MediaTek Advanced Research Center

Call for Research

(MARC-CFR)

Research Needs

April. 2026

MARC, MediaTek Advanced Research Center

Research Needs

1. Wireless Technologies 1

2. Radio System Solutions 13

2.1 Wireless RF 13

2.2 6G FR3 Antennas 15

3. Analog Circuits 19

4. High-Performance Compute and AI 22

5. Multimedia 28

6. Heterogeneous Integration for 2.5D/3D Packaging 31

7. EDA 34

8. Data Center 36

9. Special Topic 41

9.1 GPU 41

9.2 Design for X 46

9.3 Security 49

9.4 Virtualization for Functional Safety 51

Appendix: Wireless Technologies 53

Appendix: Analog Circuits 55

Appendix: Multimedia 56

Appendix: Design for X 58

1. Wireless Technologies

Technology exploration for cellular network, space system, wireless LAN, sensing, and other wireless applications

Research Needs Label: [WT]

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

UCP	Universal Commerce Protocol
UE	User Equipment
UL	Up Link
UP	User Plane
XR	Extended Reality

Motivation

The landscape of wireless technologies is undergoing a profound transformation, driven by emerging paradigms and the convergence of diverse technological ecosystems. The Wireless Technologies [WT] area seeks innovative solutions across next-generation mobile cellular networks, short-range and last-mile radio access technologies—including Wireless LAN, Wireless PAN (Bluetooth), UWB, and others—and the ubiquitous space systems such as Non-Terrestrial Networks that provide universal and seamless coverage. We also invite contributions addressing novel wireless applications, such as integrated sensing and communication, as well as high-speed interconnects in both ground- and space-based data centers.

With the opportunities offered by these technologies come inevitably many critical challenges that threaten to impede progress. The list includes, for example, the adoption of higher carrier frequencies above 7GHz in mobile cellular networks that require advanced multi-antenna techniques to make up for path loss. At the same time, short-range and last-mile radio access technologies must contend with rapidly densifying and interference-prone unlicensed environments, where efficient spectrum sharing, coexistence, and interference mitigation become increasingly complex.

Compounding these difficulties is the growing demand for wireless systems that deliver higher reliability, lower latency, and more deterministic performance. User equipment and network devices alike are expected to meet stringent security requirements and achieve greater energy efficiency, all while supporting an expanding array of applications and use cases. Addressing these multifaceted challenges is crucial for ensuring that next-generation wireless networks and emerging applications not only meet, but exceed, evolving user expectations and societal needs.

Recognizing the limitations of traditional methods in furthering the capability boundary of existing technologies [1] [2] [3], we encourage submissions that apply cutting-edge AI and machine learning techniques to address the challenges inherent in next-generation wireless systems. Equally important, on the other hand, are the proposals

that apply wireless technologies in supporting AI applications, particularly in addressing the transformative change in traffic characteristics and contents arising from agentic and physical AI.

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

Physical Layer Air Interface and Medium Access Control [4]

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) UL enhancements
    I. Joint RF/baseband design for UL enhancement, including co-optimization of PA/DPD, RF front-end constraints, and baseband waveform/processing to maximize UL efficiency, linearity, and coverage
    II. 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
    III. Develop PHY mechanisms for utilizing the TX switching to enhance the UL performance

IV. ACLR framework to enable maximum TX power transmission and reduce guard band in 6G carriers
V. Extended coverage design aligning TN and NTN
VI. Techniques to mitigate the impact of Tx impairments on the UL performance, including techniques at the transmitter/UE and techniques at the receiver/base-station (e.g., non-linearity compensation)
VII. UL MIMO schemes that account for user-induced effects on handheld antennas (e.g., grip/body blockage and adaptive antenna selection)
VIII. Joint source–channel coding for UL coverage enhancement, optimizing for UL traffic characteristics (e.g., bursty/short-packet transmissions, latency/reliability targets, and service-specific payload structures)

4) AI/ML for performance enhancement of air interface
  I. AI-empowered Physical Layer (PHY): Use of AI for channel estimation, signal detection, equalization, and adaptive modulation
  II. Intelligent End-to-End Communication: AI-driven source coding, joint source-channel coding, and semantic communication systems
  III. Edge Intelligence and Distributed AI: Federated learning, edge inference, and collaborative AI for distributed communication
  IV. Cross-domain AI with Communication: Integrating AI with IoT, V2X, satellite, and cross-layer system optimizations
  V. AI-based Communication System Prototyping: Demonstration and validation of AI-assisted communication systems in lab/testbed setups
  VI. AI/ML for the optimization of Wi-Fi user experience

5) Unlicensed-band channel access mechanisms
  I. Novel channel access mechanisms for deterministic, sub-millisecond latency applications (e.g., XR or industrial control)
  II. AI-native or sematic-aware channel access
  III. Decentralized channel access for many-to-many information exchanges in volatile ad hoc networks
  IV. PHY/MAC Adaptation and Optimization for next generation Wi-Fi

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 [5]

1) 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

2) Proximity Network

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

II. Device collaboration for collaborative MIMO, positioning or sensing

3) Wireless LANs

I. Distributed protocols for sensor, automotive, and robotic networks, enabling efficient real-time information sharing (e.g. public safety, rural connectivity, robot collaboration, or collaborative sensing)

II. Zero-trust security architectures and lightweight cryptographic protocols tailored for dynamic wireless environments, from the IoT endpoint to the core network

III. Predictive network optimization with novel AI approaches (e.g., digital twin)

4) Seamless heterogeneous networks

I. Frameworks for cross-layer and cross-network Quality of Experience (QoE) mapping and enforcement, ensuring predictable application performance across heterogeneous network segments

II. Architectures for seamless service continuity, security enforcement, and proactive mobility management across the full technology stack, from NTN to WAN, LAN, and PAN domains

III. Design and implementation of "X-as-a-Service" frameworks that can fuse information from multiple wireless technologies (e.g., CSI, RTT, AoA) into a unified, queryable world model such as "Location-as-a-Service" and "Sensing-as-a-Service

IV. AI-driven traffic steering, aggregation, distribution, and load balancing to enhance end-to-end Quality of Experience (QoE) in heterogeneous networks

V. Coexistence protocols for intra-device and inter-device radio coexistence

5) AI Framework and Life Cycle Management

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)

V. Distributed AI frameworks, including federated learning, split inference, and protocols designed for achieving robust, self-organizing networks

VI. Algorithms for self-organizing and self-healing networks that can maintain robust connectivity in the absence of centralized control and mitigate challenges endemic to ad-hoc topologies, such as the hidden node problem and dynamic interference patterns.

Emerging Applications and Transformative Technologies

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. Validation and verification of propagation channel modeling of sensing targets and environment for parameter characterization by real measurements to, at least partially, account for diffraction, scattering, and/or penetration/refraction effects, plus analytical work on near-field sensing signal processing and system design

III. 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

IV. Sensing by the fusion of data from multiple devices such as mobile phones,

watches and glasses, to enhance user experience and/or assist communication

V. Data processing for environment perception, including target identification (automobile or drone) and gesture recognition, using AI/ML techniques

VI. RF fingerprinting based on multi-dimensional MIMO channel for high precision positioning, CSI compression and other novel applications

VII. Integration of sensing, communications and energy harvesting for applications such as indoor navigation, environmental monitoring, biometric identification, or robot collaboration

VIII. AI-native heterogeneous multi-modal information fusion (e.g., from Wi-Fi data/sensing/ranging, cellular data/sensing, GPS) for human-like perception and cognition

IX. Wi-Fi Radar and use cases

X. 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, UCP, OpenClaw, and other competing solutions for agentic AI at cloud and at device

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 for device agent and UE foundation model, 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. Proof of concept to demonstrate device agent for service brokage, UE foundation model for service/UX/performance optimization, and their interaction by intent
VI. Final report summarizing findings, insights, logs (e.g., wireshark), and recommendations

4) Semantic and Token Communication
I. Characterization and modeling of network traffic generated by agentic AI, and the abstraction and exploitation of contextual and semantic information embedded within
II. Semantic Encoding and Decoding: Frameworks for encoding/transmitting semantic content, transcending traditional bit-pipe models
III. Token-based Communication: Methods for mapping messages to tokens (e.g., using LLMs), transmitting them, and reconstructing meaning
IV. AI-driven Semantic Compression and Inference: Using deep learning, LLMs, or graph neural nets for semantic understanding, compression, and tokenization
V. Context-aware and Knowledge-driven Communication: Leveraging context, side information, or knowledge graphs to optimize communication
VI. System Architecture and Prototyping: End-to-end system design, proof-of-concept demonstrators in 6G scenarios including token communication
VII. Performance Metrics and Benchmarking: Defining standards and datasets for evaluating semantic and token-based transmission systems
VIII. Cross-layer Integration: Integrating semantic and token communication into PHY, MAC, and network layers

5) Physical AI
I. Any enabling wireless technologies for physical AI applications, including but not limited to robotics and autonomous navigation
II. Wireless Infrastructure for Autonomous Navigation
III. Vision-Language-Action (VLA) training and inference for robots, preferably considering multi-modal sensor fusion, tactile actions, complex actions by understanding physical ground truth, robot-pose comprehension of itself and other robots/agents
IV. Tokenized communications (or joint sensing and communication) between collaborative robots to accomplish complex physical actions, or safety and resilience of distributed control/operation/coordination among autonomous vehicles, automated guided vehicles, understanding physical ground truth

V. Networking architecture, multiple access, error control, source coding, and traffic characterization for IV

Environmental Sustainability

1) Energy Efficiency
   I. Ultra-Low-Power Radio System that achieves ≤1 µW (≤100 µW) power consumption with ≥−80 dBm (−90 dBm) sensitivity, enables tunable frequencies and robust co-channel interference mitigation, supports multi-user operations and withstands time/frequency variations (potentially with reference signal design and assistance) in severe environment
   II. Distributive Energy-Efficiency Optimization, including but not limited to the optimization of transmission and processing power consumption across base stations and UEs for holistic energy efficiency, minimizing computation and communication overhead by avoiding fully centralized strategies, and the adaptation to 6G networks with cooperative UEs and intermediate nodes, ensuring broad applicability
   III. AI-Assisted System-Wide Energy Efficiency Optimization including, but not limited to, employing AI for cross-domain (time/frequency/spatial/power domains), cross-layer (PHY/MAC/RRC) and BS-to-UE energy efficiency optimization, leveraging distributed learning or shared intelligence to minimize the energy footprint, and balancing 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 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

VI. Secure, low-footprint co-design of communication protocols and endpoint computation for resource-constrained devices, ensuring both efficiency and data integrity

VII. Novel protocols and frameworks for decentralized AI agents that enable autonomous power-aware operations at the network edge

UE Architecture

1) Cellular 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 [8] [9]

   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

Space Technology

1) NTN

   I. Spectrum sharing mechanisms 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 [10]

2) Space Data Center Communication

   I. Inter-satellite or satellite-ground free-space high-speed optical/laser/quantum/radio communications, preferably considering security and alignment of transmitter-receiver, and interference/frequency management in the case of radio communication

   II. Computing and high-speed optical networking/switching architecture design of a space data computing cluster, preferably with physical coordination mechanism of computing containers and clusters

   III. System architecture, orchestration mechanism and protocol design to integrate computing and communication functionalities for AI services

delivery to the massive terminals

IV. Network architecture to integrate multi-orbit NTN connectivity with space data center or AI satellite for efficient data/token distribution

Reference for Wireless Communication: (please see page 53)

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, ultra-high-bandwidth, and extreme power efficient analog circuit continue to be the backbone of next-generation technologies, including 6G/5G-Advanced wireless, AI-driven wireline communications, autonomous automotive, smart home, and Edge AI/GenAI hardware. The key areas include power management, data converters and high speed Serdes. The focus includes innovations in architectures, circuits, and systems tailored for the AI and advanced packaging era.

- Power management: Explore integrated circuits and/or application circuits that could improve power conversion efficiency for application processors, RF power amplifiers, mobile devices, IoT, wearable applications and handle extreme dynamic current transients for AI accelerators (XPUs).

- 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 & Optical): techniques to support 224G/400G+ data rates, ultra-low-power die-to-die (UCIe) links, and Co-Packaged Optics (CPO) / Silicon Photonics over advanced 2.5D/3D heterogeneous integration are of paramount 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/algorithm, TX drivers and low jitter clocking, clock recovery, etc. with state-of-the-art performance (upon normalization over process technology if needed). Silicon Photonics (SiPh) and Co-Packaged Optics (CPO) transceiver

circuits with state-of-the-art energy efficiency are of high interest.

2) 2.5/3D (INFO/CoWoS) interconnect with data rate 64+ Gb/s/wire. Power efficiency <=0.25 pJ/bit @ N3 process and could have normalization over e.g., process technology if needed.
3) Chip-to-chip single-ended communications on substrate, with data rate 64+ Gb/s/wire. Innovative architecture to achieve best power efficiency is highly interested. Simultaneously bi-directional signaling die to die link is also highly interested.
4) 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).
5) Time-interleaved analog-to-digital converter calibration techniques for sampling rate >20Gs/s. (>=10bits)
6) IVR (Integrated Voltage Regulator) for SoC
   I. May include hybrid SC, LDO, etc. Vin=1.2V-1.8V, Vout=0.3V ~ 1V, Iout > 2A [2]
7) 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
   III. Multi-phase fast transient (>2A/0.1us) and high-efficient Buck converter with coupled inductors
8) PMU for ultra-low power wearable applications
   I. Ultra-low quiescent Buck, Boost and Buck-Boost converter with >90% peak efficiency @ I_lout=0.5A and small BOM area
   II. Single-inductor-multiple-output (SIMO) buck converter with high peak efficiency and small BOM area
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) Compact current monitor for dynamic digital circuits in high-performance computing, with <5% measurement error while considering the RLC effects of the power delivery network.
12) Circuits and systems for analog AI, CIM, etc. that support AI computing acceleration and non-conventional computing.

13) Reliable and functional safety circuit design for automotive applications.
14) Generative AI-Powered Analog Design Methodology: Research utilizing LLMs or Generative AI agents for automated analog circuit sizing, layout generation, parasitic extraction prediction, and performance verification boost to significantly reduce design cycle time.

■ Reference for Analog Circuit Research Needs: (please see page 55)

4. High-Performance Compute and AI

■ Research Needs Label: [HPC]

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

■ Motivation: Powering the AI Revolution: Scaling High-Performance Compute for Next-Generation Intelligence

The contemporary landscape of Artificial Intelligence and advanced computing is undergoing a profound and accelerating transformation, impacting nearly every facet of modern life, from global communication and immersive entertainment to autonomous systems and intricate industrial automation. Recent breakthroughs in Generative AI (GenAI), Large Language Models (LLM), Large Vision Models (LVM), Agent AI, and Physical AI are fundamentally reshaping the capabilities and applications of AI, leading to an unprecedented and exponential demand for computational power. MediaTek's Edge AI platform is strategically positioned at the vanguard of this revolution, integrating sophisticated AI functionalities into a broad spectrum of edge devices, including smartphones, wearables, televisions, smart speakers, tablets, computers, and automobiles. Furthermore, our scope extends to data center applications, encompassing Cloud/Edge AI inference servers. smart speakers, tablets, computers, and automobiles. Furthermore, our scope extends to data center applications, encompassing Cloud/Edge AI inference servers.

However, the escalating computational intensity of these emerging foundation models is rapidly widening the gap between AI’s escalating demands and the inherent constraints of semiconductor scaling. As Edge AI applications proliferate, the required computing power frequently exceeds what current semiconductor process scaling alone can provide. To effectively address this critical challenge, MediaTek is actively advancing hardware architectures and optimizing the synergistic co-design of software and algorithms through continuous improvements. Our unwavering commitment is to foster groundbreaking innovation, cultivate exceptional talent, and proactively shape the future of AI and computing technology. We invite esteemed academic and industrial experts to engage in collaborative research programs that push the boundaries of this dynamic field.

MediaTek has already demonstrated significant industry leadership, with our NPUs powering over 800 million intelligent devices since 2019, spanning smartphones, TVs, surveillance, and AIoT applications. Our robust NeuroPilot ecosystem boasts over 10,000

registered developers from more than 600 companies, and continues to grow. The NeuroPilot platform's user-friendly SDK is engineered for rapid deployment of various AI models and GenAI applications on MediaTek NPUs. This capability extends across diverse applications, including camera AI, voice processing, display enhancements, gaming, communication, and system control.

We cordially invite proposals for transformative research that drives the development of next-generation algorithms and architectures across diverse application domains. Our focus is on advancing applied algorithms, optimizing system-level performance, and exploring novel compute architectures for emerging AI models. We particularly welcome innovative research in data collection, synthesis, benchmark methodology, and algorithm-hardware co-design for edge devices. Proposals integrating scalable machine learning cores with energy-efficient hardware architectures, extending to advanced NPU system designs integrated with CPUs, GPUs, MCUs, and memory architectures, are highly encouraged. Interdisciplinary collaborations that offer high innovation value and originality, with the potential to create robust, efficient, and scalable AI platform solutions, are also of strong interest. The following sections outline our critical research needs for High-Performance Compute & AI platform technology.

■ Areas of Interest (including but not limited to)

Model Architecture for Emerging Applications

The top level focuses on the design, optimization, and efficient deployment of advanced AI model architectures, addressing the computational and performance demands driven by GenAI, LLM, LVM, Agent AI, and Physical AI across both edge and cloud environments.

1) **New Foundation Model/Architecture (e.g., Mamba, Gated/Linear Attention, MoE)**

   We seek exploration into novel foundation models and attention mechanisms, such as Mamba and Gated/Linear Attention, which offer significant promise for enhanced efficiency and performance in sequence processing, critical for scaling AI capabilities on diverse platforms. Research should focus on their inherent advantages, limitations, and architectural adaptations for both training and inference.

2) **Generative AI on Edge Devices**

   Research is paramount for enabling robust Generative AI capabilities directly on

resource-constrained edge devices. This includes, but is not limited to, supporting high-fidelity video and 3D content generation, handling complex multimodality, processing long contexts efficiently, facilitating sophisticated reasoning capabilities, developing agent-based systems, and establishing LLM operating systems. Efficient representations and architectures that minimize memory footprint and computational overhead are key.

3) **Autonomous Driving and Smart Cockpit Applications**

Advancements are crucial in developing end-to-end ADAS (Advanced Driver-Assistance Systems) models and sophisticated AI for smart cockpit functionalities, such as advanced automotive AI assistants. This requires robust, real-time, and energy-efficient perception, prediction, and decision-making models.

4) **Embedded AI and Vision-Language-Action (VLA) Foundation Models**

This area is crucial for developing highly efficient and high-performance embedded vision solutions for edge devices, including advanced computational photography techniques that enhance camera AI and egocentric vision capabilities. This foundational work enables robust real-time image/video enhancement and understanding. Building upon these capabilities, this domain further involves creating AI systems that can interact intelligently with the physical world through advanced perception, language understanding, and actionable outputs. Research on VLA foundation models is therefore essential for enabling adaptive behaviors in robotics, autonomous systems, and other physical AI entities, requiring seamless integration of vision and language on resource-constrained platforms.

5) **Test-time Computing for Language and Vision**

Test-time computing is essential for dynamically enhancing AI capabilities during inference, pushing the boundaries of what edge device can achieve. This includes improving inference quality by enabling Chain-Of-Thought inferences to generate intermediate prompts and refine final outputs, aligning with the paradigm shift towards Reasoning AI. It is also critical for performance improvement, specifically for accelerating token generation per second through optimal utilization of computational and memory resources via techniques like speculative decoding. Beyond inference enhancement, test-time training for personalized or specialized capabilities enable rapid model adaptation and personalization directly on edge devices without extensive re-training. Finally, exploring dedicated hardware support for test-time scaling is vital, for edge devices to efficiently manage

dynamically scaling AI workloads within complex multi-device architectures or high-performance inference servers for multi-user leveraging heterogeneous runtimes.

6) **Agentic AI and Multi-Agent Systems**

This research area focuses on the development and optimization of autonomous, goal-oriented AI agents and their applications. Building upon breakthroughs in Agent AI and the paradigm shift towards Reasoning AI, this domain explores advanced capabilities such as intelligent perception, sophisticated reasoning, adaptive planning, and autonomous execution. A key emphasis is placed on multi-agent collaboration and coordination for complex tasks, spanning from individual intelligent assistants to interconnected agent ecosystems. Research should address the architectural principles, interaction protocols, and learning mechanisms necessary for robust and efficient agentic behaviors, particularly when deployed on resource-constrained edge devices and in mixed edge/cloud environments. This includes advancements in agent architectures that leverage advanced foundation models for enhanced autonomy and decision-making.

ML Core / HW-SW Co-optimization

The middle level focuses on optimizing the core machine learning computations and the synergistic co-design of hardware and software to maximize efficiency, performance, and scalability, especially for compute-intensive AI models.

1) **LLM/LVM Acceleration**

Dedicated research into accelerating Large Language Models and Large Vision Models is essential to meet their demanding computational requirements and enable their widespread deployment across edge and cloud platforms. This includes exploring novel dataflows, parallelism strategies, and architectural enhancements.

2) **Low-Bit and Efficient Representation for Generative AI on Edge Devices**

To enable sophisticated GenAI on resource-constrained edge devices, research is needed on low-bit precision (e.g., Floating, Integer, or specialized numerical representations) and highly efficient data representations for edge intelligence. This includes investigating techniques for improved performance, reduced memory bandwidth, and enhanced power efficiency without significant accuracy degradation.

3) High-performance Inference Server

Developing technologies for high-performance inference servers, including multi-user multi-device architecture/algorithm co-design, is critical for both dedicated edge inference servers and scalable cloud-based AI solutions. This involves managing concurrent workloads and optimizing resource allocation.

4) **Compiler Optimization and Scalability**

Optimizations for advanced AI compilers, such as TVM/MLIR, including their integration with frameworks like Vulkan ML, are crucial for efficiently mapping diverse AI models onto heterogeneous hardware accelerators. This area is vital for bridging the gap between high-level model descriptions and low-level hardware execution. Furthermore, research into technologies that enable these AI compilers to scale effectively is paramount. This addresses the imperative to handle increasingly complex AI models, larger datasets, and diverse hardware architectures, thereby tackling the challenges posed by growing model sizes and architectural heterogeneity in the evolving AI landscape.

5) **Multi-core Computing Technology and Runtime Optimization**

Enhancing multi-core computing technologies, including advanced runtime optimization and intelligent schedulers for multi-tasking AI workloads, is vital for efficient resource utilization on NPUs and other heterogeneous platforms. This includes dynamic task scheduling and power management.

6) Algorithm-Hardware Co-design for Emerging AI Models

This involves a holistic approach where algorithms and hardware architectures are developed in conjunction to achieve optimal performance, power efficiency, and scalability for emerging AI models.

Hardware

The bottom level focuses on foundational hardware advancements that underpin high-performance AI computing, with a particular emphasis on NPU architectures, memory systems, and advanced packaging technologies.

1) NPU Multi-core Scalability

Research on scaling up NPU multi-core architectures is crucial to meet the rapidly growing computational demands of modern AI, ensuring optimal performance across various workloads and device classes. This includes investigating inter-core communication, synchronization, and load balancing.

2) **On-the-fly Activation Compression/Decompression for Edge Devices**

Techniques for dynamic compression and decompression of activations are vital for efficiently managing memory bandwidth and reducing latency on edge devices, particularly for large AI models. This minimizes the burden on the memory subsystem and improves overall power efficiency.

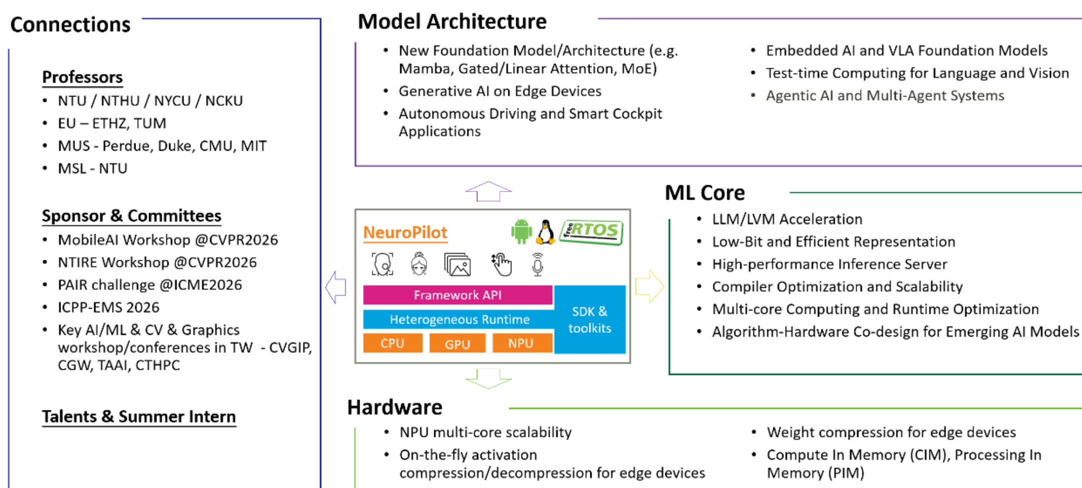
3) **Compute-In-Memory (CIM) and Processing-In-Memory (PIM)**

Exploring and developing CIM and PIM technologies is essential for achieving ultra-low power AI computing by significantly reducing data movement between memory and processing units. Current academic projects demonstrate particular interest in CIM for Transformer Networks and LLMs, highlighting its potential for substantial power and latency reductions.

We firmly believe that collaborative research in these pivotal areas will enable MediaTek to continue pioneering new frontiers in foundational research and system design. This strategic approach will allow us to overcome the inherent challenges posed by advanced AI systems and sophisticated hardware, thereby solidifying our leadership in the High-Performance Compute & AI platform technology.

■ Reference for High-Performance Compute and AI Research Needs:

(HPC) Research Needs 2026



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]
6) PINN for image/video coding or processing [16]

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 56)**

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. 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/...)

8. Data Center

■ Research Needs Label: [DC]

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

■ Motivation

The rapid proliferation of artificial intelligence workloads has fundamentally transformed the landscape of modern data centers, creating unprecedented demands that far exceed the capabilities of conventional infrastructure. As AI models grow in scale and complexity — from large language models (LLMs) with hundreds of billions of parameters to multimodal foundation models requiring massive training runs — data centers have become the critical backbone upon which the entire AI ecosystem depends. In the other end, the demand of inferencing also grows dramatically, it brings the critical challenges on service level requirements on long context length, latency, massive concurrent requests and so on. For token-economics, the pursuit for optimizing power & performance efficiency, TCO & operation cost is the never-ending topic. This surge in AI-driven computation has exposed a constellation of pressing challenges that motivate cutting-edge research across multiple dimensions.

■ Potential areas of interest (but not limited to)

Interconnect & Optics:

The exponential growth of AI demands a paradigm shift in data interconnects. We invite research proposals on novel physical layer technologies for ultra-high-density optical links, including but not limited to CPO, micro-LED, and VCSEL arrays. The primary goal is to achieve breakthroughs in I/O density, power efficiency, and cost-effectiveness, enabling the massive connectivity required for future-generation, large-scale AI systems and overcoming the limitations of current electrical interconnects. Examples topics for this category: DSP algo in EIC for optical impairments in CPO, Advanced modeling/method in E2O2E link simulators by considering EIC & PIC integrated in CPO, Low power architecture for multi-rate Serdes DPMA across electrical and CPO systems.

DC Networking:

1) **Network fabric innovations for AI/HPC**, including ultra-low latency design, collectives operation optimization, advanced load balancing, congestion control mechanism, large fabric management & orchestration, early fault prediction and

resolution, advanced network fabric simulator and so on.

2) **Network service integration for AI/HPC**, including compute in the network, memory tiering and pooling over network, storage optimization & tiering over network.

DC grade thermal solution:

Thermal design for data centers has specific problems that need to be addressed as we towards 2 types of package designs – larger packages and stacked packages. These have to be tackled along with an increase in power density for those designs. Towards that end, we need to specifically target certain areas for research.

1) **Higher conductive filler materials**
   For stacked packages, heat needs to be effectively transmitted across the different layers of the packages to the thermal solution. Currently, filler material conductivity increases thermal resistance to heat transfer and improvements to filler material conductivity can help increase cooling capability of stacked packages.
2) **Local hot spot improvements for single phase cold plate designs**
   For col plate solutions, there are limitations from overall cooling standpoint as well as due to localized hot spots. The localized hot spots end up tending to limit the overall solution. Improvements to the cold plate designs for local hot spots ranging from a power density of 3W/mm2 to 6W/mm2 will be critical for future DC package design.
3) **TIM designs for high warpage packages**
   With increasing total power, the temperature rise in TIM is still critical. While there have been improvements in TIM performance, they do not consider the effect of warpage on performance. Designing for the thinnest BLT alone does not improve performance, nor does increasing conductivity without improving compressibility. The new materials would need to be able to for the size and warpage expectations of DC packages.

System level performance & TCO modeling:

The advanced modeling system aimed to provide insights on sensitivity analysis and design space exploration for the combination of components (e.g. CPU, XPU, network switch, storage) inside the large AI/ML system. Research topics like innovations on E2E ML/AI system performance, workload trace collection and correlations, AI based DSE, LLM serving/scheduling strategy exploration, power and carbon aware modeling, heterogenous computing, heterogenous workloads modeling, computation and

communication overlapping and so on.

Advanced memory technology & solutions:

Proposals are encouraged across a wide spectrum of research areas. The following technical domains are of particular interest to MTK. This list is illustrative and not exhaustive; investigators are encouraged to propose research in adjacent or emerging areas that align with the program's objectives.

1) **Emerging Memory Device Technologies**

   Research into new materials, device structures, and physical phenomena that enable memory behavior beyond DRAM and NAND flash, including but not limited to:

   I. Phase-Change Memory (PCM) and chalcogenide-based storage class memory
   II. Resistive RAM (ReRAM / RRAM) and oxide-based memristive devices
   III. Spin-Transfer Torque Magnetic RAM (STT-MRAM) and Spin-Orbit Torque MRAM (SOT-MRAM)
   IV. Ferroelectric RAM (FeRAM) and ferroelectric field-effect transistors (FeFET)
   V. Electrochemical RAM (ECRAM) for analog and neuromorphic applications
   VI. Two-dimensional and van der Waals material-based memory devices
   VII. Molecular, DNA, and biological-substrate memory for archival storage

2) **Memory Architecture and System Design**

   Novel architectural approaches that reshape the relationship between processors and memory subsystems:

   I. Processing-in-Memory (PIM) and near-data computing architectures
   II. Compute Express Link (CXL) memory pooling and disaggregated memory fabrics with compute in CXL buffers
   III. 3D-stacked memory integration and heterogeneous memory hierarchies
   IV. Universal memory concepts bridging volatile and non-volatile storage
   V. Memory controller design for quality of service and latency predictability
   VI. Storage-class memory integration with operating systems and runtimes

3) **AI and Neuromorphic Memory**

   Memory systems tailored for the unique demands of machine learning, neural networks, and biologically inspired computing:

   I. Analog/in-memory computing for efficient matrix-vector multiplication
   II. On-chip memory optimization for large language model inference
   III. High-bandwidth memory (HBM) architectures for AI accelerators
   IV. Non-volatile weight storage for continual and on-device learning

4) **Reliability, Security, and Endurance**

   Techniques to improve the dependability and security of memory

technologies across their full lifecycle:

I. Error correction and fault tolerance for high-density memory arrays
II. Wear leveling and endurance management for non-volatile memories
III. Hardware-enforced memory encryption and integrity verification
IV. Row hammer mitigations and physical attack resistance
V. In-field reliability modeling and predictive maintenance for memory systems

5) **Photonic and Quantum Memory**

Exploratory research into fundamentally new modalities for information storage:

I. Optical and photonic RAM leveraging phase-change or electro-optic materials
II. Quantum memory for entanglement storage and quantum repeater applications
III. Cryogenic memory compatible with superconducting quantum processors
IV. Hybrid classical-quantum memory interfaces

6) **Sustainability and Energy Efficiency**

Approaches to dramatically reducing the energy and environmental footprint of memory technologies:

I. Ultra-low-power standby and retention mechanisms
II. Energy harvesting and self-powered non-volatile memory
III. Life cycle analysis and sustainable materials for memory manufacturing
IV. Thermally efficient packaging and thermal management for dense memory

DC Silicon manufacture & operations:

1) **Advanced Optical DFx (Quality & Reliability) – To aid CPO testing at scale (Low-Volume Photonics to High-Volume Semiconductor Manufacturing)**

I. Ex: Silicon Photonics BIST – On-Chip BIST to test MZI (Mach-Zehnder Interferometers) & Ring Resonators
II. Ex: Optical Telemetry – RAS capability at Rack & beyond
III. Optical Loop-Back Testing – Non-Contact optical health check as part of Boot sequence
IV. Etc.

2) **Agentic AI in DFT – Faster TTM, Improved Test Coverage, Less Resource, Continuous Improvement**

I. Analyze codebase, generate test plans, execute tests, debug failures and self-heal test scripts
II. Ex: Autonomous Test Generation & Self-healing Test Scripts
III. Ex: Debug & Root-Cause Analysis
IV. Ex: Test Monitoring & Optimization etc.

3) **Silent Data Corruption (SDC) – In containing growing FIT (gen over gen)**

   I. Enhanced Testing/Screening vs SDC screening Efficacy: ATE Efficacy vs SLT Efficacy vs Field Escapes

   II. Periodic In-Field Testing

4) **Enhanced DPPM model – For more realistic estimation of Pre-Si SIP DPPM**

   I. Models beyond existing ones (Williams-Brown: 1981, Agrawal: 1981)

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 radiance cache
8) Neural physic collision avoidance
9) Neural post-processing/Neural AS LOD
10) Neural Shading w/ cooperative matrix and cooperative vector

Some Opportunities

1) Modern GPU architecture to accelerate neural graphics applications.
2) Algorithm/network model trade-off between performance/power/area and picture quality
3) Machine learning become important to act like a universal function walkthrough all samples and do subsample for a similar path traced, texture sampled, material sampled, or geometry subdivision, and MLP is suitable for such purpose and shall be work well in newly supported neural shading functions to be called in a shader such as cooperative matrix and vector to provide high detailed quality rendering.
4) Neural shading is quite essential to insert neural computing inside shaders and minimize bandwidth and latency between GPU and neural computes. We seek for algorithms fulfill key applications to boost rendering as a head light in front of our GPU architecture design and development.
5) Cooperative matrix and vector are very key to neural shading and supported by our GPU to make threads work together to come out a matrix or vector multiply and add.

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

High-Efficency GPU Physic System on Mobile

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

High-Efficiency GPU-Driven Geometry Rendering on Mobile

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. Design-for-X (where X stands for any of the previously stated requirements) is a well-established and effective approach to close the pre- and post-silicon gap.

Today, with the trend towards bespoke architecture/packaging/die 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, validation, and testing 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.

■ Research Need – defect coverage of power distribution networks

Current DFT methodology for digital designs has focused on using structural fault models to test logic circuits via scan access and using BIST for memories (MBIST) targeting memory-specific fault models. A neglected area is the network that supplies power to logic and memory circuits. At advanced nanometer nodes, defects in the power distribution network are having a noticeable impact on quality and reliability resulting in higher SLT and customer DPPM.
[*BIST := built-in-self-test*] [*DFT := design-for-test*]
[*SLT := system-level test*] [*DPPM := defective-parts-per-million*]

A typical power distribution network contains thousands of transistor power switches controlled by a parallel set of series-connected inverter delay chains merging into a single digital output to acknowledge power-on/off. The chains cap the amount of rush current flowing during power-on/off to maintain stability and avoid stress damage. Inherent delay of the chains gradually powers on/off sections of compute circuitry spread out over a period of time. Though gross stuck-at faults in the delay chains can be detected by a few simple functional patterns that observe the acknowledge output, many kinds of physically marginal defects can have a much subtler impact without causing obvious functional failure. Specifically, one can view the power switch (PSH) network as a circuit block with thousands of analog outputs. The in-line resistance of each analog output during power-on is expected to lie beneath a low impedance threshold. However, defects in the switch itself or the switch-control logic can cause the in-line resistance to become abnormally high, thus increasing local voltage drop for

nearby logic/memory circuits. The degraded power integrity can result in intermittent failures when adverse workload and operating conditions are triggered. These failures are almost impossible to predict and replicate; and make root-cause diagnosis notoriously difficult and costly. Marginal defects in power distribution circuitry contribute to SDC in data centers and impede zero-DPPM goal for automotive products.

To a rough first approximation, the Williams-Brown equation relates DPPM to defect coverage: $DPPM/10^6 = 1 - Y^{(1-T)}$ where Y and T are fractional numbers in [0, 1]. Y is manufacturing yield and T is defect coverage. Overall T is an area-weighted sum of logic, memory, and PSH coverages: $T = A_{logic} \times T_{logic} + A_{memory} \times T_{memory} + A_{PSH} \times T_{PSH}$ where $A_{logic} + A_{memory} + A_{PSH} = 1$. For illustration, assume $Y = 90\%$, $A_{logic} = 60\%$, $T_{logic} = 98\%$, $A_{memory} = 35\%$, $T_{memory} = 99\%$, and $A_{PSH} = 5\%$. One can compute DPPM for $T_{PSH} = 0\%$ versus 70%. When $T_{PSH} = 0\%$, then $T = 93.5\%$ and DPPM = 6877. When $T_{PSH} = 70\%$, then $T = 97\%$ and DPPM = 3208. Raising PSH defect coverage from 0% to 70% can cut DPPM by more than half in this illustrative case with the assumed yield, area, and coverage parameters.

Research related to raising PSH defect coverage should address the following:

1) Comprehensive analysis and characterization of defect behavior in PSH networks. This will involve extensive SPICE-level defective circuit simulation.
2) Develop PSH DFT and production test methods to achieve high defect coverage.
3) Due to the analog nature of defects, PSH test methods will likely involve statistical outlier analysis and decision-making. Such a method may require in-line calibration to accommodate process variability.
4) Adding custom PSH-BIST circuitry to the design is one possible DFT approach. Careful consideration must be given to the impact of PSH-BIST on area, performance, test time, and yield. Enabling in-field PSH-BIST would also boost overall comprehensiveness of the solution.

■ Research Need – DFT methodology for silent-data corruption (SDC)

The SDC issue in high-performance computing for AI training and inference has reached an alarming level. Existing test methods are no longer adequate as data center DPPM is more than 10× industry target and threatens reliable computing [5]. New novel methods in design and test need to be explored to address this critical issue: (1) quick diagnosis of SLT fails, (2) in-field detection of bad chips under varying conditions, and (3) test experiments to fill knowledge gaps and validate new approaches. A promising direction is the idea of system health profiling which is being developed by cloud

service providers. This can be "shift-left" by collecting fine-grain chip-level margin profiles and correlating to system failures [6, 7].

Reference for Design for X Research Needs: (please see page 58)

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:

    III. 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;

    IV. 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;

V. 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;

VI. 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: Wireless Technologies

■ Reference

[1] D. W. Tseng and H. Chu, "Pioneering the Future with Wi-Fi 8: Part 1," October 2024. [Online]. Available: https://www.mediatek.com/hubfs/MediaTek%20Assets/Pdfs/White_Papers/MDT3011_Pioneering_the_Future_with_WiFi8.pdf?hsLang=en

[2] D. W. Tseng and H. Chu, "Pioneering the Future with Wi-Fi 8: Part 2," 15 February 2025. [Online]. Available: https://www.mediatek.com/hubfs/MediaTek%20Assets/Pdfs/White_Papers/Wi-Fi%208%20Part%202%20Whitepaper.pdf?hsLang=en

[3] D. W. Tseng and H. Chu, "Pioneering the Future with Wi-Fi 8: Part 3," 25 November 2025. [Online]. Available: https://www.mediatek.com/hubfs/MediaTek%20Assets/Pdfs/White_Papers/Pioneering%20the%20Future%20with%20Wi-Fi%208_%20Part%203.pdf?hsLang=en

[4] 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

[5] 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

[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] 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

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

[10] 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.

[11] D. W. Tseng and H. Chu, "Pioneering the Future with Wi-Fi 8: Part 1," October 2024. [Online]. Available: https://www.mediatek.com/hubfs/MediaTek%20Assets/Pdfs/White_Papers/MDT3011_Pioneering_the_Future_with_WiFi8.pdf?hsLang=en

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

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

[3] G. Nishant et al., “Silent Data Corruption in AI,” Open Compute Project 2025. [online] Available: https://www.opencompute.org/documents/sdc-in-ai-ocp-whitepaper-final-pdf

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

[5] S. Mitra et al., “Silent Data Corruption by 10x Test Escapes Threatens Reliable Computing,” Google 2025. [online] Available: https://arxiv.org/abs/2508.01786v4

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

[7] H. H. Chen, “Test Challenges and Direction in the Age of AI Everywhere,” Keynote Talk, TestConX China 2025. [online] Available: https://www.testconx.org/premium/wp-content/uploads/2025china/TestConXChina2025s3p0ChenKeynote_8233.pdf