MediaTek Advanced Research Center

Call for Research

(MARC-CFR)

Research Needs

April. 2025

MARC, MediaTek Advanced Research Center

Research Needs

1. 6G Communication Systems

Key technology explorations for 6G Communication System Design

■ Research Needs Label: [6GSys]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Abbreviations

3DGS	3D Gaussian Splatting
3GPP	Third Generation Partnership Project
ACLR	Adjacent Channel Leakage Ratio
AI/ML	Artificial Intelligence and Machine Learning
AIXP	AI-Exchange Protocol
API	Application Programming Interface
BS	Base Station
CA	Carrier Aggregation
CSI	Channel State Information
ICC	Integrated Communication and Computing
IP	Intellectual Property
MAC	Medium Access Control
MCP	Model Context Protocol
MIMO	Multiple Input Multiple Output
PHY	PHYsical layer
QoE	Quality of Experience
QoS	Quality of Service
RAT	Radio Access Technology
RLM	Radio Link Monitoring
RRM	Radio Resource Management
RX	Receive/Reception
SBFD	Sub-Band Full Duplex
SCell	Secondary Cell
SLA	Service Level Agreement
TN/NTN	Terrestrial/Non-Terrestrial Network
TX	Transmit/Transmission

UE	User Equipment
UL	Up Link
UP	User Plane
XR	Extended Reality

Motivation

The advent of 6G wireless communication systems is poised to bring significant advancements over previous generations, particularly in areas such as spectral efficiency, data throughput, end-to-end latency, power consumption, security and privacy preserving. These improvements will pave the way for groundbreaking applications, including cloud gaming, XR, sensing, and high-precision positioning.

6G will also require a network architecture and cross-layer optimized protocols that are more secure and energy-efficient. Additionally, the development of UE with cost-effective, low power consumption, and evolution-friendly modem architecture will be critical to the success of 6G deployment [1].

Moreover, 6G will venture into new mid-band carrier frequencies and non-terrestrial deployments, where the propagation characteristics will differ dramatically from current systems. The integration of the latest AI/ML technologies will be essential in addressing these challenges and achieving our goals [2].

We invite proposals that address these key areas, with a particular emphasis on the application of AI/ML techniques where appropriate. Proposals that include proof of concept prototyping, when applicable, will be given additional consideration. We are interested in projects spanning short-term and long-term horizons:

- Short-term projects: These correspond to 3GPP Releases 20 and 21. Release 20 marks the study phase of 6G, set to commence in the summer of 2025. A comprehensive overview of the scope for Release 20 can be found in the technical document contributions of 3GPP 6G workshop [3] [4]. The findings from Release 20 will be the foundation of Release 21, during which formal normative work for 6G will be undertaken, starting at the beginning of 2027.
- Long-term projects: Pertaining to Release 22 and beyond, these may be referred to as 6G Advanced, with a focus on future advancements and innovations.

In light of the above, we propose the following list of research areas, though it is by no means exhaustive. We encourage submissions that explore these and other relevant topics:

■ Specific Areas of Interest

Air Interface [5]

1) Modulation, Waveform and Channel Coding
   I. Modulation, waveform and multiple access schemes for coverage and spectral efficiency enhancement in terrestrial and non-terrestrial channels
   II. High throughput channel coding
2) Multi-Antenna Techniques
   I. CSI aging mitigation
   II. CSI acquisition and compression, for both downlink and uplink, using classical and/or AI/ML approaches, based on channel reciprocity and/or measurement of forward link reference signal
   III. Metasurface fabrication, operational control, system performance assessment, channel modeling and prototyping
   IV. Massive and Massively Distributed MIMO system design including, but not limited to, transmission schemes, CSI acquisition, beam management, reference signal overhead reduction, energy-efficiency and network nodes synchronization etc., considering the surge in the number of elements in upper-mid band (7 to 15 GHz) antenna arrays
   V. Joint design of transmit scheme and receiver algorithm for Low-Complexity, High-Rank MIMO connections
3) UE-SBFD
   I. UE-SBFD Demo/Research: This involves researching and demonstrating the feasibility and performance of UE-SBFD. The scope could cover advanced self-interference suppression mechanisms at the UE, such as low-power and low-complexity mechanisms for higher TX-RX antenna isolation, TX-blocker suppression, TX-leakage reduction, and digital self-interference cancellation.
4) Joint RF/base-band design for UL enhancements
   I. UL Throughput and coverage enhancement via fast TX reconfiguration/switching cross carriers/bands. Explore the possible designs for TX switching and simulate of the performance gains with TX switching.
   II. Develop PHY mechanisms for utilizing the TX switching to enhance the UL performance.
   III. ACLR framework to enable maximum TX power transmission and reduce

guard band in 6G carriers.

IV. Extended coverage design aligning TN and NTN.

5) Communication for AI
   I. Traffic analysis and modeling for native AI radio networks
   II. Corresponding air interface optimization supporting the traffic pattern of native AI radio networks
6) Application of AI/ML techniques (including semantic communication) for any of the above

New Applications

1) Integrated Sensing and Communication [6] [7]
   I. Identification of novel use cases, particularly those involving a UE, and the design of their enabling technologies and overall system architectures
   II. Sensing assisted communication, including estimation of positions of obstacles and scatterers in the channel by the UE and/or the base stations, to assist beam management, cell reselection/handover and CSI acquisition, among other functional elements in communication
   III. Sensing by the fusion of data from multiple devices such as mobile phones, watches and glasses, to enhance user experience and/or assist communication
   IV. Data processing for environment perception, including target identification (automobile or drone) and gesture recognition, using AI/ML techniques
   V. Low-overhead, multi-dimensional reference signal design for both mono-static and bi-static sensing
   VI. RF fingerprinting based on multi-dimensional MIMO channel for high precision positioning, CSI compression and other novel applications
   VII. Proof of concept prototyping for any of the above
2) Cross-layer optimizations for future applications
   I. 3DGS Avatar Platform: Develop AI algorithms to implement 3DGS avatars on XR devices, enabling photorealistic virtual avatars that enhance social interaction in the Metaverse
   II. Context-Aware Edge-Cloud AI Assistant: Design an edge-cloud AI-powered XR assistant (e.g. AI agent) capable of preserving context and enabling intelligence collaboration across multiple devices and platforms (e.g. multi-AI agents)
   III. Remote XR Application Platform: Develop a proof-of-concept remote XR application to evaluate 6G system requirements, and user experience

improvements achieved through cross-layer optimization

3) AI Agent Communication
   I. Comprehensive study of AI Agent Communication Protocols, including MCP, Responses API, AIXP, and other competing solutions
   II. Literature review of legacy and new solutions with detailed examination of the protocols, including their architecture, use cases, and implementation specifics such as protocol encoding format, mandatory and optional Information Elements, etc.
   III. Analysis and comparison of different solutions, considering factors such as scalability, security, ease of integration, supporting companies, potential challenges and benefits for various industries, and performance in terms of computing complexity and memory and storage requirements
   IV. Collection of empirical data through experiments, simulations, or case studies
   V. Final report summarizing findings, insights, logs (e.g., wireshark), and recommendations

Higher Layer Protocol, RRM and Security

1) User plane protocol design
   I. Streamlining data transfer through the user plane to reduce latency bottlenecks in the stack, to address 6G use cases (ICC, XR, etc.) while keeping complexity low
   II. Proof of concept prototyping new UP design to determine its impact on end-to-end applications in real-time
2) Measurements framework enhancements
   I. Mobility and measurements framework enhancements:
      - Explore new mechanism (including AI solutions) to minimize the number of measurements occasions needed by the UE without impacting UE operations, such as mobility.
      - Develop a design that requires less interruption during measurements via exploring possible scenario of embedding the interruption time within the scheduled symbols.
      - Develop a faster and accurate handover design between cells (including AI solutions).
      - Enable AI based handover without measurements allowing the UE to handover to another cell based on AI module that could be developed according to the UE movements and position.

II. CA and Dual RAT operation enhancement
- Develop faster techniques for SCells measurements ensuring seamless operation of carrier aggregation including SCell activation/deactivation and addition/release.
- Develop solutions to address the synchronization issue of concurrent RAT operation.

III. Radio link monitoring
- Explore new mechanisms to improve power saving.
- Explore new mechanisms to reduce false-alarm radio link failure.

IV. Multi-device RRM grouping:
- Explore, develop and design how a group of devices, of a single user or multi-users, can utilize and share RRM operations, such as RLM/measurements/handover.

3) Next generation security, trustworthy and privacy preserving systems
   I. Novel approaches to overcome privacy concerns in distributed Machine Learning wireless systems.
   II. Novel algorithmic frameworks for communication-efficient and differentially private federated learning wireless systems with applications to real-world use cases.
   III. Holistic 6G network security architecture planning.

Network and System Architecture [8]

1) NTN
   I. Spectrum sharing mechanism between TN and NTN for improving overall spectral efficiency while managing the interference
   II. Non-classical techniques for the mitigation of challenging NTN propagation characteristics such as long delay, fast-changing and high interference level [9]

2) New network architecture and protocols for enabling computing, sensing and intelligence services
   I. Explore architecture and protocols to enable computing service through integrating cloud platform, e.g., Kubernetes, with 3GPP service architecture, including performance management of compute services to meet SLA, e.g., combined QoE control mechanism for communication and computing
   II. Explore architecture and protocols to enable data services, including collection, storage, distribution, analysis, life cycle management, automation, etc., to facilitate network AI, network automation,

positioning/sensing service, or any other service requiring continuous data

III. Explore architecture and protocols to enable intelligence services, especially with GenAI capability, like network agent, multi-agent collaboration, etc. Service requirements for human or AI agent as customer

IV. Proof of concept to demonstrate new use cases and enabling technology, especially ones impacting consumer and device capability, e.g., UE triggered computing offload

3) Proximity Network

I. Short-range communication (e.g. for joint AI inference/training across devices)

II. Device collaboration for collaborative MIMO, positioning or sensing

4) AI Framework for 6G

I. Design and evaluate solutions for AI-enhanced mechanisms in protocols to achieve system optimization, e.g., further enhance mobility performance with advanced AI technologies, including but not restricted to Reinforcement Learning, on-line training.

II. Study new AI system architecture or framework integrated with 3GPP network for distributed intelligence using, for example, federated learning and intelligence plane for an AI application, to protect user's consent and privacy

III. Novel approaches to overcome the constraints of mobile devices and enable AI generated content services in future mobile networks, including cooperation among multiple mobile devices and network clouds (e.g., edge clouds).

IV. Novel approaches for dynamic creation of native AI and computing networks (e.g., creating networks on the fly based on required services/applications/compute power).

Environmental Sustainability

1) Energy Efficiency

I. Ultra-Low-Power Radio System

- Achieve ≤1 μW (≤100 μW) power consumption with ≥−80 dBm (−90 dBm) sensitivity
- Enable tunable frequencies and robust co-channel interference mitigation
- Support multi-user operations and withstand time/frequency variations (potentially with reference signal design and assistance)

II. Distributive Energy-Efficiency Optimization
- Optimize transmission and processing power consumption across base stations and UEs for holistic energy efficiency
- Minimize computation/communication overhead by avoiding fully centralized strategies
- Adapt to 6G networks with cooperative UEs and intermediate nodes, ensuring broad applicability

III. AI-Assisted System-Wide Energy Efficiency Optimization
- Employ AI for cross-domain (time/frequency/spatial/power domains), cross-layer (PHY/MAC/RRC) and BS-to-UE energy efficiency optimization
- Leverage distributed learning or shared intelligence to minimize the energy footprint
- Balance AI-driven enhancements with stringent power constraints

2) Carbon-Aware System Operation

I. Develop carbon related performance metrics and end-to-end evaluation methodology for measuring carbon emissions of mobile communication systems, considering different spatial and temporal scales.

II. Explore the integration of power grids with future mobile communication systems to achieve carbon reduction, by examining how these two systems work together to effectively lower carbon emissions.

III. Investigate a carbon and QoS-driven service architecture aimed at providing green and user-centric services that prioritize both environmental sustainability and user satisfaction in service delivery.

IV. Develop mechanisms for monitoring energy consumption, energy supply mix, and carbon intensity in mobile communication systems, considering spatial and temporal granularities.

V. Design carbon-aware resource management strategies for next-generation communication systems, incorporating computing and sensing aspects among others, focusing on minimizing carbon emissions through energy-related criteria.

UE Architecture

1) 6G modem system architecture

I. Architecture exploration of 6G modem IP with the key directions identified, including processors, platform, and hardware/firmware/software partitioning, pursuing leading position in performance, low-power, and cost

effectiveness as a product.

II. Methodology and tools for supporting evaluation, simulation and profiling of the IP architecture design and implementation.

2) Neuromorphic processing [10] [11]

I. Explore application of neuromorphic processing architectures to enable wireless AI applications in a power efficient manner.

II. Quantify potential power consumption gains of such architectures over existing ones

■ Reference for 6G Communication Systems: (please see page 50)

2. Radio System Solutions

■ Research Needs Label: [RSS]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

2.1 Wireless RF

■ Motivation

New communication standards such as 5G beyond and WiFi8 increases throughput, reduces latency, while for commercialization, transceivers need to consume low power and have smaller form factors. To fulfill these demands, advancement of the following technologies is required.

First is the power amplifier which is usually the most power consuming circuitry in a transceiver. Both 5G and WiFi8 adopts OFDM whose signal peak-to-average-power-reduction ratio (PAPR) is typically more than 6dB. Therefore, maintaining high efficiency at both peak output power and >=6dB power backoff is desirable. Also, for 5G and beyond, new mmWave frequency bands are being opened up. Multi-band mmWave power amplifiers are needed to reduce phased-array module and system sizes.

Second is receiver architecture and components for multi-mode operations. Compared to the conventional architecture, a direct sampling RF receiver offers greater flexibility, easier for integration and occupies smaller area in advanced process nodes. By removing mixers and using a wide-bandwidth ADC to digitize RF waveforms directly, signals can be processed in the digital domain. ADC with wide bandwidth and high sampling rate is the essential component for such a receiver architecture. If signals of interested RF bands can be sampled and digitized in ADC's first Nyquist zone, complicated filtering, signal processing and frequency planning can be greatly simplified.

Finally, because a higher-order modulation is required, for example, from 1024QAM to 4096QAM for WiFi6 and WiFi7/8, respectively, multi-mode, wide tuning range, high resolution, and high-quality signal sources, such as crystal oscillator and voltage-controlled-oscillator (VCO) are necessary.

Low power consumption, wide bandwidth, high performance, and small form factor are generally required for all circuits and systems.

■ Specific areas of interest

High efficiency sub-6GHz power amplifier simultaneously achieving the following targets:

1) Switch-cap PA with >15% 3dB fractional bandwidth; center frequency between 2-6GHz using 1.8V supply.
2) QFDM-64QAM average power > 20dBm, average PAE > 25%, while passing FCC emission requirement. If necessary, develop pre-distortion tailored for this specific PA and apply.

Wide-bandwidth ADC:

1) Class 1: Sampling rate >3GS/s, over-sampling ratio between 2-8, dynamic range >57dB, interleaved paths <=4.
2) Class 2: Signal bandwidth >6GHz (preferably >13GHz), SNR/SFDR 55-60dB, Nyquist sampling preferred but the second Nyquist zone is possible. Emphasis on power efficiency.
3) Clocking, and driving and reference buffers for the ADC need to be included.
4) Specific interest in architectures that employ digital calibration/compensation e.g. AI/machine learning to improve performance in advanced process technologies and overcome bottlenecks in traditional architectures

Direct IF bandpass receiver:

1) Sampling rate >3GS/s; IF signal bandwidth > 400MHz; dynamic range > 57dB.
2) The receiver needs to deal with anti-aliasing without using bandpass filter at its input, at least up to 5th harmonics of the sampling clock.

VCOs (5-80GHz), DCOs (5-80GHz) and Crystal oscillators (<150MHz) exploring the following:

1) Wide tuning range (continuous or banded operation), high-performance and low-power
2) Phase noise suppression techniques for 10kHz~10MHz frequency offset away from the carrier frequency
3) Switched-cap array with >20000ppm tuning range, <0.05ppm resolution, and DNL < 0.5LSB with sufficient quality factor compared to those of other tank elements
4) Low power techniques to trade power with performance while satisfying key communication system requirements at respective operating modes of interest

Frequency synthesis

1) Power efficient frequency synthesizers with <40 fs integrated jitter
2) Focus on mmW frequency generation 28-150GHz and/or at 5-15GHz

Novel building blocks, subsystems or architecture for Sub-THz applications: VCO, LNA, PA, low resolution ADC / DAC, phased array, integrated antenna and circuitry.

Process technology for sub-THZ or high output power applications: e.g., GaN.

AI-Assisted transceiver design:

Example coverage of topics of interest include but not limited to the following:

1) Creation of nonconventional transceiver component, circuit, subblock, or subsystem using AI: passive synthesis, exploration of circuit topology, alternative subsystem achieving the same function, and so on.
2) Design flow refinement through AI: using AI to reduce design time / license requirement during transceiver design, integration of AI to existing transceiver design flow to enhance productivity.

[Reference] ISSCC 2025 Paper 25.3, "AI enabled Design Space Discovery and End to end Synthesis for RFICs with Reinforcement Learning and Inverse Methods Demonstrating mmWave /sub THz PAs between 30 120 GHz".

■ Special information:

If specific process is required to achieve required circuit performance, the research proposal needs to explicitly request and provide sufficient justifications. Access to such process can be discussed with corresponding MediaTek owners.

2.2 6G FR3 Antennas

■ Motivation

The relentless evolution of wireless communication systems is driving the need for more advanced and efficient antenna technologies. As we transition from 5G to 6G, the demand for higher data rates, lower latency, and more reliable connections continues to grow. The introduction of new frequency ranges, such as Frequency Range 3 (FR3), which encompasses higher frequency bands, presents unique opportunities and challenges for User Equipment (UE) antenna design, particularly in the context of Multiple Input Multiple Output (MIMO) systems.

Higher Frequency Bands and Bandwidth: FR3 operates at higher frequency bands, which offer wider bandwidths and the potential for faster data transmission rates. However, these higher frequencies also experience greater propagation loss and are more susceptible to blockage and absorption by obstacles. New MIMO antenna designs for UE must be optimized to operate efficiently within these bands, ensuring robust signal reception and transmission.

Enhanced Spatial Multiplexing: MIMO technology leverages multiple antennas at both the transmitter and receiver to increase the capacity of a radio link through spatial multiplexing. With the advent of 6G, the need for advanced MIMO techniques becomes even more critical to meet the expected exponential growth in data traffic. New UE MIMO antennas must support enhanced spatial multiplexing capabilities to deliver the multi-gigabit per second data rates envisioned for 6G.

Digital beamforming utilizing antenna arrays is especially crucial in FR3, where the coherent combination of signals can mitigate the challenges associated with increased path loss at higher frequencies. It is imperative that new UE antennas integrate a specific number of arrays to enhance both link reliability and spectral efficiency.

Device Size and Integration: As UE devices continue to shrink in size, integrating multiple antennas without compromising performance becomes increasingly challenging. The design of new 6G FR3 UE MIMO antennas must consider form factor constraints, ensuring that antennas are not only compact but also capable of coexisting with other device components without causing interference.

User Experience and Coverage: The ultimate goal of 6G is to enhance the user experience by providing ubiquitous coverage and seamless connectivity. New MIMO antenna designs must ensure consistent performance across diverse environments, from dense urban areas to rural locations, enabling a seamless user experience regardless of location.

In summary, the motivation for developing new 6G FR3 UE MIMO antennas lies in addressing the unique challenges posed by higher frequency bands while capitalizing on their potential to deliver unprecedented data rates and connectivity. The design of these

antennas will play a pivotal role in realizing the ambitious goals of 6G and shaping the future of wireless communication.

■ Specific areas of interest

Antenna topology study covering the following FR3 frequency ranges:

1) 5.9 to 8.4GHz
2) 12.7 to 13.25GHz
3) Dual band antenna covering both 5.9-8.4GHz and 12.7 to 13.25GHz
4) Dual band antenna covering S-band and C-band

The study of antenna miniaturization and strategic placement

1) This is crucial across different product platforms, including smartphones, tablets, and notebooks.
2) Modern smartphones, for instance, already incorporate over ten antennas within their compact frames. Consequently, it is essential to ensure that the FR3 antenna is sufficiently miniaturized to integrate seamlessly, particularly within the constrained space of a smartphone.

A high-performance antenna equipped with the following features to improve MIMO T-put performance

1) Antenna isolation: > 20dB
2) Antenna mismatch: < -15dB
3) ECC over FOV: < -15dB
4) Other features could also be studied and proposed from this research

Omi-directional antenna

1) 25% gain CDF and 75% gain CDF delta: < 2dB

Compact modular antennas with 4x or 8x antenna ports that could fit along the edge side of the phone

1) modular antenna with 4x antenna ports
2) modular antenna with 8x antenna ports
3) Preliminary size constraints for the 2x ports modular antenna: 3.8 x 20 x 2.5 mm^3
4) Preliminary size constraints for the 4x ports modular antenna: 3.8 x 40 x 2.5 mm^3

5) Frequency range of interest: 12.7 to 13.25GHz

A study and verification of FR antenna MIMO T-put performance

1) Development of a MIMO T-put simulation platform
2) Development of a MIMO T-put measurement and verification platform including in-house testing lab set up
3) Investigation of MIMO T-put capabilities within the constraints of a smartphone enclosure

3. Analog Circuits

■ Research Needs Label: [Analog]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

- High performance, high bandwidth, power efficient analog circuit continue to play important role for wireless, wireline communications, automotive, smart home and AIoT applications. The key areas include power management, data converters and high speed Serdes. The focus includes innovations in architectures, circuits, and systems.

- Power management: Explore integrated circuits and/or application circuits that could improve power conversion efficiency for application processors, RF power amplifiers, mobile devices, IoT and wearable applications

- Data converter: Analog-to-digital converters and Digital-to-analog converters are fundamental and enabling building blocks for a wide range of applications from meter, audio to communications and beyond. The techniques to improve resolution, dynamic range, sampling rate, and energy efficiency (FoM) are highly demanded.

- High speed interface (e.g., serdes): Techniques to support high data rate, power efficient data links and high density I/O system over advanced 2.5D/3D package are of interest.

■ Specific areas of interest

1) High speed interface Serdes: Power and area efficient circuits including but not limited to AGC, equalizers, high-bandwidth amplifiers, analog and ADC/DAC-based front ends, TX drivers and low jitter clocking, clock recovery, etc. with state-of-the-art performance (upon normalization over process technology if needed). Optical communication circuits and systems are also of interest.
2) 2.5/3D (INFO/CoWoS) interconnect with data rate 32+ Gb/s/wire. Power efficiency <=0.3 pJ/bit @ N4 process and could have normalization over e.g., process technology if needed.

3) Chip-to-chip single-ended communications on substrate, with data rate 16+ Gb/s/wire. Innovative architecture to achieve best power efficiency is highly interested.
4) High dynamic range, low power data converters and analog front end for audio and sensor applications. (>120dB, preferably >140dB)
5) High sampling rate, power efficient data converters for wireless applications (>=10bits, >2GSPS/channel) [1] and wireline applications (7-8 bits, 2GSPS/channel and power <=2mW).
6) Time-interleaved analog-to-digital converter calibration techniques for sampling rate >20Gs/s. (>=10bits)
7) IVR (Integrated Voltage Regulator) for SoC
   I. May include hybrid SC, LDO, etc. Vin=1.2V, Vout=0.3V ~ 1V, Iout > 2A [2]
8) XPU Power Delivery
   I. Multi-phase fast transient (>2A/0.1us) area efficient Buck converter with >90% peak efficiency @4-to-0.8V, Iout_max > 10A
   II. Multi-phase fast transient (>2A/0.1us) area efficient Buck converter with per-phase >90% efficiency @1.8-to-0.8V and Iout_max/2, Iout_max > 2.5A, phase number >= 8 and inductor <20nH
9) RF PA Power Delivery / Modulator
   I. >100MHz (200MHz is preferred) ETM with efficiency >90% and low noise (e.g. spur noise -49dBm/MHz) [3]
10) Ultra-low voltage, low power analog circuits for bandgap, temperature sensor, oscillators and clocking with high stability, etc. (<=0.5V, nW)
11) Circuits and systems for analog AI, CIM, etc. that support AI computing acceleration and non-conventional computing.
12) Reliable and functional safety circuit design for automotive applications.
13) AI-powered design methodology for analog design productivity and performance boost.

■ Reference for Analog Circuit Research Needs: (please see page 52)

4. Edge AI System

■ Research Needs Label: [Edge AI]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation: Shaping the Future of Edge AI – Empowering Frontier Innovation

AI and computing power are transforming every aspect of our lives - from learning and working to driving, communication, and entertainment. Breakthroughs in large language models, generative AI, and intelligent agents - enhanced by multimodality and advanced reasoning capabilities - are reshaping the AI landscape. The MediaTek Edge AI platform is at the forefront of this revolution, integrating sophisticated AI into a diverse range of edge devices, including mobile phones, wearables, televisions, smart speakers, tablets, computers, automobiles, and IoT devices.

Advancements in emerging foundation models are rapidly widening the gap between AI's escalating computation demands and the constraints of the semiconductor scaling. As edge AI applications expand, the required computing power far exceeds what current semiconductor process scaling can provide. To close this gap, we are advancing hardware architectures and streamlining software and algorithm complexity through continuous improvements, paving the way for next-generation AI applications. Despite challenges such as increasing SoC complexity, limited memory bandwidth, and thermal constraints, our commitment to innovation remains unwavering. We invite academic experts to join our collaborative research programs to overcome these challenges, cultivate talent, and shape the future of AI and computing technology.

We cordially invite research proposals that drive the development of next-generation algorithms and architectures across diverse application areas. Our focus is on advancing applied algorithms, optimizing systems, and exploring compute architectures for emerging AI models. We welcome innovative research proposals in data collection, synthesis, and benchmark methodology, as well as algorithm-hardware co-design for edge devices. In addition, we encourage proposals that integrate scalable machine learning cores with energy-efficient hardware architectures, extending to advanced NPU system designs integrated with CPUs, GPUs, MCUs, memories, and more. We also encourage interdisciplinary collaborations that bring high innovation value and originality,

with the potential to create robust, efficient, and scalable AI platform solutions.

Join us as we pioneer new frontiers in foundational research and system design to overcome the inherent challenges of advanced AI systems and hardware.

■ Areas of Interest (including but not limited to)

Application

1) Generative AI on Edge devices (video and 3D, multimodality, long context, reasoning, agent, LLM OS)
2) Embedded vision and computational photography
3) Autonomous driving, end-to-end ADAS models, smart cockpit
4) Edge-Cloud collaboration
5) Multi-user multi-device architecture/algorithm co-design for Edge inference server
6) Embodied AI, VLA (vision-language-action) foundation models

Machine-Learning Core

1) Low-bit representation for edge intelligence, efficient representation for GenAI on edge devices (Floating, Integer, or special representation)
2) New foundation model/architecture (e.g. Mamba)
3) AI compiler optimization (TVM/MLIR) (e.g. Vulkan ML)
4) Multi-core computing technology (e.g. runtime optimization)
5) Compiler scale-up technology

AI Hardware

1) NPU multi-core scalability
2) On-the-fly activation compression/decompression for edge devices
3) Weight compression for edge devices
4) Ultra low power AI: e.g. CIM (Compute In Memory), PIM (Processing In Memory)

■ Reference for Edge AI System Research Needs:

MEDIATEK

Connections

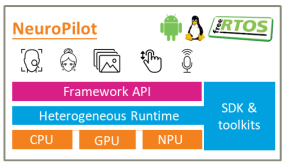
Professors
- NTU / NTHU / NCTU
- EU – ETHZ
- MUS-Perdue, Duke, CMU, MIT
- MSL-NTU

Sponsor & Committees
- MobileAI Workshop @CVPR2025
- NTIRE Workshop @CVPR2025
- PAIR challenge @ICME2025
- ICPP-EMS 2025
- TVM Conference 2025
- Key AI/ML & CV & Graphics workshop/conferences in TW - CVGIP, CGW, TAAI, CTHPC

Talents & Summer Intern

Applications

- Generative AI on Edge devices (video and 3D, multimodality, reasoning, agent, LLM OS)
- Embedded vision and computational photography
- Autonomous driving, end-to-end ADAS models, smart cockpit
- Edge-Cloud collaboration
- Multi-user multi-device architecture/algorithm co-design for Edge inference server
- Embodied AI, VLA (vision-language-action) foundation models



ML Core

- Low-bit for edge intelligence, efficient representation for GenAI on edge devices (Floating, Integer, or special representation)
- New foundation model/architecture (e.g. Mamba)
- AI compiler optimization (TVM/MLIR) (e.g. Vulkan ML)
- Multi-core computing technology (e.g. runtime optimization)
- Compiler scale-up technology

Hardware

- NPU multi-core scalability
- On-the-fly activation compression/decompression for edge devices
- Weight compression for edge devices
- Compute In Memory (CIM), Processing In Memory (PIM)

5. Multimedia

Research Needs Label: [MM]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

Motivation

The fields of Image Processing and Computer Vision (CV) play a pivotal role in enhancing the convenience of daily life through a myriad of applications, ranging from consumer electronics and surveillance cameras to advanced driver-assistance systems (ADAS). With the rising popularity of edge devices, such as mobile phones, TVs, and tablets, there is an increasing demand for the integration of these applications into these platforms. The advent of Artificial Intelligence (AI) has marked a new era of progress, offering advancements that surpass traditional methods. Despite these achievements, current AI methodologies face significant challenges. High computational and memory requirements often hinder the practical deployment of AI solutions, rendering them less feasible for real-world edge applications. Additionally, the data-driven nature of AI necessitates extensive datasets for training, which poses substantial hurdles in data collection and annotation. It is, therefore, imperative to develop efficient AI strategies that not only address these limitations but also maintain a balance between performance and efficiency, ultimately facilitating their application in product development.

In light of the aforementioned challenges, we are inviting research proposals that aim to devise practical AI solutions capable of enriching our lives through diverse applications, including but not limited to smartphone cameras, ADAS, surveillance systems, and edge devices. Proposals may focus on various aspects such as application development, algorithmic innovation, methodological advancements, or domain specific HW accelerator design. We are particularly interested in research that ventures into untapped areas, promising high levels of innovation and potential impact. Below are some key areas of interest, although proposals are not limited to these topics alone. We encourage the submission of research that explores novel territories, striving for groundbreaking advancements in the field.

Specific areas of interest

Real-world AI image/video restoration and enhancement, with complexity and power consumption considerations

1) Efficient AI image/video restoration (denoising, super-resolution, …), video stabilization, video frame interpolation, … etc.
2) Real-world video streams from TV, streaming or social media
3) Real-world RAW images from cameras sensors
4) Perceptual image/video quality assessment
5) Hardware-optimized AI/GenAI accelerators/functions/techniques design

Efficient network for vision applications and scene/intention analysis

1) Joint training of visual perception systems (depth, detection, segmentation, …), with temporal stability
2) Scene/intention analysis for surveillance and ADAS system
3) Domain adaptation approaches (unsupervised or semi-supervised domain adaptation is preferred)
4) A simulator or a real platform for validating the proposed ADAS approach

Visual attention and transformers for low level image processing and visual recognition

1) Practical vision applications with visual attention or transformers
2) Feasible complexity for edge devices
3) Domain adaptation consideration
4) Self-/semi-supervised learning is encouraged

AI video compression

1) AI loop filtering [1][2][3][4]
2) AI intra prediction [5][6][7]
3) AI super resolution [8][9][10]
4) Other AI video coding tool(s) [11][12]
5) End-to-end AI video coding [13][14][15]

Extended Reality (XR)

1) Simultaneous localization and mapping (SLAM)
2) Object/scene 3D reconstruction
3) Natural user interface

No-reference video quality assessment

1) Evaluate texture details over time: measure the smoothness and naturalness perceived by the human eye

2) Identify and locate camera ISP issues observed in the video, including their type and position

■ Reference for Multimedia Research Needs: (please see page 53)

6. Heterogeneous Integration for 2.5D/3D Packaging

Novel Material, Architecture, Interconnection, Co-Packaged Optics, Silicon Photonics, Reliability, and Thermal Management for 3D-IC & Chiplet Package Applications

■ Research Needs Label: [HI for PKG]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

As the size of chips continues to shrink, modern integrated circuit (IC) design is facing many challenges. One of the challenges is how to effectively integrate multiple chips in terms of power, performance, area, cost, and reliability (PPACR). 3D-IC and 2.5D chiplet technologies are two promising solutions, but there are also some unique challenges, especially in package.

3D IC and 2.5D IC technologies are advanced techniques used to integrate multiple chips into a single package, enhancing performance and efficiency. Each chip can contain different functions, such as processors, memory, and sensors. Here are some challenges that may be encountered in 3D IC and 2.5D IC packaging:

- Heat dissipation issues: When multiple chips are stacked together in 3D ICs or placed closely in 2.5D ICs, the heat they generate will accumulate. This may cause excessive heat buildup, resulting in system crashes or performance degradation. To address this issue, more efficient heat dissipation solutions (such as cooling strategies) and novel thermal interface materials (TIM) need to be developed.
  Power supply issues: When multiple chips are integrated, they require higher power supply. This may result in unstable power supply, leading to system performance degradation. To address this issue, more efficient power management technology and power supply solutions need to be developed.

- Signal interference issues: When multiple chips are close to each other, signal interference issues may arise. This may cause signal distortion or system crashes. To address this issue, more effective signal paths and signal shielding technology need to be developed.

- Mechanical (Warpage) issue: Mechanical warpage in 3D ICs poses a significant challenge to the semiconductor industry, affecting the structural and SIPI integrity and functionality of multi-layered devices. As the demand for more compact and powerful electronic devices grows, the need to address the warpage issue becomes increasingly critical.

- High Bandwidth Memory (HBM) integration: Integrating HBM into 3D ICs and 2.5D ICs can significantly enhance memory bandwidth and performance. However, this integration also introduces challenges such as thermal management, power delivery, and signal integrity. To address these issues, innovative solutions in thermal dissipation, power management, and high-speed interconnects are required to ensure the efficient and reliable operation of HBM within these advanced packaging technologies.

- Large package size issue: The semiconductor industry is on the cusp of a transformative shift towards larger and more complex package designs, driven by the escalating requirements of acceleration chips in AI servers and the demand for High Bandwidth Memory (HBM). There is an urgent call for innovation in large package technology, particularly for reticle sizes expanding to larger than 6.0x with the integration of more than 12 ~16 HBM.

- Heterogeneous integration issues: Chiplets may be produced by different manufacturers using different technologies and materials. This may lead to heterogeneous integration issues, such as thermal expansion mismatch and different mechanical properties. To address this issue, more effective bonding and interconnect technologies need to be developed.

- Interconnect density issues: Since chiplets are smaller in size than traditional chips, they may require higher interconnect density. This may lead to interconnect density issues, such as signal crosstalk and power supply noise. To address this issue, more efficient interconnect design and signal shielding technology need to be developed.

- Test and debug issues: Since chiplets are produced separately and then combined, testing and debugging may be more challenging. To address this issue, more effective test and debug technologies need to be developed to ensure the reliability and quality of the final product.

- Co-packaged optics (CPO): The integration of co-packaged optics (CPO) with semiconductor devices represents a pivotal advancement in data communication technology. As data center bandwidth requirements continue to escalate, the need for efficient, high-speed optical interconnects within close proximity to electronic chips has become critical. To address this request, we are seeking innovative solutions that combine co-packaged optics with advanced packaging technologies to resolve the challenges of next-generation data transfer and processing.

■ Specific areas of interest

1) Innovative package architecture/technique integrated with HBM for large package design (>6.0x reticle size)
2) Effective thermal management, innovative cooling strategy, optimal thermal design
3) Novel anisotropic thermal interface material (TIM)
4) High thermal conductivity molding compound
5) Backside power via for PDN layout application
6) Hybrid OX bonding scheme development for bonding interface strength and thermal performance optimization.
7) The thermal-mechanical stress evaluation of 3D-IC stacking chip/monolithic SoC in advancing packaging
8) Die-to-Die interconnect design
9) Innovative decoupling capacitor solutions in Packaging to meet ultra-high di/dt request
10) Cutting-edge co-packaged optics solutions that can be seamlessly integrated with advanced packaging techniques.
11) Novel EIC and PIC integration and fiber attach technology
12) Innovative approaches to integrate optical components such as lasers, photodetectors, and waveguides with IC packages.
13) Thermal management solutions to address the heat dissipation challenges of CPO.
14) Signal integrity analysis and optimization for high-speed optical data transmission.
15) High die stacking development (>20die stacking) for HBM 3D package
16) Thin die solution and methodology (with contactless die pick up, thin die strength enhancement (plasma grinding/ dicing) for HBM applicaiton
17) HBM cube themal/ stress/ warpage solution with high die stacking
18) Advanced and novel Cu-Cu Hybrid bonding technology.
19) Contact-less testing methodology for advanced packaging

7. Software Engineering

■ Research Needs Label: [SE]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

7.1 AI-Driven Attack Paths Prediction and Defense

■ Motivation

In the contemporary landscape of software development, security has emerged as a critical concern. As software systems grow in scale and complexity, the number and variety of vulnerabilities also increase, providing malicious attackers with ample opportunities to exploit these weaknesses. Traditional security measures often involve reactive approaches, addressing vulnerabilities only after they have been exploited. However, the integration of AI-driven technologies offers a proactive solution by predicting potential attack paths and implementing defensive strategies in advance.

The inclusion of SDL (Security Development Lifecycle) in this topic underscores the importance of embedding security considerations throughout the software development lifecycle. From requirement analysis to design, implementation, testing, deployment, and maintenance, each phase presents unique opportunities to integrate AI-driven security measures. By leveraging AI to predict attack paths, developers can identify and mitigate potential vulnerabilities during the early stages of development, thereby enhancing the overall security posture of the system.

AI-driven attack path prediction involves utilizing machine learning models to analyze historical attack data and predict future attack vectors. This predictive capability enables organizations to adopt a proactive defense strategy, focusing resources on the most critical areas of vulnerability. The application of AI in this context not only enhances the efficiency of security measures but also reduces the resource-intensive nature of traditional security testing methods.

Moreover, the integration of AI into SDL ensures that security is no longer an afterthought but an integral part of the design process, such as threat modeling. By automating the identification of potential attack paths, AI tools can assist developers in making informed decisions, thereby fostering a culture of proactive security management.

In conclusion, the motivation for this topic lies in addressing the challenges of modern software development by harnessing the power of AI to predict and defend against potential attacks. By embedding AI-driven security measures within the SDL, this topic aims to enhance the resilience of software systems, safeguarding them against evolving threats and ensuring the development of more secure and reliable software solutions.

■ Potential areas of interest (but not limited to)

Leverage AI for threat modeling, tasks may include:

1) Identify and analyze attack paths by reviewing system artifacts like diagrams, flowcharts, or descriptions. A key challenge is enabling AI to comprehend these diagrams and perform attack path analysis.
2) Provide mitigation strategies for identified threats.
3) Evaluate the updated risk landscape post-mitigation using models like DREAD or other risk assessment frameworks.

Utilize AI for codebase assessment, tasks may include:

1) Use AI to analyze the codebase, predict potential attack paths, and filter out high-risk modules.
2) Detect vulnerabilities in high-risk modules using AI, focusing on issues like those in the CWE TOP 10.
3) Assess the severity of vulnerabilities using the Common Vulnerability Scoring System (CVSS).
4) Generate potential remediation solutions for these vulnerabilities with AI.

■ Reference:

[1] Franco Terranova et al., "Leveraging Deep Reinforcement Learning for Cyber-Attack Paths Prediction: Formulation, Generalization, and Evaluation," 2024, https://dl.acm.org/doi/abs/10.1145/3678890.3678902

I. Evan Crothers et al., "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," 2023, https://arxiv.org/abs/2210.07321

II. Lara Mauri et al., "Modeling Threats to AI-ML Systems Using STRIDE," 2022, https://doi.org/10.3390/s22176662.

7.2 AI FUZZ

■ Motivation

With the rapid development of artificial intelligence (AI) technology, enhancing the efficiency and coverage of fuzz testing has become a significant issue. By introducing AI technology to explore the construction of a more efficient fuzz testing system, AI can automatically optimize and adjust fuzz testing rules, thereby improving system efficiency and coverage. Moreover, it can identify as many security vulnerabilities as possible within a limited time and provide remediation suggestions, ultimately enhancing software security.

■ Potential areas of interest (but not limited to)

Enhancing fuzzing efficiency and coverage using AI.

1) Combine Coverage-guided fuzzing techniques and AI technology to build a more efficient fuzz testing system. For example, using large language models (LLMs) to generate fuzz test seed inputs and other methods to increase efficiency and fuzz coverage.
    I. AI-driven patching.

Use multiple interactive LLM agents to generate patches.

■ Reference:

[1] Google oss-fuzz-gen, https://github.com/google/oss-fuzz-gen
    I. Trail of Bits' Buttercup heads to DARPA's AIxCC, https://blog.trailofbits.com/2024/08/09/trail-of-bits-buttercup-heads-to-darpas-aixcc/.

7.3 Developing the Standard for Embedded System Development Benchmark for Large Language Models

■ Motivation

The rapid advancement of Large Language Models (LLMs) has opened new frontiers in various domains, including natural language processing, code generation, and automated reasoning. However, the evaluation of these models, particularly in the context of embedded systems, remains a significant challenge. Current benchmarks, such as HumanEval and SWE-Lancer, primarily focus on general-purpose programming tasks

and do not adequately address the unique requirements and constraints of embedded systems.

Constraints of Embedded Systems:
Unlike general-purpose computing environments, embedded systems face strict limitations on memory, computational speed, and power consumption. Existing benchmarks like HumanEval do not systematically account for these constraints, limiting their relevance to real-world embedded applications.
Need for Standardized Evaluation:
Without a cohesive benchmark standard, it becomes challenging to compare different LLM architectures or fine-tuning strategies specific to embedded contexts. Inconsistent or incomplete evaluation methods lead to confusion about which models are truly best-suited for real-time, resource-limited, and safety-critical operations.

Ensuring Safety and Reliability:
Embedded systems often power mission-critical devices (e.g., automotive, aerospace, medical). A robust benchmark must not only measure a model's code generation accuracy but also consider safety guarantees and reliability in edge cases.

Bridge Academia and Industry:
Developing a standardized evaluation suite helps unify academic research and industrial deployment strategies, accelerating innovation in embedded AI applications. Universities and industry labs can use a common platform for measuring progress, thus improving reproducibility and fostering collaborative advancements.

■ Potential areas of interest (but not limited to)

Benchmark Design and Development:

1) Define the criteria and metrics for evaluating LLMs in embedded systems.
2) Develop a diverse set of tasks and challenges that reflect real-world embedded system scenarios.
3) Ensure the benchmark covers various aspects, including code generation, optimization, verification, and debugging.

Resource Constraints and Optimization:

1) Investigate how LLMs can generate code that meets the resource constraints of embedded systems (e.g., memory, processing power, energy consumption).

2) Explore techniques for optimizing code generated by LLMs to enhance performance and efficiency.

Reliability and Real-Time Performance:

1) Assess the reliability of code generated by LLMs in embedded systems, focusing on error rates, fault tolerance, and robustness.
2) Evaluate the real-time performance of LLM-generated code, ensuring it meets the timing requirements of embedded applications.

Security and Safety:

1) Investigate the security implications of using LLM-generated code in embedded systems.
2) Develop benchmarks to assess the safety and security of LLM-generated code, particularly in critical applications.

Human-AI Collaboration:

1) Study the interaction between human developers and LLMs in the context of embedded systems.
2) Develop tools and interfaces to facilitate effective collaboration and code review processes.

■ Reference:

[1] https://arxiv.org/abs/2107.03374

[2] https://github.com/openai/human-eval

[3] https://openai.com/index/swe-lancer/

7.4 Enhance LLM Comprehension for Wireless Debug Logs

■ Motivation

- LLMs face challenges in interpreting non-linguistic Wireless logs.
- Wireless log structure differs from typical NLP inputs, creating a gap.
- AP network logs or Wireshark logs can offer initial insights.
- Academic collaboration may boost LLM interpretation precision.

■ Potential areas of interest (but not limited to)

Data Annotation & Processing:

1) Automatic labeling/annotation for non-linguistic Wireless logs

Domain-Specific Model Training:

1) Fine-tune LLMs on annotated Wireless log datasets.
2) Develop approaches to map technical log syntax to natural language.

Hybrid Analytical Approaches:

1) Combine rule-based parsing with data-driven LLM methods.
2) Evaluate models on real-world trace and module log challenges.

7.5 LLM Management & Application Framework for Domain-Specific Knowledge

■ Motivation

- LLMs require layered management due to varied data trust and sensitivity during pre-training/fine-tuning.
- Current practices drive up deployment costs, demanding efficient management strategies.
- Industry-Academia Collaboration can innovate new data stratification and LLM governance methods.

■ Potential areas of interest (but not limited to)

Data Trust Assessment: Evaluate source reliability and sensitivity for improved pre-training.

Layered Model Governance: Develop frameworks to manage diverse data layers in LLM deployment.

Efficiency Optimization: Propose techniques for cost reduction and scalable LLM frameworks.

Security and Compliance: Examine regulations and risks to protect sensitive information.

7.6 Latest 3GPP Spec Evaluation with reference software with LLM

■ Motivation

- Rapid 3GPP changes demand agile evaluation
- Quick performance verification is essential.
- Collaboration can reduce evaluation impact risks

■ Potential areas of interest (but not limited to)

Code Adaptation Analysis:

1) Examine 3GPP modifications on OpenAirInterface C/C++ code.
2) Use LLM to understand a big software base and apply the corresponding code modifications and verification

Automated Performance Testing:

1) Develop rapid, automated test suites.
2) Simulate standard change scenarios for validation.

7.7 Automating Document Consistency and Traceability in SDLC Using Large Language Models (LLMs)

■ Motivation

The software development lifecycle (SDLC) is essential for guiding teams in creating high-quality software efficiently. Managing SDLC is strenuous due to the extensive documentation and numerous work products that require constant updates. Changes in code, driven by client feature requests or other reasons, necessitate updates across multiple documents to maintain consistency and ensure traceability. This process is labor-intensive and prone to errors, highlighting a critical need for automation (Omer & Sahraoui, 2021).

Recent advancements in large language models (LLMs) have shown promise in automating tasks such as document generation and updates. Leveraging LLMs for this purpose within chip design processes can reduce the manual labor required and enhance the accuracy of documentation updates. By utilizing accurate and relevant data from document databases and code repositories, LLMs can provide the necessary context in prompts to facilitate automated traceability and document consistency (Goknil et al., 2014; Cleland-Huang et al., 2015). This research aims to investigate the application of LLMs in automating the documentation process in a fabless chip design company, focusing on achieving improved consistency and traceability.

■ Potential areas of interest (but not limited to)

LLMs for Automated Document Updates

1) Examination of existing LLMs and their capabilities in document generation and update.
2) Methods to retrieve and provide the most relevant data to LLMs from document databases and code repositories to ensure context accuracy.
3) Case studies showcasing successful implementation of LLMs in updating product documentation according to industry standards (e.g., ASPICE).

Ensuring Document Consistency and Traceability

1) Techniques to integrate LLMs into the SDLC process to maintain consistency across multiple documents when code changes occur.
2) Development of algorithms that enable LLMs to identify and update upstream and downstream tasks affected by code changes.
3) Evaluation of the impact of LLMs on reducing manual labor and increasing the accuracy of documentation updates.

■ Reference:

[1] Goknil, A., Kurtev, I., Van Den Berg, K., & Veldhuis, J. (2014). Traceability and consistency in model-driven engineering: A survey of techniques, tools, and trends. Journal of Systems and Software, 99, 258-280. https://doi.org/10.1016/j.jss.2014.08.030

[2] Cleland-Huang, J., Gotel, O., Zisman, A. (2015). Requirements Engineering: Foundation for Software Quality. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-16101-3_1

[3] Omer, M. & Sahraoui, H. (2021). Requirements and Design Consistency: A Traceability and NLP Assisted Approach. IEEE Transactions on Software Engineering. https://doi.org/10.1109/TSE.2021.3059883

[4] Yang, Y., Huang, L., Yuan, Z., & Song, M. (2021). A survey of deep learning techniques in software documentation. ACM Computing Surveys, 54(3), 1-36. https://doi.org/10.1145/3448400

7.8 Enhancing Change Request Classification and Analysis with Large Language Models (LLMs) for Improved Efficiency and Accuracy

■ Motivation

The motivation for adopting large language models (LLMs) in the classification and

analysis of change requests (CRs) is derived from improving efficiency, accuracy, and scalability in the software development and maintenance processes. Traditional approaches of handling CRs are often labor-intensive and susceptible to human errors, resulting in delays and inconsistencies that adversely affect project schedules and quality. By leveraging the advanced natural language processing capabilities of LLMs, we can automatically classify and analyze CRs, which greatly reduces manual efforts and speeds up SDLC cycle time.

Moreover, LLMs have demonstrated remarkable abilities in understanding and interpreting complex language patterns, which can help lead to precise categorization and insightful analysis of CRs. This not only improves the reliability of the overall issue triage process, but also identifies potential trends and problems that human may ignore. Furthermore, the scalability of LLMs allows them to handle large volumes of CRs efficiently, making them an ideal solution for organizations dealing with extensive and continuously growing datasets. In summary, the adoption of LLMs for CRs classification and analysis promises to transform the way software development teams manage change requests, driving improvements in productivity, accuracy, and overall project outcomes.

■ Potential areas of interest (but not limited to)

Automated Classification Systems

1) Explore how LLMs can be integrated into the existing SDLC workflows to automate the CRs classification (e.g., issue type)
2) Discuss the potential of reducing manual errors and boost issue triage cycle time.

Semantic & Sentiment Analysis

1) Investigate the application of LLM in conducting NLP analysis on CRs to gauge the urgency or importance of requests.
2) Analyze how sentiment analysis can help prioritize CRs and improve processing time.

Trend Identification

1) Examine how LLMs can identify trends and issue patterns in CRs over time, providing insights into common issues or recurring problems (e.g., duplicated).
2) Discuss the potential for proactive problem-solving and preventive measures based on the trend analysis.

Reference:

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

[2] Zhang, Y., & Yang, Q. (2021). A Survey on Multi-Task Learning. IEEE Transactions on Knowledge and Data Engineering, 34(12), 5586-5609.

[3] Zhang, Y., & Yang, Q. (2022). Root Cause Analysis in Software Systems Using Large Language Models. Proceedings of the 2022 IEEE/ACM International Conference on Automated Software Engineering (ASE), 123-134.

[4] Liu, X., Li, Y., & Wang, J. (2023). Leveraging Large Language Models for Automated Root Cause Analysis in IT Operations. Journal of Systems and Software, 192, 110-123.

8. EDA

■ Research Needs Label: [EDA]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

- People continue to discover how to apply AI and ML to IC design. This leads to higher productivity and higher product quality. Certain areas of IC design, of high interest to MediaTek, are under-served by commercial tool vendors. Physical design plays a crucial role in various aspects of integrated circuit (IC) design. However, current approaches still heavily rely on manual tuning, and institutions and companies are investing more resources in solutions and academic articles to address this challenge. Nevertheless, there are still some missing pieces that need to be included in the reality IC design flow, such as design rule handling and data transmission timing minimization. Therefore, there is a need for efficient continuous and combinatorial optimization methodologies to handle the increasingly extreme design complexity and design rules.

- Moreover, the continuous scaling of semiconductor technology nodes presents
- significant challenges in achieving optimal Power, Performance, and Area (PPA). Design Technology Co-Optimization (DTCO) and System Technology Co-Optimization (STCO) are critical methodologies that address these challenges by integrating system, design, and technology considerations early during implementation. However, DTCO and STCO involve many design stages, are highly complex, and require significant time and manual intervention. Recent advancements in AI and ML for IC design have demonstrated promising benefits for productivity and quality. Therefore, we aim to leverage AI and advanced algorithms to develop in-house capabilities for DTCO and STCO tasks at MediaTek.

■ Potential areas of interest but not limited to

1) Sign-off corner reduction

2) APR runtime reduction w/ GPU acceleration (for >4M design)

3) RTL Verilog coding Spec2C, Spec2Test

4) DVFS Governor to optimize DVFS policy and reduce Automotive latency

5) Multi-Chip floorplanning in 2.5D/3D IC

6) Applying defect-based fault models to enhance defect coverage and diagnosis capability.

7) EDA for Design Technology Co-Optimization (DTCO) and System Technology Co-Optimization (STCO)
   I. AI and ML Algorithms for DTCO and STCO for PPA optimization
   II. Transistor-based custom and standard cell design flows and optimization including transistor synthesis, transistor P&R and modelling
   III. Timing, IR and circuit analysis and optimization for pre-silicon and pre-silicon database
   IV. 2.5D/3DIC Power Integrity Optimization, PDN Resource Optimization, and PDN & Timing Co-Optimization. (Note that the PDN resources include hybrid/micro-bump, C4 bump, TSV/TIV, and all metal connections within chiplets, interposer, organic/glass-core substrates, and the PCB)

8) Multi-objective constrained optimization with low optimization budget

9) Multi-physics (Front-End Variation, Layout Dependent Effect, IR, Thermal, Aging, Stress) aware timing prediction

10) GPU-based design methodology and approach

11) Early SI/PI/Congestion Analysis and Prevention

12) Early routing layer/NDR budgeting for high-speed design (for EMI/...)

9. Special Topic

9.1 GPU

■ Research Needs Label: [GPU]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

- **Intense Competition in GPU Design:** The development of GPU design for mobile graphics has reached a highly competitive stage. To maintain competitiveness, it is necessary to simultaneously output realistic lighting and shadow rendering, high geometric detail, and support intelligent resolution and frame rate.

- **Research on Realistic Algorithms for Ray Tracing:** To improve realism, it is essential to enhance various realistic algorithms for ray tracing, including direct lighting and shadows, multiple reflections and refractions, atmospheric diffraction, and camera-like handling of motion blur and focus.

- **Large-Scale Scenes:** The trend towards seamless, freely explorable open worlds presents time challenges for preparing and updating the memory data structures required for graphics. More comprehensive IP functionality and specification designs are needed to meet usage requirements, whether for ray tracing intersection calculations, initial ray emission position preparation, intersection material lookup, or reconstruction of these structures.

- **Detailed Character Animations:** The precision of character animations is also gaining attention. Lifelike skin and skeletal animations require the use of neural network algorithms and changes to the GPU's information flow to reduce DRAM exchanges, thereby lowering BUS and DRAM power consumption and focusing power on critical computations. These are new challenges for edge computing in the new era.

- **Achieving Ultra-High-Quality Graphics**: After obtaining ultra-high-quality graphics, it is important to address noise reduction, super-resolution, anti-aliasing, and generating new frames from historical frame content. These are crucial GPU functionalities that also require the integration of neural network algorithm research and processing capabilities.

Therefore, we need various research projects to explore the development of these important directions and nurture talents who can lead MTK GPU to the forefront of competition

■ Specific areas of interest

Modem GPU Architecture for Neural Graphics

Interest

1) Neural Super Resolution
2) Neural Frame Rate Generation
3) Neural Super Sampling & Anti-Aliasing
4) Neural clothing/mesh animation
5) Neural materials/Neural texture compression/neural displacement
6) Neural lighting/Neural ray denoising
7) Neural post-processing/Neural AS LOD
8) Neural Shading

Some Opportunities

1) Modern GPU architecture to accelerate neural graphics applications.
2) Algorithm/network model trade-off between performance/power/area and picture quality

Raytracing on Mobile

Interest

1) The raytraced-based importance sampling / guiding method for faster lighting converge or denoising friendly.
2) Data compression of raytracing data at different level, including AS layout, AS depth reduction by geometry representation change.
3) Bandwidth reduction by smart caching policy or design.
4) Evolution of ray traversal and acceleration structure to handle large seamless game scene, skin animation, and far objects.
5) Compact & condense primitive data representation & intersection test algorithm, such as opacities and displacements to plot a triangle to a rich detailed geometry.

Some Opportunities

1) Algorithms with lesser and lesser sampling
    I. The developer can reduce the ray jobs by the importance sampling for the advance effects, such as indirect lighting and color bleeding etc…., this may also reduce the cost of denoiser (such as the size of filters and number of

filtering).

II. The better initialization of ray jobs with a cache or guiding algorithms, such as **ReSTIR**, path guiding, **radiance caching** and so on, those methods may keep coherence between frames and increase the quality of 1st iteration.

2) Data compaction and compression of raytracing geometry
   I. Raytracing algorithms are reported as the bandwidth bound problem. How to reduce the bandwidth for different data:
   II. **Data size** of geometry in Acceleration structure, including the depth of AS, or the size of geometry data at the leaf nodes. (Triangles data layout optimization)
   III. Data in AS with **non-lossless truncation or compression**.
3) Cache policy for each memory hierarchy to reduce bandwidth.
   I. Memory footprint is a big problem for raytracing job, we may need a new cache policy for each level of memory cache system.
4) Evolution of ray traversal and **acceleration structure** to handle game scene and animation smartly.
   I. Avoid unnecessary rebuild and update of AS structure.
   II. Skin and skeleton support
   III. Level of detail support to reduce aliasing
   IV. Faster building algorithms for faster traversals
5) Evolution of **material** system
   I. Neural BRDF models.
   II. Neural materials
   III. Level-of-detail filtering
   IV. Important sampling

<u>High-Efficency GPU Physic System on Mobile</u>

Interest

1) Many game logic reply on collision detection and physical reaction for game scene, game objects, and character skeleton system.

Some Opportunities

1) May reduce CPU loading
2) May reduce CPU to GPU writes
3) May chain to GPU based skin-skeleton animation

<u>High-Efficiency GPU-Driven Geometry Rendering on Mobile</u>

Interest

1) Compute based skin-bone animation update for geometry inputs and memory

bandwidth discard.

2) Geometry culling to minimize overdraw and maximize HW geometry capacity utilization
3) Minimize replicate material read and compute
4) Use compute work graph to eliminate barrier sync, empty submits, and avoid worst case allocation.
5) Compute rasterization to beat HW rasterization.
6) Support continuous LOD
7) Support skin vegetation
8) Support dynamic tessellation

Some Opportunities

1) Algorithms like BVH culling and cluster traversal may effectively select visual precision closest clusters and lock edges to prevent from continuous LOD crack problem, and also minimize overdraw via HZB occlusion culling.
2) Deferred material shading pipeline may eliminate replicated material I/O and compute and overdraw to G buffers and minimize material draw calls.
3) Workgraph may eliminate the barrier sync between rasterization and material shading, eliminate empty material shading submits, prevent from worst case memory allocation for material draw parameters and buffers.

Game Frame Interpolation Using Neural Graphics

Interest

1) Occlusion handling for complex scene
2) Game integration
3) Super resolution integration

Some Opportunities

4) Occlusion detection and handling
    I. This is the most common challenging, the edge of the object covers background or is covered by another object. For games, semi-transparent objects are widely used for special effects, such as damage text, HP bar, and NPC icon. A robust method is needed to handle them.
5) Cooperation with in-game information
    I. Unlike static video, game can provide additional information such as depth, opacity, object label, or even in-game motion.
6) Real-time segmentation or object tracking
    I. Fast and small moving objects tend to disappear or get ignored by frame interpolation. Tracking these objects may solve this kind of problem.

7) Super Resolution Integration
    I. Both frame interpolation and super resolution can reduce the power for mobile devices. It is possible to integrate them into an advance system.

Game Frame Interpolation Using AI

Interest

1) Focus on AI MC (Motion Compensation) and target to improve PQ
2) Combine with 3DRS (3D Recursive Search block matching) ME
3) Realtime, low latency and low power consumption for AI model
4) Occlusion handling for MC
5) Game integration
6) Super resolution integration

Some Opportunities

1) Replace MFRC MC by AI to improve known PQ problem, like halo and stair situation.
2) MFRC ME was efficient and low cost, MTK want to leverage this portion know-how.
3) Realtime, low latency and low power consumption
    I. Mobile devices have limited power and latency budget compared to desktop computers. Designing an efficient and realtime NN architecture is the top priority.
4) Occlusion detection and handling
    I. This is the most common challenging, the edge of the object covers background or is covered by another object. For games, semi-transparent objects are widely used for special effects, such as damage text, HP bar, and NPC icon. A robust method is needed to handle them.
    II. Expect AI can repair occlusion hole.
5) Cooperation with in-game information
    I. Unlike static video, game can provide additional information such as depth, opacity, object label, or even in-game motion.
6) MNSR (MTK Neural Super Resolution) Integration
    I. Both frame interpolation and super resolution can reduce the power for mobile devices. It is possible to integrate them into an advance system.
    II. Both MNSR and AI MC will use HW neural engine. Is it possible to merge them for better efficiency?

9.2 Design for X

■ Research Needs Label: [DFX]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

As transistor device scaling and multi-die heterogeneous integration via package innovations continue to drive the rise in design complexity, it becomes ever more challenging to meet multiple requirements in yield, quality, reliability, and energy efficiency. These requirements can be in conflict and trade-off optimizations must be made. For example, to be energy efficient and decrease power density to avoid thermal issues, reduced operating voltage is desired. However, lowered operating voltage also reduces the margin of noise tolerance which increases the potential of failure thus becomes a reliability problem. Design methods, flows, and tools are deployed in both pre-silicon and post-silicon stages to meet the multitude of interacting requirements. In pre-silicon, using models of devices and the manufacturing process, design goals are verified by timing and power integrity analysis tools. Due to modeling inaccuracies and unpredictability, post-silicon validation is done to check consistency with pre-silicon predictions. Inconsistencies encountered are then used to drive improvements in pre-silicon processes for the next iteration. The constant pace of technology change forces continuous iterations of learning between pre-silicon verification and post-silicon validation.

Today, with the trend towards bespoke multi-core/chiplet architecture/silicon optimized for specific application markets, it becomes imperative to adopt a system view comprised of the full hardware (HW) and software (SW) stack. Optimization of individual components without the system perspective is no longer sufficient. Functional safety is an example of a system-level requirement that have cross layer connections with those at the device level. Cloud hyper-scalers started reporting incidences of “silent data corruption” (SDC) in 2021 [1, 2] that can be traced to weak HW components which managed to escape device-level quality assurance. It triggered broad interest in industry and academia [3]. Device voltage/timing marginality is identified as one of the potential root causes [4]. As complex digitization extends into all manners of systems, the issue of SDC can have severe economic and life-threatening consequences. The traditional brute-force approach in fault tolerance such as triple-modular redundancy is cost-prohibitive for most modern systems. The solution can only be developed with a full-stack approach

and collaboration between component suppliers and end-system users. A major “shift-left” direction is called for to bring end-system perspectives into the design, verification, and validation of SW/HW components. As full-system iterative learning is likely to increase greatly in complexity, advances in machine learning and generative AI holds promise to help manage productivity and boost effectiveness.

■ Areas of Research Need

1) Minimum operating voltage Vmin is a key design consideration given today’s emphasis on energy efficiency. Vmin also plays a key role in reliability assurance since sufficient margin must be maintained to tolerate noise, but not too much to waste power. A complex set of interacting factors affect Vmin including frequency, workload, environment, fab variation, and defects. Voltage and timing are also inseparable design parameters. We need a holistic and integrated approach in both pre-silicon and post-silicon methods of Vmin analysis and prediction. Pre-silicon IR-drop analysis needs to model dynamic local effects to accurately capture power grid noise which impacts Vmin margin. In post-silicon, embedded sensors such as process variation-sensitive ROSC and fine-grain timing-margin sensitivity derived from scan OCC patterns [5] provide deep data about device internal conditions. These kinds of information have many applications including (1) replacing pattern-based Vmin binning by fast sensor-based prediction during volume production [6], (2) learning systematic design and fabrication features that limit further Vmin reduction which informs pre-silicon flow improvements, and (3) identifying characteristic signatures of marginal weakness which may cause SDC-related issues in the end-system. Accurate and efficient methods for the above-mentioned applications using the latest ML and AI advancements combined with domain knowledge should be a major focus.

2) Application domains such as automotive and AI/HPC data centers have stringent reliability requirements. As the likelihood of failure increases with rising system complexity, SDC has become a serious unsolved obstacle that impedes reliability assurance. Hyper-scalers spend an inordinate amount of effort to detect and prevent SDC in data centers that operate computing cores in the millions [7]. As effective as DFT-based structural test (ST) has been in achieving high fault coverage of HW components, it is still unable to fully prevent defects escaping into the field [8, 9]. System-level test (SLT) is found to be a necessary complement to ST to fully address SDC [7, 10, 11] even though SLT incurs high development

cost and lacks the methodological rigor of ST. Diagnosing root causes of SLT failures is also a major challenge for design teams. System-level fault modeling is a basic foundation to enable analysis of SLT effectiveness, enhance failure diagnosis, and evaluate fault-tolerant design approaches. As failing behavior at system level tends to be transient in nature, transient fault modeling becomes an important research topic [12]. Current low-level ST only considers static faults for practical reasons. At system level including dynamic clock cycles, the transient fault space explodes enormously. Thus, practical and efficient ways to assess transient fault coverage are needed. Possible approaches include raising the level of abstraction, fast and accurate fault coverage estimation, and development of novel joint HW/SW fault tolerance schemes to achieve optimal results.

■ Reference for Design for X Research Needs: (please see page 55)

9.3 Security

■ Research Needs Label: [Security]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

- **Post-Quantum Cryptography (PQC)** – also known as quantum resistant cryptography, are cryptographic algorithms able to stand up to quantum computing power. In September 2022, the NSA (national security agent) announced CNSA 2.0 (Commercial National Security Algorithm Suite), which announce its first recommendation of post-quantum cryptographic algorithms, and reveal the plan for the transition to quantum resistant cryptography.

- **Physically Unclonable Function (PUF)** – is a physical function that cannot be reproduced in physical way, that for a given input and conditions (challenge), provides a “digital fingerprint”-like output (response), that served as a unique identifier for the chip. It is an important technique to further enhance HW security system.

- **Confidential Compute in SDV** – The central compute in Software Defined Vehicles (SDV) is transitioning from closed system to open architecture where significant volume of additional software is over-the-air (OTA) upgraded for feature enhancement, bug fixes and new service deployment. This means not all post-production software

are trusted by OEM, instead, Independent Software Providers (ISP) may deploy their software via the cloud via App Stores. Additionally, vehicles may be required to interact with their digital twins in the cloud, to resolve long-tail challenges in autonomous driving. Frequently, these ISPs require protection of data-in-use, to prevent proprietary or confidential information from being accessed by OS super user or hypervisor. There are multiple challenges to be resolved to create secure and cost-effective system architecture between virtualization, root-of-trust and confidential computing. We list these identified topics in Section (3) below.

■ Specific areas of interest

1) we are seeking the novel HW (or SW) implementation or co-processor architectural implementation of the following algorithms:

Algorithm	Function	Specification	Parameters
Module-Lattice-Based Key-Encapsulation Mechanism Standard (ML-KEM aka CRYSTALS-Kyber)	Asymmetric algorithm for key establishment	FIPS PUB 203	Use ML-KEM-1024 parameter set for all classification levels.
Module-Lattice-Based Digital Signature Standard (aka CRYSTALS-Dilithium)	Asymmetric algorithm for digital signatures	FIPS PUB 204	Use ML-DSA-87 parameter set for all classification levels.
Leighton-Micali Signature (LMS)	Asymmetric algorithm for digitally signing firmware and software	NIST SP 800-208	All parameters approved for all classification levels. SHA256/192 recommended.
Xtended Merkle Signature Scheme (XMSS)	Asymmetric algorithm for digitally signing firmware and software	NIST SP 800-208	All parameters approved for all classification levels.

2) We are seeking the novel implementation, architecture and/or procedure of PUF that is friendly for mass production and/or easily migrating to different technology nodes.

3) We are seeking studies of confidential computing implementations and focused studies in Automotive central compute applications, such as, but not limited to the following topics:

I. Scalable system architecture to efficiently address requirements of multiple categories of applications, such as Automotive ADAS/AD, Data Center, and in-vehicle entertainment/AI-assistance;

II. Hardware/software system models to enable confidential computing in distributed edge/cloud systems. An sparking example is that multiple cloud-deployed services may utilize sensors in/around a vehicle concurrently, without hindering personal information or causing privacy concerns from the user, and, without hindering individual IPRs of providers;

III. System implementation of TEE Device Interface Security Protocol (TDISP) protocols as defined in PCIe Gen 6, in multi-chip or chiplet devices. Scope of study is not limited in CCA, instead shall expand to complete data chain of in-storage, in-use and in-transport;

IV. Comprehensive analysis and modeling of threats due to intertwined MCS quality of service and Arm Realm-based computing mechanism, identify potential risks and create universal or specific test strategies;

9.4 Virtualization for Functional Safety

■ Research Needs Label: [Safety]

If the proposal is related to more than one Research Need area, please place the label of the primary area upfront.

■ Motivation

- **Virtualization for Functional Safety** – Hypervisor based architecture has become a main solution for automotive central compute processors, where ADAS/AD, Digital Cluster, and Infotainment functions can be integrated into a single SOC via monolithic die or chiplets. In these systems, virtualization provides resource isolation and Quality-of-service for tasks deployed on different virtual machines, and in some cases (with Arm MPAM support), provides task-level isolation based on process IDs. Across all integrated functions, we can classify tasks as safety-critical with FFI (Freedom-from-interference) requirement, real-time, safety-critical, and QM (quality-managed) only. Due to the high-level of safety integration, it remains a major roadblock for silicon designers to create fully integrated central-compute SOCs that support all three applications.

- **Virtualization for mixed-criticality Systems (MCS)** – Some times MCS architecture is regarded as a path to achieve desired safety integration level. It is important to point out that MCS focuses on Worst Case Execution Time (WCET), these techniques help the system to achieve some aspects of safety mechanisms, but it can not address FFI, and WCET is not full representation of Fault-Tolerant Time Interval (FTTI). In this topic, we study challenges of achieve WCET in a high-performance MCS system via virtualization, without the burden to jointly addressing FFI.

■ Specific areas of interest

We are seeking architecture, modeling and performance vs. efficiency tradeoffs studies of state-of-art virtualization architecture in a heterogeneous SOC or chiplet system. Areas of focus may be one or multiple of the following topics:

1) Scalable interconnect technologies to efficiently support safety FFI, large realtime traffic (up to 16 camera inputs and 16 HD displays), and high-bandwidth processors. The interconnect shall efficiently support assumed quality of service with stable performance while still support mixed safety tasks up to ASIL-D level.
2) Scalable architectural proposal across Processing Element (PE), Interconnect and Memory Slave Controllers (MSCs). State-of-art virtualization technologies shall be utilized.
3) System modeling of MCS systems. Popular tools such as Platform Architect [1] [2] can be used, or, if for specific subsystems, System C or equivalent maybe used.
4) Survey, benchmarking, and in-depth studies of field-proving central compute architecture. At the time of this writing, only a small number of Automotive Original Equipment Manufactures (OEMs) deployed such systems. So the comprehensive study will shed light on the practicality and effectiveness of various techniques deployed and enhance of confidence level of such solutions.

■ Reference

[1] Platform Architect from Synopsis. https://www.synopsys.com/verification/virtual-prototyping/platform-architect.html
[2] VLAB works. https://vlabworks.com/

Appendix: 6G Communication Systems

■ Reference

[1] MediaTek Inc., "6G White Paper: MediaTek's vision for next-gen cellular technologies," 31 January 2022. [Online]. Available: https://www.mediatek.com/tek-talk-blogs/6g-white-paper-mediateks-vision-for-next-gen-cellular-technologies

[2] Next G Alliance, "6G Technologies – Next G Alliance," July 2022. [Online]. Available: https://nextgalliance.org/white_papers/6g-technologies/

[3] 3GPP, "Directory Listing /ftp/workshop/2025-03-10_3GPP_6G_WS/Docs/," March 2025. [Online]. Available: https://www.3gpp.org/ftp/workshop/2025-03-10_3GPP_6G_WS/Docs/

[4] MediaTek Inc., "MediaTek Views on 6G Day-1," 3GPP 6G Workshop, 10-11 March 2025. [Online]. Available: https://www.3gpp.org/ftp/workshop/2025-03-10_3GPP_6G_WS/Docs/6GWS-250070.zip

[5] MediaTek Inc., "MediaTek Views on 6G Day-1, Radio Aspects," 3GPP 6G Workshop, 10-11 March 2025. [Online]. Available: https://www.3gpp.org/ftp/workshop/2025-03-10_3GPP_6G_WS/Docs/6GWS-250071.zip

[6] 3GPP, "Specification # 22.837," 28 June 2024. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/22_series/22.837/22837-j40.zip

[7] IMT-2030(6G)推进组, "6G 通信感知一体化评估方法研究报告," 27 October 2023. [Online]. Available: https://www.imt2030.org.cn/html//default/zhongwen/xinwendongtai/1718874199256264706.html?index=4

[8] MediaTek Inc., "MediaTek Views on 6G Day-1, System Architecture Aspects," 3GPP 6G Workshop, 10-11 March 2025. [Online]. Available: https://www.3gpp.org/ftp/workshop/2025-03-10_3GPP_6G_WS/Docs/6GWS-250072.zip

[9] P. Jiang, C.-K. Wen, X. Li, S. Jin and G. Y. Li, "Semantic Satellite Communications Based on Generative Foundation Model," 18 April 2024. [Online]. Available: https://arxiv.org/abs/2404.11941

[10] J. Chen, "Reliable Neuromorphic Computing and Wireless Communication, Ph.D Thesis," August 2024. [Online]. Available: https://kclpure.kcl.ac.uk/ws/portalfiles/portal/286321753/2024_Chen_Jiechen_21054327_ethesis.pdf

[11] brainchip, "Products – Akida Neural Processor SoC - BrainChip," 2025. [Online]. Available: https://brainchip.com/akida-neural-processor-soc/

Appendix: Analog Circuits

■ Reference

[1]

DAC		ADC		PLL	
Parameter	Specification	Parameter	Specification	Parameter	Specification
Resolution	14 bits	Resolution	12 bits	Ref. frequency	491.52MHz
Clock rate	16GHz	Clock rate	16GHz	o/p frequency	3.93 ~ 15.72GHz
o/p impedance	100ohm(diff.)	i/p impedance	100ohm(diff.)	R.M.S. jitter	100fs (10k~100M)
o/p bandwidth	8GHz	i/p bandwidth	8GHz		
o/p power	2dBm	i/p swing	$1.2V_{dpp}$		
IM3	-62dBc@7GHz	IM3	-62dBc@7GHz		
NSD	-156dBm/Hz	NSD	-153dBFS/Hz		

[2] S. T. Kim, et al., “Enabling wide autonomous DVFS in a 22nm graphics execution core using a digitally controlled hybrid LDO/switched-capacitor VR with fast droop mitigation,” ISSCC, pp. 154-155, 2015

[3] J. -S. Paek et al., "A − 137 dBm/Hz Noise, 82% Efficiency AC-Coupled Hybrid Supply Modulator with Integrated Buck-Boost Converter," in IEEE Journal of Solid-State Circuits, vol. 51, no. 11, pp. 2757-2768, Nov. 2016, doi: 10.1109/JSSC.2016.2604296

Appendix: Multimedia

Reference

[1] Y.-H. Lam, A. Zare, F. Cricri, J. Lainema, and M. M. Hannuksela. 2020. Efficient Adaptation of Neural Network Filter for Video Compression. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 358–366. DOI: https://doi.org/10.1145/3394171.3413536

[2] Y.-H. Lam, M. Santamaria, J. Lainema, F. Cricri, R. Ghaznavi-Youvalari, A. Zare, H. Zhang, H. R. Tavakoli, and M. Hannuksela. AHG11: Content-adaptive neural network post-processing filter. Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 Document JVET-V0075. April 2021.

[3] H. Wang, J. Chen, K. Reuze, A. M. Kotra, and M. Karczewicz. EE1-1.4: Test on Neural Network-based In-Loop Filter with Large Activation Layer. Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 Document JVET-W0130. July 2021.

[4] Y. Li, K. Zhang, and L. Zhang. AHG11: Deep In-Loop Filter with Adaptive Model Selection and External Attention. Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 Document JVET-W0100. July 2021.

[5] M. Meyer, J. Wiesner, and C. Rohlfing, “Optimized convolutional neural networks for video intra prediction,” in Proc. of IEEE International Conference on Image Processing ICIP '20, IEEE, Piscataway, Oct. 2020

[6] M. Meyer, J. Wiesner, J. Schneider, and C. Rohlfing, “Convolutional neural networks for video intra prediction using cross-component adaptation,” in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP '19, pp. 1607–1611, IEEE, Piscataway, May 2019

[7] Y. Hu, W. Yang, M. Li, and J. Liu, “Progressive spatial recurrent neural network for intra prediction,” Computing Research Repository (CoRR), 2018

[8] B. Lim, S. Son, H. Kim, S. Nah and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, 2017, pp. 1132-1140, doi: 10.1109/CVPRW.2017.151.

[9] C. Lin, L. Zhang, K. Zhang, and Y. Li. AHG11: CNN-based Super Resolution for Video Coding Using Decoded Information. Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 Document JVET-W0099. July 2021.

[10] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “ESRGAN:

Enhanced super-resolution generative adversarial networks," in Proceedings of the European Conference on Computer Vision (ECCV) workshops, 2018.

[11] F. Galpin, P. Bordes, T. Dumas, A. Robert, P. Nikitin, and F. Le Leannec. AHG11: Deep-learning based inter prediction blending. Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29 Document JVET-V0076. April 2021.

[12] Huo, D. Liu, F. Wu and H. Li, "Convolutional neural network-based motion compensation refinement for video coding", Proc. IEEE ISCAS, pp. 1-4, May 2018.

[13] D. Minnen, J. Ballé, and G. Toderici, 'Joint Autoregressive and Hierarchical Priors for Learned Image Compression', arXiv:1809.02736.

[14] Yoojin Choi, Mostafa El-Khamy, Jungwon Lee, 'Variable Rate Deep Image CompressionWith a Conditional Autoencoder', Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3146-3154

[15] Fei Yang, Luis Herranz, Joost van de Weijer, José A. Iglesias Guitián, Antonio López, Mikhail Mozerov, "Variable Rate Deep Image Compression with Modulated Autoencoder", arXiv: 1912.05526.

[16] Horgan et al, "Vision-based Driver Assistance Systems: Survey, Taxonomy, and Advances, "in 2015 IEEE 18th international conference on Intelligent Transportation Systems

[17] Yurtsever et al, "A survey of Autonomous Driving: Common practices and emerging technologies, "in IEEE Access, Mar. 2020

[18] Xu et al, "Dynamic video segmentation network, "in the IEEE conference on computer vision and pattern recognition 2018

[19] Hong et al, "Virtual-to-real: Learning to control in visual semantic segmentation, "in International Joint Conferences on Artificial Intelligence (IJCAI) 2018

Appendix: Design for X

Reference

[1] H. D. Dixit et al., “Silent Data Corruptions at Scale,” 2021. [online] Available: https://arxiv.org/abs/2102.11245

[1] P. H. Hochschild et al., “Cores That Don’t Count,”, HotOS 2021. [online] Available: https://dl.acm.org/doi/10.1145/3458336.3465297

[2] B. Parthasarathy, “Computing’s Hidden Menace: The OCP Takes Action Against Silent Data Corruption (SDC),” 2024. [online] https://www.opencompute.org/blog/computings-hidden-menace-the-ocp-takes-action-against-silent-data-corruption-sdc

[3] A. Singh et al., “Silent Data Errors: Sources, Detection, and Modeling,” IEEE VTS 2023.

[4] H. H. Chen, “Analysis of Vmin Variability in Complex Digital Logic via Post-Silicon Profiling,” IEEE VLSI-DAT 2023.

[5] C. Chen et al., "Vmin Prediction Using Nondestructive Stress Test," IEEE VTS 2023.

[6] H. D. Dixit, "Detecting silent errors in the wild: Combining two novel approaches to quickly detect silent data corruptions at scale, " 2022. [online] https://engineering.fb.com/2022/03/17/production-engineering/silent-errors/

[7] W. Li et al., "PEPR: Pseudo-Exhaustive Physically-Aware Region Testing," IEEE ITC 2022.

[8] C. Nigh et al., "Faulty Function Extraction for Defective Circuits," IEEE ETS 2024.

[9] D. Lerner et al., "Optimization of Tests for Managing Silicon Defects in Data Centers," IEEE ITC 2022.

[10] M. Shamsa and D. Lerner, "Defect Mechanisms Responsible for Silent Data Errors," IEEE IRPS 2024.

[11] A. Yang et al., "Transient Fault Pruning for Effective Candidate Reduction in Functional Debugging," IEEE ITC 2022.