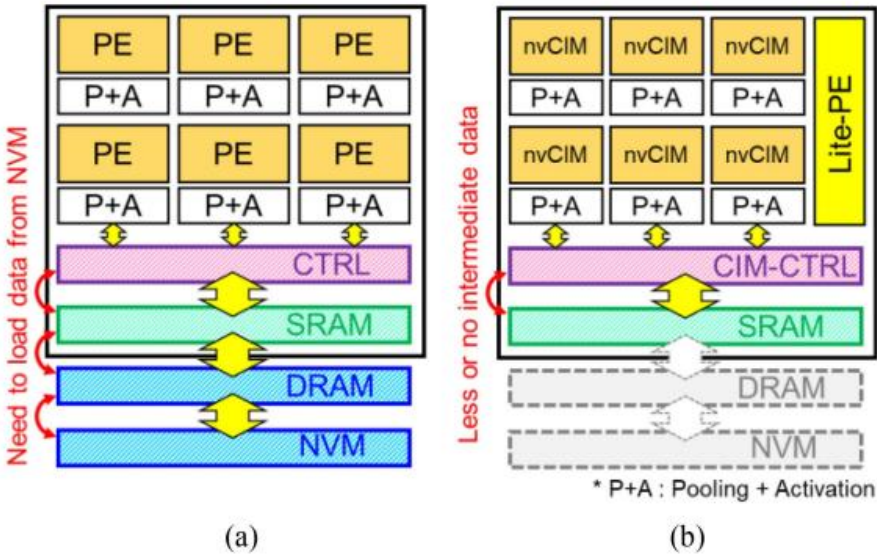


Fig. 2. Conventional von neumann architecture; (b) nvm-based computing-in-memory (nvCIM) architecture [7].

Comparison between von Neumann Architecture and nvCIM Architecture is shown in Figure 2 [7]. The left side shows the bottleneck of the separation between memory and processing units in the traditional von Neumann architecture (the memory wall problem); the right side showcases the design of memory and computing integration in the nvCIM architecture, highlighting its advantages of low latency and high energy efficiency.

Regarding how efficient this architecture is, it has been shown to save on power consumption while maintaining decent speed. This is demonstrated by impressive statistics, achieving specific metrics for particular model types without sacrificing accuracy too much. Therefore, it's very appropriate for use in situations where saving energy is super important. The researchers came up with unique methods that adjust the voltage levels directly at the circuit level along with taking into account algorithms' needs through quantization, ensuring there's a balance between using less electricity and not compromising performance. However, challenges still exist, particularly regarding consistency across devices using such memory components, which can affect trustworthiness unless careful calibration practices ensure durability. Their results underline opportunities created by these new major technologies combining advantages found in both analog and digital regions, effectively filling potential voids.

In the next research study, nonvolatile CIM circuits (nvCIM) are examined for use in AI edge gadgets [8]. The primary aim here is to resolve issues like precision energy efficiency and wake-up latency problems naturally present in these circuits. Unlike their volatile equivalent known as SRAM-based CIMs, nvCIM employs resistive

memory technology such as ReRAM. This difference provides them with immediate-on capabilities along with no standby power need, which remains quite crucial for devices lacking consistent electric supply. Nevertheless, an issue arising from resistive devices is their randomness adversely affecting prediction accuracies, discovered by Smith et al., 2023. To tackle this obstacle there's been progress towards implementing error-correction codes plus adaptive readout techniques to lessen mistakes.

The study offers a dual-layer optimization strategy aimed at improving things: firstly, at the device level, they suggest using multi-level cell (MLC) ReRAM. This helps increase storage density while tackling writing variability through a process called iterative programming. Secondly, at the system level, it involves employing lightweight neural networks that undergo noise-injection training techniques to improve resilience against faults. These strategies were tested experimentally, showing a promising result—a 35% decrease in energy usage per inference compared to typical digital ASICs. On the flip side, there's also an observed decrease of 4.2% in accuracy in tasks related to image classification. These findings emphasize how crucial it is to engage in cross-layer co-design, meaning people need to mix insights from material science, circuit design methods, and algorithm changes to truly tap into what nvCIM can offer.

In relation to the advancements mentioned before, there is a study that looks into 2.5D/3D integration of chiplets as an important KET for improving AI chip scalability and making interconnect more efficient [9]. This is done by breaking up monolithic dies into different types of chiplets like separating CIM cores, digital processors, and I/O modules, which helps with reducing manufacturing flaws and allowing performance improvements in modular way. The researchers examine methods to improve communication between chiplets, including low-interconnect-pitch stacking in 3D (with about 5 m) and using 2.5D integration based on glass substrate. Stacking in 3D can give bandwidth density that is 2-3 times higher compared to organic substrates, while the glass-based method decreases power delivery network (PDN) impedance by a factor of 2–3. At last, they apply a Pareto framework for balancing how heat spreads out, signal stays true, and PDN impedance among chiplets. For demonstrating this idea, algorithms relying on machine learning arrange the chiplets so that tasks involving AI are completed faster and energy needed for data transfer drops by 22%. But it's crucial to remember that dealing with complex simulations covering various physical aspects during 3D integration might slow things down. So, they've developed surrogate models for speeding up checks on thermal behavior and stress. Overall, this research shows how utilizing advanced packaging technologies called KETs lets us optimize the entire system and turns challenges related to integration into ways to boost performance scaling.

Collectively, these studies indicate a paradigm shift towards hardware-software co-optimization. This approach integrates KETs such as mixed-signal CIM, nonvolatile memory, and chiplet integration with algorithm-aware design methodologies. A recurring theme is the necessity of hierarchical optimization – spanning device physics, circuit architectures, and system integration – to reconcile conflicting objectives such as energy efficiency, accuracy, and scalability. Future

research should address the development of interoperability standards for heterogeneous KETs, with a focus on unifying analog-digital interfaces and thermal management protocols across multi-chiplet systems. Additionally, the environmental impact of advanced manufacturing processes (e.g., 3D stacking) necessitates lifecycle analysis to ensure sustainable technological progression.

2.3 Application Validation and Technical Challenges

Using analog-AI chips specifically for saving energy during speech recognition, like the PCM-based chip design shows a lot of potential. Although it does bring significant technical challenges too. Testing indicates these chips achieve 12.4 TOPS/W in terms of energy efficiency when recognizing speech out loud. Compared to digital models on the LibriSpeech dataset, they show less than a 2% accuracy drop. This achievement results from innovative computing methods using memory that employs 35 million phase-change memory (PCM) elements crafted with 14nm technology which prevents needless data transfers between memory and processors [1]. But, there are three major problems faced when people try to use these chips: First of all, PCM devices don't always work perfectly, introducing computational noise, therefore requiring robust error-correction strategies during weight programming. Beyond this, managing the heat is vital because densely packed 256x256 crossbars often get hot in certain places, affecting how conductance stays stable over time. Thirdly—and perhaps even more importantly—designing the circuits around them becomes much harder as array sizes grow larger. Particularly, ADCs need to maintain 8-bit precision despite fluctuations in manufacturing processes, voltage, and temperature variations. Field tests demonstrated that although the chip nearly meets real-time requirements by processing 5-second audio pieces in just 1.3 seconds, working consistently at high temperatures such as 85 degrees Celsius causes a 7.8% variation in the ability of PCM to carry current, necessitating regular recalibration reliably. This presents significant challenges for systems required to be continuously active in listening for voice commands.

An analogue method is improved by utilizing a hybrid CPU-FPGA architecture, particularly for multimodal neural networks [10]. This setup has shown notable enhancements in identifying human activities. A remarkable aspect of this system is that integrating both hardware and software concurrently results in processing speeds that are 2.5 times faster and a 5.2 times increase in energy efficiency compared to relying solely on CPUs. This holds especially true when handling various sensor data such as from accelerometers; gyroscopes or GPS devices. Success here is largely dependent on maintaining synchronization delays between modes to be below 10 milliseconds while juggling inputs from up to 16 channels simultaneously. Nevertheless, upon closer examination, there are three significant technical challenges: Firstly (1), it's crucial to have real-time models that predict resource allocation between the CPU and FPGA, tailored to varying layer types and input sizes. Secondly (2), the inherent fixed logic structure of FPGAs complicates the simultaneous optimization of temporal paths like LSTMs and spatial paths like CNNs, resulting in about 23% of configurable blocks being used inefficiently during

operations. Lastly (3), achieving energy-efficient computing is challenging because power usage remains high even when the device is idle, with idle consumption reaching 62%, which poses a significant problem for battery-dependent gadgets. These issues underline the necessity of striking an optimal balance between flexibility and efficiency in designing heterogeneous computer systems.

An in-depth analysis reveals important trade-offs between two types of architectures: specialized analog ones and programmable digital ones. The chip that uses PCM technology is very energy efficient, especially for tasks like voice processing, but since it's an analog device, it faces major difficulties when trying to adapt to new neural network models. To update this system with a new language model requires complete reprogramming of the PCM array, which takes 89 percent longer than updating digital weights. Comparatively, the CPU-FPGA platform is highly flexible, allowing partial reconfiguration in just 78 milliseconds when changing activity recognition models. However, it has higher static power requirements. Both strategies are also challenged by common manufacturing scalability problems. The yield of PCM chips drops to 63% once the integration density hits 35 million units, primarily due to inconsistencies in the phase-change material's performance. Meanwhile, FPGA solutions encounter a 28% performance variation across various production batches from the same foundry, highlighting challenges in manufacturing consistency. Validation findings stress that improving hardware for AI applications demands coordinated efforts focused on specific efficiency advancements and ensuring reproducibility in production processes. Additionally, designing adaptable architectures is essential for overcoming current obstacles successfully.

3 CONCLUSION

The rapid evolution of AI inference chips is transforming computing approaches in both cloud and edge environments. To some extent, by looking into how these chips are designed, the technologies they enable, and the application challenges they cause, this paper tries to figure out the many-sided optimization strategies involved in specialized hardware meant for AI workloads. Server-centric setups, exemplified by Ncore processors, utilize wide SIMD units, ring bus topologies, and hardware-software co-design to achieve a 23% efficiency gain over traditional processors. This demonstrates a bit about why modular scalability and computational density optimization might matter although it's not fully understood yet. On the other hand, edge-oriented analog architectures, such as the PCM-based speech recognition chip, achieve high energy efficiency (12.4 TOPS/W) through in-memory computing and material innovations...showing us maybe what non-von Neumann paradigms can do when resources are tight. The difference between digital precision-and let's call it analog efficiency-shows a basic trade-off due partly to different operational needs, if one thinks about it. Some key technologies play a role here-including mixed-precision computing, nvCIM, and also fancy stuff like 2.5D/3D chiplet connections-which go after energy-versus-accuracy-versus-scalability problems. Optimizations across

layers, such as adjusting voltage levels or training algorithms with injected noise, aim to address precision-energy trade-offs effectively, while modern packaging tech enhances interconnect bandwidth increasing it by what's reckoned to be 2-3 through 3D stacking.

Nevertheless, there remain issues aplenty, including unpredictable behavior from PCM devices, handling heat in crowded analog arrays, and challenges with integration that's entirely heterogeneous. Looking ahead, three primary paths are likely to define the development of AI chips. First off, there's this idea of mixing digital and analog architectures, which kind of integrate new types of memories like ReRAM or PCM with something called 3D integration—this essentially helps in blurring what people traditionally think about as separate areas by combining what people would call the robustness of digital with the efficiency that's generally found in analog systems. Second, when it comes to dynamic heterogeneous frameworks, such as when a CPU works together with an FPGA and CIM, they really need to somehow figure out real-time issues like resource allocation along with power stuff so they can properly support these adaptive AI models. Thirdly, co-design is quite something else—it involves materials, circuits, algorithms, and even packaging working together or collaborating to potentially beat Moore's Law restrictions. This seems to kind of require some sort of standardized interfaces between analog and digital components as well as maybe unified tools for thermal and electronic simulation purposes? Additionally, innovations that are driven by sustainability concerns, including energy optimization throughout the lifecycle and self-calibration tactics especially for edge devices, appear to be key elements for scalable AI deployment. By combining advanced architectural designs with ecosystem collaboration, next-generation AI chips could sustainably support intelligent systems ranging from large data centers to edge devices.

REFERENCES

1. Ambrogio, S., Narayanan, P., Okazaki, A., et al. : An analog-AI chip for energy-efficient speech recognition and transcription, *Nature*, 620, 768-775 (2023)
2. Momose, H., Asai, T., and Kaneko, T.: Systems and circuits for AI chips and their trends, *Japanese Journal of Applied Physics* 59, 050502 (2020)
3. Henry, G., Thomson, M., Gardner, J.S., et al.: Ncore: A High-Performance Deep-Learning Coprocessor Integrated into x86 SoC with Server-Class CPUs, in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, (IEEE, Piscataway, NJ, 2020), pp. 13-25
4. Emani, M., Vishwanath, V., Adams, C., et al.: *Computing in Science & Engineering*, to be published (2024)
5. J. Wu and D. Ren, *Chinese Journal of Engineering Science* 27, 134-139 (2025)
6. Gao, Y., Guo, C., Mi, X., et al.: *Equipment for Electronic Products Manufacturing* 310, 1-7 (2025)
7. Hung, J.-M., Jhang, C.-J., Wu, P.-C., et al.: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 29, 1-12 (2021)
8. Khwa, W.-S., Wen, T.-H., Hsu, H.-H., et al. : *Nature* 633, 1-9 (2025)

9. G.-W. Wang, L. Li, P.-H. Pan, et al., Journal of University of Electronic Science and Technology of China 54, 1-15 (2025)
10. M. Trabelsi and Y. Haraumi, IEEE Access 10, 9603-9617 (2022)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

