MediaTek Advanced Research Center

Call for Research

(MARC-CFR)

Research Needs

April. 2024

MediaTek

Strategic Technology Exploration Platform (STEP)

Research Needs

1. 6G Communication Systems

Key technology explorations for 6G Communication System Design

■ Research Needs Label: [6GSys]

■ Motivation

1) 6G cellular communications network is expected to support higher spectral efficiency, higher data throughput, lower end-to-end latency, lower power consumption, more robust waveform and more secure than 5G.
2) Evolutionary and/or revolutionary techniques are required to mitigate current NR technology gap in terms of capacity, coverage and efficiency for eMBB/vertical use cases and new application drivers such as cloud gaming and XR in FR1, FR2 and even higher frequency spectrum [1].
3) To achieve these objectives of 6G, several potential research areas of interest are listed below for reference.

■ Potential areas of interest but not limited to

1) New MIMO and multiplexing techniques [2]
   I. Channel prediction for outdated CSI due to UE mobility
   II. Design a communications system with BS/UE-controlled RIS (Reconfigurable Intelligent Surface) for developing RIS channel model used in simulations and for evaluating system coverage and throughput improvement especially for the deployment in frequency bands higher than FR1.
      - Proof of concept prototyping of above system to validate the system gain
      - Consideration of near-field communication effects and its impact on system design
   III. Design a communication system with UE-side full duplex, including single-frequency full duplex and sub-band full duplex to enhance system latency and spectral efficiency with form factor, cost and power constraint.
      - Proof of concept prototyping of above system to validate the system gain
   IV. MIMO system design (incl. transmission schemes and CSI reporting, etc.) and its evaluation, taking into account the scalability of the dimension of antenna arrays, for both centralized and distributed deployment topology. For the latter case, also consider practical impairments such as timing/frequency synchronization error among the geographically separated arrays.

V. Efficient beam management and CSI acquisition framework for (centralized and distributed) massive MIMO for RS/feedback/latency overhead reduction.

VI. Radio propagation channel measurement and modeling for new frequency bands such as 7 to 15 GHz band and sub-THz band, taking into account the factors such as near-field effect and very large array dimension.

2) Fusion of communications, computing and sensing (including high-precision positioning) [3][4]

I. Sensing assisted communication: identify typical use cases (especially UE sides or UE assisted use cases), evaluate feasibility, summarize challenges and potential solutions. For example, UE estimates positions of obstacles, and combine the position information of BS and UE, to assist beam management, cell reselection/handover operations. Statistical property of channel is obtained during sensing tasks, and which can assist communication side to improve performance or reduce complexity.

- Proof of concept prototyping of above system to validate the system gain

II. Near-field effects and their impact to sensing system design: characterize and model near-field channel mathematically and develop corresponding signal processing and sensing techniques for channel estimation, beam focusing, and localization.

- Proof of concept prototyping of above system to validate the system gain

III. Sensing of device-to-device: design system architecture and signal processing procedure, summarize challenges and solutions. For example, modify sidelink system to support integrated sensing and communication (ISAC).

- Proof of concept prototyping of above system to validate the system gain

IV. UE cooperative sensing: multiple UEs use their positioning information and sensing results to cooperatively identify objects and estimate locations, and further use the information to assist some sensing applications, or to assist communications.

- Proof of concept prototyping of above system to validate the system gain

V. AI assisted sensing: apply ML-based methods sensing algorithm, or enhance sensing performance with AI, for example, motion or gesture recognition.

- Proof of concept prototyping of above system to validate the system gain

3) Non-terrestrial network

I. Design the satellite/cellular spectrum sharing mechanism to improve overall spectral efficiency while managing the interference

II. Design new waveform applicable for 6G non-terrestrial networks which can well coexist with terrestrial network and optimized for satellite channels.

4) Artificial intelligence for communications
   I. Design and evaluate solutions for AI-enhanced physical layer performance/ robustness or AI-enabled features to achieve system optimization, targeting for 3GPP Rel-19 AI study item or beyond
   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 networks (e.g., creating networks on the fly based on required services/applications).
5) Cross-layer optimizations for future applications
   I. Codec Avatar Platform
      - Design algorithm to implement Codec Avatar using AR/VR devices aiming to connect people in the Metaverse with photorealistic virtual avatars
      - Design AI enhanced Super-Resolution using machine learning to clarify, sharpen, and upscale 3D Avatars to increase user experience
   II. SLAM Technology development
      - Develop algorithm between edge server/device or device/device for split rendering and SLAM to improve network capacity and E2E latency
   III. XR Platform
      - XR applications proof of concept. To verify user experience enhancement for cross-layer optimization
      - Develop cross-layer optimization algorithm for streaming adaptation to improve user experience in mobile network and standardize cross-layer API for B5G/6G
6) New network architecture for Integrated communication and computing (ICC)
   I. Explore architecture options to integrate cloud platform, e.g. kubernetes, with 3GPP service architecture
   II. ICC proof of concept to demonstrate UE triggered computing offload.
7) Energy efficiency and harvesting
   I. Design ultra-low-power radio system based on radio technology of <= 1 μW active power consumption with at least -80 dBm receiver sensitivity [5]. The system should achieve tunable frequency, mitigation of interference from/to legacy co-channel users, multiple access functionality and robustness against

time/ frequency uncertainty.

- Develop proof of concept platform to verify performance

II. Design of distributive optimization for system energy efficiency, where joint optimization is over the usage of the transmission power and processing power of all base-stations and UEs, to resolve the excessive complexity for a centralized optimization (computation burden, communications overhead). Given 6G network evolution includes distributed network and more intermediate network nodes, the distributive optimization scheme should be applicable to networks featuring distributed MIMO and/or UE collaboration operations.

III. AI-assisted optimization of system energy efficiency: Focus on leveraging AI to optimize energy efficiency at various levels, from task scheduling and benchmarking to energy harvesting, collaborative architectures, and hardware design. By exploring these novel approaches, researchers can develop innovative solutions that minimize the energy footprint of AI systems while maintaining high performance.

8) Next generation security, trustworthy and privacy preserving systems

I. Design information-theoretic/physical layer-based mechanism and secret key distribution protocol, such as QKD (Quantum Key Distribution), to integrate with asymmetric security using novel PQC (Post Quantum Cryptography) algorithm to guarantee E2E security level even after Quantum computer is available.

II. Analyze and evaluate 3GPP standard and implementation impacts based on the lightweight cryptography algorithm to be defined by NIST.

III. Novel approaches to overcome privacy concerns in distributed Machine Learning wireless systems.

IV. Novel algorithmic frameworks for communication-efficient and differentially private federated learning wireless systems with applications to real-world use cases.

9) Carbon-Aware System Operation

I. Develop carbon related KPIs and create an end-to-end evaluation methodology (from mobile device to network) to accurately determine the 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. This involves examining how these two systems can work together to lower carbon emissions effectively.

III. Investigate a carbon and Quality of Service (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 the energy-related characteristics of mobile communication systems. This includes tracking energy consumption, the mix of energy supply (such as renewable and non-renewable sources), and carbon intensity, while considering various spatial and temporal granularities.

V. Design carbon-aware resource management strategies for next-generation communication systems, incorporating computing and sensing aspects, among others. This involves using energy-related criteria to guide the allocation and management of resources, with a focus on minimizing carbon emissions.

10) 6G modem system architecture

I. Architecture exploration of 6G modem IP with the key directions identified, including processors, platform, and HW/FW/SW 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.

■ Reference for 6G Communication Systems: (please see page 31)

2. Radio System Solutions

■ Research Needs Label: [RSS]

2.1 Wireless RF

■ Motivation

New communication standards such as 5G beyond and WiFi7 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 WiFi7 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, 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

1) High efficiency sub-6GHz power amplifier simultaneously achieving the following targets:
   I. Switch-cap PA with >15% 3dB fractional bandwidth; center frequency between 2-6GHz using 1.8V supply.
   II. 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.
2) Multi-band mmWave power amplifier simultaneously achieving the following targets:
   I. > 30% fractional 3dB bandwidth, center frequency between 20-250GHz, using 1.2V or lower supply voltage.
   II. OP1dB >20dBm (if between 20-50GHz), >15dBm (if between 50GHz-100GHz), or > 10dBm (if >100GHz). Linear (small-signal) gain higher than 20dB.
   III. Antenna integration in module, in package, or on chip is optional but will be a significant plus.
3) Wide-bandwidth Nyquist rate ADC:
   I. Class 1: Sampling rate >3GS/s, over-sampling ratio between 2-4, dynamic range >57dB, interleaved paths <=2.
   II. 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.
   III. Clocking, and driving and reference buffers for the ADC need to be included.
   IV. 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
4) Direct IF bandpass receiver:
   I. Sampling rate >3GS/s; IF signal bandwidth > 400MHz; dynamic range > 57dB.
   II. 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.
5) VCOs (5-80GHz), DCOs (5-80GHz) and Crystal oscillators (<150MHz) exploring the following:
   I. Wide tuning range (continuous or banded operation), high-performance and low-power
   II. Phase noise suppression techniques for 10kHz~1MHz (preferably 5MHz) frequency offset away from the carrier frequency

III. 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

IV. Low power techniques to trade power with performance while satisfying key communication system requirements at respective operating modes of interest

6) Frequency synthesis

I. Power efficient frequency synthesizers with <40 fs integrated jitter

II. Focus on mmW frequency generation 28-150GHz and/or at 5-7GHz

7) Novel architectures for mmW / Sub-THz applications employing low resolution ADC / DAC

8) Process technology supporting ft/fmax higher than those from CMOS may be considered to design high mmWave and sub-mmWave (sub-THz) frequency bands for low power, high PA Pout and efficiency, and small form factors. Availability (commercialization viability) should be considered.

9) Ultra-broadband RF Phase shifter for Future RF beamformers:

Main part of this project is the design of a broadband quadrature signal generator and then using vector sum to generate the phase shift with the specific accuracy and resolution. In recent years, there has been a come back to lumped couplers for mmWave frequencies but more ideas/inventions are needed to make them compact and broadband. Spec[negotiable]:

I. Vector modulator based on phase shifter

II. Compact (less than 2 differential inductor area consumption or similar)

III. Ultra broadband covering n257:n262 [24.2GHz 48.2GHz]

IV. Gain variation< 0.5dB across the band

V. Phase accuracy >2 degrees

VI. Phase resolution 5 bit

VII. Loss < 3dB

10) Ultra-broadband, compact True Time Delays Future RF beamformers:

This project is the design of a compact time delay with the specific accuracy and resolution in the band of interest. There have been attempts to use switched transmission lines for such delays but there is a limitation such as huge area consumption (inversely proportional to the frequency) and lossy switching systems.

I. Spec[negotiable]:

- Compact: this will be used within a transceiver beamformer IC, hence it should be as compact as possible
- Ultra broadband covering n257:n262 [24.2GHz 48.2GHz]

- Gain variation< 0.5dB across the band
- Time accuracy > 2 degrees equivalent

II. Time resolution 5 bit

III. Loss < 4dB

■ 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.

■ Reference: (mmWave power amplifiers)

1) https://images.samsung.com/is/content/samsung/p5/global/business/networks/insights/white-paper/samsung-5g-fwa/white-paper_samsung-5g-fixed-wireless-access.pdf
2) https://www.techplayon.com/5g-nr-ue-power-classes/
3) S. N. Ali, et al. "A 25–35 GHz Neutralized Continuous Class-F CMOS Power Amplifier for 5G Mobile Communications Achieving 26% Modulation PAE at 1.5 Gb/s and 46.4% Peak PAE," IEEE Trans, Circuits Syst. I, Reg. Papers, vol. 66, no. 2, pp. 834-847, Feb. 2019.

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.

1) 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.

2) 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.

3) 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.

4) 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.

5) 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

1) Antenna topology study covering the following FR3 frequency ranges:
   - 5.9 to 8.4GHz
   - 12.7 to 13.25GHz
   - Dual band antenna covering both 5.9-8.4GHz and 12.7 to 13.25GHz
   - Dual band antenna covering S-band and C-band
2) The study of antenna miniaturization and strategic placement
   - This is crucial across different product platforms, including smartphones, tablets, and notebooks.
   - 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.
3) A high-performance antenna equipped with the following features to improve MIMO T-put performance
   - Antenna isolation: > 20dB
   - Antenna mismatch: < -15dB
   - ECC over FOV: < -15dB
   - Other features could also be studied and proposed from this research
4) Omi-directional antenna
   - 25% gain CDF and 75% gain CDF delta: < 2dB
5) Compact modular antennas with 4x or 8x antenna ports that could fit along the edge side of the phone
   - modular antenna with 4x antenna ports

- modular antenna with 8x antenna ports
- Preliminary size constraints for the 2x ports modular antenna: 3.8 x 20 x 2.5 mm^3
- Preliminary size constraints for the 4x ports modular antenna: 3.8 x 40 x 2.5 mm^3
- Frequency range of interest: 12.7 to 13.25GHz

6) A study and verification of FR antenna MIMO T-put performance
   - Development of a MIMO T-put simulation platform
   - Development of a MIMO T-put measurement and verification platform including in-house testing lab set up
   - Investigation of MIMO T-put capabilities within the constraints of a smartphone enclosure

3. Analog Circuits

■ Research Needs Label: [Analog]

■ Motivation

1) High performance, high bandwidth, power efficient analog circuit continue to play important role for wireless, wireline communications, automotive, smart home and AIoT applications. The key areas include power management, data converters and high speed Serdes. The focus includes innovations in architectures, circuits, and systems.
2) Power management: Explore integrated circuits and/or application circuits that could improve power conversion efficiency for application processors, RF power amplifiers, mobile devices, IoT and wearable applications
3) 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.
4) High speed interface (e.g., serdes): Techniques to support high data rate, power efficient data links and high density I/O system over advanced 2.5D/3D package are of interest.

■ Specific areas of interest

1) High speed interface Serdes: Power and area efficient circuits including but not limited to AGC, equalizers, high-bandwidth amplifiers, analog and ADC/DAC-based front ends, TX drivers and low jitter clocking, clock recovery, etc. with state-of-the-art performance (upon normalization over process technology if needed). Optical communication circuits and systems are also of interest.
2) 2.5/3D (INFO/CoWoS) interconnect with data rate 32+ Gb/s/wire. Power efficiency <=0.3 pJ/bit @ N4 process and could have normalization over e.g., process technology if needed.
3) Chip-to-chip single-ended communications on substrate, with data rate 16+ Gb/s/wire. Innovative architecture to achieve best power efficiency is highly interested.
4) High dynamic range, low power data converters and analog front end for audio and sensor applications. (>120dB, preferably >140dB)
5) High sampling rate, power efficient data converters for WiFi, 5G and base station

applications. (>=10bits, >2Gs/s/channel) [1]

6) Time-interleaved analog-to-digital converter calibration techniques for sampling rate >20Gs/s. (>=10bits)
7) IVR (Integrated Voltage Regulator) for SoC
   I. May include hybrid SC, LDO, etc. Vin=1.2V, Vout=0.3V ~ 1V, Iout > 2A [2]
8) XPU Power Delivery
   I. Multi-phase fast transient (>2A/0.1us) area efficient Buck converter with >90% efficiency @4-to-0.8V, Iout_max > 10A
   II. Multi-phase fast transient (>2A/0.1us) area efficient Buck converter with >90% efficiency @1.8-to-0.8V, Iout_max > 10A, and inductor <10nH
   III. ZCS/ZVS or resonant
9) RF PA Power Delivery / Modulator
   I. >100MHz (200MHz is preferred) ETM with efficiency >90% and low noise (e.g. spur noise -49dBm/MHz) [3]
10) Ultra-low voltage, low power analog circuits for bandgap, temperature sensor, oscillators and clocking with high stability, etc. (<=0.5V, nW)
11) Circuits and systems for analog AI, CIM, etc. that support AI computing acceleration and non-conventional computing.
12) Reliable and functional safety circuit design for automotive applications.
13) AI-powered design methodology for analog design productivity and performance boost.

■ Reference for Analog Circuit Research Needs: (please see page 32)

4. AI Systems and Hardware

■ Research Needs Label: [AIHW]

■ Motivation: Pioneering the Future of AI Technologies

In an era where AI is reshaping the landscape of technology and society, we stand at the forefront of innovation, seeking to push the boundaries of what AI can achieve. The rapid evolution of AI applications, fueled by breakthroughs in algorithms, the proliferation of big data, and exponential growth in computing power, has revolutionized user experiences and spawned a multitude of novel applications. Social media, multimedia, and gaming industries continue to drive technological advancements, while the emergence of generative AI, also known as AI-generated content (AIGC), has catapulted AI's utility, empowering individuals to enhance their productivity like never before.

Despite these advancements, the escalating complexity of System-on-Chip (SoC) designs outpaces the predictions of Moore's Law, confronting us with resource limitations, particularly in memory bandwidth and thermal budget. To navigate these challenges and continue our trajectory of innovation, we are calling upon the academic community to join forces with us in a collaborative research endeavor.

We are actively seeking research proposals that address a wide range of AI-related challenges, with a particular interest in the development and optimization of foundation models for systems and systems tailored for foundation models across diverse application fields such as mobile technology, automotive systems, AR/MR HMDs, surveillance, TV, and AIoT. Our call for collaboration extends to innovative data collection, generation, and benchmarking methodologies, algorithm-hardware co-design for edge devices, and the complexities of machine-learning cores and hardware architecture, including RISC-V design. We encourage cross-disciplinary proposals that demonstrate high innovation value and originality potential to create robust, efficient, and scalable AI platform solutions, aiming to advance the state-of-the-art in both foundational model development and system design to surmount the challenges inherent in AI systems and hardware.

■ Specific areas of interest

Application

1) Generative AI on edge devices: video and 3D, multimodality, LAM architecture, LLM

OS

2) Autonomous driving: end-to-end ADAS models
3) Low-bit efficient representation for Generative AI on edge devices: floating, integer, or special representation
4) Embedded vision and computational photography
5) Generative AI risk miscellaneous: adversarial attack, jail break, prompt injection

Machine-Learning Core

1) New foundation model/architecture: e.g. RWKV, Mamba
2) AI compiler optimization: e.g. TVM/MLIR, Vulkan ML related
3) Multi-core computing technology: e.g. runtime optimization
4) Android system optimization on RISC-V
5) Edge-cloud collaboration

AI Hardware

1) On-the-fly activation compression/decompression for edge devices
2) Weight-compression accelerator for edge devices
3) Ultra low power AI: e.g. CIM (Compute In Memory)

■ Reference for AI Systems and Hardware Research Needs: (please see page 30)

5. Multimedia

Research Needs Label: [MM]

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

1) Real-world image/video restoration and enhancement, with complexity and power consumption considerations
   I. Image/video restoration (denosing, super-resolution, …), video stabilization, video frame interpolation, … etc.

II. Real-world RAW images from cameras sensors
III. Real-world video streams from TV, streaming or social media
IV. Perceptual image/video quality assessment
V. Hardware-optimized AI/GenAI accelerators/functions/techniques design

2) Vision applications and scene/intention analysis
I. Joint training of visual perception systems (detection, segmentation, …), with temporal stability
II. Scene/intention analysis for surveillance and ADAS system
III. Domain adaptation approaches (unsupervised or semi-supervised domain adaptation is preferred)
IV. A simulator or a real platform for validating the proposed ADAS approach

3) Visual attention and transformers for low level image processing and visual recognition
I. Practical vision applications with visual attention or transformers
II. Feasible complexity for edge devices
III. Domain adaptation consideration
IV. Self-/semi-supervised learning is encouraged

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

5) Extended Reality (XR)
I. Simultaneous localization and mapping (SLAM)
II. Object/scene 3D reconstruction
III. Natural user interface

■ Reference for Multimedia Research Needs: (please see page 33)

6. Modern GPU

■ Research Needs Label: [GPU]

■ Motivation

1) Nowadays as ecosystem grows, plenty of new applications drive GPUs to new limits in different ways. MediaTek is seeking research proposals to optimize those state-of-the-art rendering techniques. As first step, bottleneck analysis is fundamental for developing better architecture, algorithm, and user experience. MediaTek urges for methods or cross comparison results of modem GPU architecture on new graphics/computing technologies.
2) As examples, MediaTek is interested in the following rising topics. Neural grphaics technology is become popular on mobile, and MediaTek looking for research and innovations on new GPU architecture and algorithms and optimization focus on performance, power, and picture quality.

■ Specific areas of interest

Modem GPU Architecture for Neural Graphics

Interest

1) Neural super sampling
2) Neural frame rate upsampling
3) Neural clothing/mesh deformation
4) Neural materials/Neural texture compression/neural displacement
5) Neural lighting/Neual ray denoising
6) Neural post-processing/Neural LOD
7) Neural Shading

Some Opportunities

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

Raytracing on Mobile

Interest

1) The raytraced-based importance sampling / guiding method for faster lighting converge or denoising friendly.
2) Data compression of raytracing data at different level, including AS layout, AS depth

reduction by geometry representation change.

3) Bandwidth reduction by smart caching policy or design.
4) Evolution of ray traversal and acceleration structure to handle game scene and animation smartly.
5) Seeking for better primitive compression & intersection test algorithm, such as Nvidia's DMM (displacement micromap) for converting geometry to triangle & high maps.

Some Opportunities

1) Graphics Algorithms
   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 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 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

High-Efficiency GPU-Driven Geometry Rendering on Mobile

Interest

1) Geometry culling to minimize overdraw and maximize HW geometry capacity utilization
2) Minimize replicate material read and compute
3) Use compute work graph to eliminate barrier sync, empty submits, and avoid worst

case allocation.

4) Compute rasterization to beat HW rasterization.
5) Support continuous LOD
6) Support skin vegetation
7) 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 GPU

Interest

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

Some Opportunities

1) 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.
2) 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.
3) 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.
4) 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.

7. 3D-IC Chip and Package

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

■ Research Needs Label: [3DIC] [Chiplet]

■ 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 technology is a technique that stacks multiple chips in three-dimensional space. Each chip can contain different functions, such as processors, memory, and sensors. Since the chips are very close to each other, communication speed and energy efficiency can be greatly improved. However, stacking multiple chips together also brings some challenges. Here are some challenges that may be encountered in 3D IC package:

1) Heat dissipation issues: When multiple chips are stacked together, the heat they generate will also accumulate. This may cause excessive heat buildup, resulting in system crashes or performance degradation. To address this issue, more efficient heat dissipation solutions (such cooling strategies) and novel thermal interface material (TIM) need to be developed.
2) Power supply issues: When multiple chips are stacked together, 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.
3) 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.
4) 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.

On the other hand, 2.5D chiplet technology combines individual chips together to achieve a complete system. Each chip can contain different functions, such as processors, memory, and sensors. The chips can be connected through high-speed interfaces, such as advanced substrate, silicon interposers, or through-silicon vias (TSVs), etc.. Here are some challenges that may be encountered in chiplet packaging:

1) 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 server and the demand of 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.
2) 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.
3) 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.
4) 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.
5) 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 reslove 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.

8. Verification and Validation

Research Needs Label: [V&V]

Motivation

As design complexity continues to rise via Moore's Law of transistor integration and now multi-die heterogeneous integration via package innovations, 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 methodology, 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 confirm consistency of pre-silicon predictions. Inconsistencies encountered are then used to drive improvements in pre-silicon processes for the next iteration. The constant pace of technology change forces continuous iterations of learning between pre-silicon verification and post-silicon validation.

Today, with the trend toward bespoke multi-core silicon optimized for specific application markets, it becomes imperative to adopt a system view comprised of the full hardware (HW) and software (SW) stack. Optimization of individual components without the system perspective is no longer sufficient. Functional safety is an example of a system-level requirement that have cross layer connections with those at the device level. Cloud hyper-scalers started reporting incidences of “silent data corruption” (SDC) in 2021 [1, 2] that can be traced to weak HW components which managed to escape device-level quality assurance. It triggered broad interest in industry and academia [3]. Device voltage/timing marginality is identified as one of the potential root causes [4]. As complex digitization extends into all manners of systems including those with safety-critical aspects such as automotive, the issue of SDE can become life-threatening. The traditional approach in fault tolerance and redundancy is cost-prohibitive for consumer-oriented 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 verification and validation of HW components. As full-system iterative learning is likely to increase greatly in complexity, advances in machine learning and generative AI holds promise to help manage productivity and boost effectiveness.

■ Areas of Research Need

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

2) Functional safety (FuSa) is a key requirement for systems that have life-threatening impact if failures occur. Such systems employ both HW and SW-based fault tolerance schemes to meet FuSa goals where cost considerations dictate the combination and balance of HW and SW. The traditional approach of HW triple-modular-redundancy is generally deemed too expensive for consumer-oriented markets. For meeting FuSa certification ASIL levels, coverage and detection of transient and permanent faults must be determined. Traditional gate-level fault simulation of the entire HW-SW stack is simply too costly and impractical. It's also too late to wait until gate-level HW implementation becomes available. More practical, efficient, and accurate ways to assess FuSa coverage are needed including fault modeling and simulation at a higher levels of abstraction, fast and accurate

fault coverage estimation techniques, and joint development with HW/SW fault tolerance schemes to achieve optimal results.

- **Reference for Verification and Validation Research Needs: (please see page 36)**

9. AI Productivity

9.1 Generative Artificial Intelligence (GAI)

■ Research Needs Label: [GAI]

■ Motivation

Generative Artificial Intelligence (GAI) has shown significant progress in recent years, due to the emergence of massive pre-trained models such as DALL-E, Midjourney, and ChatGPT, which can unlock potential for a wide range of applications. These pre-trained models have already shown promising results in producing fluent text, producing images, and performing few-shot learning. As leading company in IC design, we foresee GAI as future tool for accelerating electronic design workflow. These GAI services can involve in **hardware code writing, netlist graph generation, test cases generation for verification and more tasks in IC design flows**. Although these models offer huge potential, they have proven difficult to train, control and comprehend, giving rise to scalability, grounding, and interpretation challenges. Therefore, we are soliciting research proposals for addressing the followings areas:

■ Fundamentals

Large model training system and theory

■ Application

Generative AI in IC design

■ Specific areas of interest

1) Self-supervised Learning methodology (unimodal, multi-modal) (theory/application)
   I. Unimodal methodology
      - Text: Program synthesis. For example, GPT-like models for generating hardware code.
      - Vision: for example, Vision transformer pretraining for thermal simulation
      - Graph: generative models in large graphs, especially for IC circuit graph generation
   II. Multi-modal methodology:

- Text + Graph: Graph-Text Multi-Modal Pre-training [1]

2) Efficient large model training technologies
   I. 3D parallelism (data parallel, tensor parallel, pipeline parallel) [2, 3]
   II. Advanced ZeRO algorithms [13, 14]
   III. Compiler technology for deep learning [4]
   IV. Communication volume reduction for distributed training [5,6]
   V. Parameter Efficient method for pretraining [15, 17]
3) Data efficient learning methods
   I. Data Selection/ Data pruning / Curriculum learning methodology for generative model training [7, 16, 18]
4) Language models for data generation and augmentation [8]
   I. Goal: generating more and higher quality data [19]
5) Grounded language model
   I. Goal: making language model ground on facts, physical law and symbolic operations
6) Research in tool augmented neuro system [9], neuro-symbolic system [10]
   I. interplay between LLM and knowledge graph [11]
7) Reinforcement Learning for aligning human intent [12]
   - Goal: making models following human intent/ instructions
8) Challenges in making LLM agent
   I. Autonomous cooperation among communicative agents [20]
   II. Modalized agent system with the following modules
      - Profiling module: for example, making the agent play the role of a domain expert
      - Memory module: enable the agent to store information perceived from the environment and leverage the recorded memories to facilitate future actions.
      - Action module: enable the agent to interact with the environment.
      - Planning module: enable the agent to solve a complex task.
   III. Other related fields such as
      - Prompt robustness
      - Addressing hallucination
      - Known the knowledge boundary
      - Inference efficiency

■ Reference for Generative AI Research Needs: (please see page 37)

9.2 AI for IC Design

■ Research Needs Label: [EDA]

■ 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. MediaTek seeks to develop in-house capability to address them.

■ Potential areas of interest (but not limited to)

1) Eliminate redundant sign-off concerns
2) Multi-factor design closure techniques
3) Multi-objective constrained optimization with low optimization budget (model-based optimization and Bayesian optimization are welcome)
4) Multi-objective constrained combinatorial optimization
5) distributional searching, and critical indices approximation

Appendix: 6G Communication Systems

Reference

[1] 6G White Paper - MediaTek's vision for the next-generation of cellular mobile technologies: https://www.mediatek.com/blog/6g-whitepaper

[2] “Chapter 3 Radio Technologies” in Next G Alliance Report: 6G Technologies, https://nextgalliance.org/white_papers/6g-technologies/

[3] 3GPP SA1 TR 22.837, “Study on Integrated Sensing and Communication”

[4] IMT-2030 研究报告: 通信感知一体化技术报告（第二版）

[5] D. D. Wentzloff, A. Alghaihab and J. Im, "Ultra-Low Power Receivers for IoT Applications: A Review," 2020 IEEE Custom Integrated Circuits Conference (CICC), Boston, MA, USA, 2020, pp. 1-8.

Appendix: Analog Circuits

■ Reference

[1]

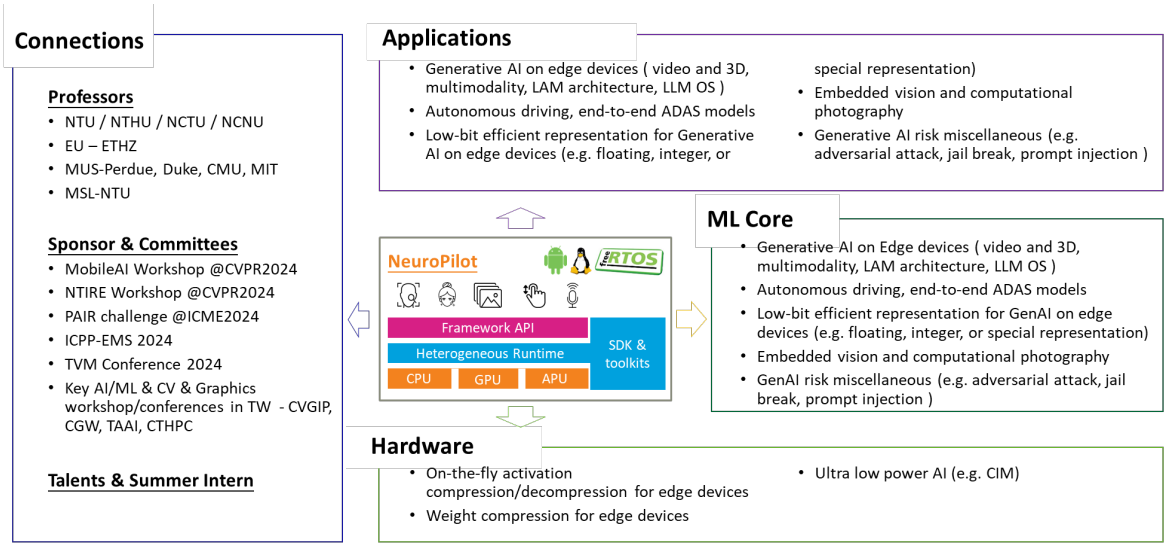
DAC		ADC		PLL	
Parameter	Specification	Parameter	Specification	Parameter	Specification
Resolution	14 bits	Resolution	12 bits	Ref. frequency	491.52MHz
Clock rate	16GHz	Clock rate	16GHz	o/p frequency	3.93 ~ 15.72GHz
o/p impedance	100ohm(diff.)	i/p impedance	100ohm(diff.)	R.M.S. jitter	100fs (10k~100M)
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: AI Systems and Hardware

■ NeuroPilot – MediaTek Edge AI Platform



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: Verification and Validation

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

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

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

Appendix: AI Productivity

Generative Artificial Intelligence (GAI)

■ Reference

[1] Park, S., Bae, S., Kim, J., Kim, T., & Choi, E. (2022, April). Graph-Text Multi-Modal Pre-training for Medical Representation Learning. In _Conference on Health, Inference, and Learning_ (pp. 261-281). PMLR.

[2] DeepSpeed: Extreme-scale model training for everyone - Microsoft Research(https://www.microsoft.com/en-us/research/blog/deepspeed-extreme-scale-model-training-for-everyone/)

[3] Li, S., Fang, J., Bian, Z., Liu, H., Liu, Y., Huang, H., ... & You, Y. (2021). Colossal-AI: A unified deep learning system for large-scale parallel training. arXiv preprint arXiv:2110.14883.

[4] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. _arXiv preprint arXiv:2205.14135_.

[5] Ryabinin, M., Dettmers, T., Diskin, M., & Borzunov, A. (2023). SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient. _arXiv preprint arXiv:2301.11913_.

[6] Gan, S., Lian, X., Wang, R., Chang, J., Liu, C., Shi, H., ... & Zhang, C. (2021). Bagua: scaling up distributed learning with system relaxations. _arXiv preprint arXiv:2107.01499_.

[7] Li, C., Yao, Z., Wu, X., Zhang, M., & He, Y. (2022). DeepSpeed Data Efficiency: Improving Deep Learning Model Quality and Training Efficiency via Efficient Data Sampling and Routing. _arXiv preprint arXiv:2212.03597_.

[8] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). Self-Instruct: Aligning Language Model with Self Generated Instructions. _arXiv preprint arXiv:2212.10560_.

[9] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., ... & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. _arXiv preprint arXiv:2302.04761_.

[10] Karpas, E., Abend, O., Belinkov, Y., Lenz, B., Lieber, O., Ratner, N., ... & Tenenholtz, M. (2022). MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. _arXiv preprint arXiv:2205.00445_.

[11] West, P., Bhagavatula, C., Hessel, J., Hwang, J. D., Jiang, L., Bras, R. L., ... & Choi, Y. (2021). Symbolic knowledge distillation: from general language models to commonsense models. _arXiv preprint arXiv:2110.07178_

[12] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. _arXiv preprint arXiv:1909.08593_.

[13] Chen, Q., Gu, D., Wang, G., Chen, X., Xiong, Y., Huang, T., ... & Sun, P. (2024). InternEvo: Efficient Long-sequence Large Language Model Training via Hybrid Parallelism and Redundant Sharding. *arXiv preprint arXiv:2401.09149*.

[14] Wu, C., Zhang, H., Ju, L., Huang, J., Xiao, Y., Huan, Z., ... & Zhou, J. (2023). Rethinking memory and communication cost for efficient large language model training. *arXiv preprint arXiv:2310.06003*.

[15] Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., & Tian, Y. (2024). GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. *arXiv preprint arXiv:2403.03507*.

[16] Chen, M., Roberts, N., Bhatia, K., Wang, J., Zhang, C., Sala, F., & Ré, C. (2024). Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, *36*

[17] Huh, M., Cheung, B., Bernstein, J., Isola, P., & Agrawal, P. (2024). Training Neural Networks from Scratch with Parallel Low-Rank Adapters. *arXiv preprint arXiv:2402.16828*..

[18] Qin, Z., Wang, K., Zheng, Z., Gu, J., Peng, X., Xu, Z., ... & You, Y. (2023). Infobatch: Lossless training speed up by unbiased dynamic data pruning. *arXiv preprint arXiv:2303.04947*.

[19] Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

[20] Li, G., Hammoud, H., Itani, H., Khizbullin, D., & Ghanem, B. (2024). Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, *36*.