Battery Technology Trailing Smartphone Innovation

February 2024

For Enovix

TIRIAS RESEARCH

Summary: Application Complexity and Usage Driving the Need for Improved Power Density

Hardware enhancements combined with increased application performance are increasing power consumption faster than enhancements in processing efficiency can offset them. GenAI is expected to push workloads over the top, making it difficult to even achieve 8 hours of battery life on a smartphone.

Constrained Form Factor with Plateau In Screen Size and Pixel Density

- Display & phone size has plateaued as high-demand use cases & component capability still increase
- Battery physical space under further pressure from demanding capabilities including cameras, folding

High Compute Use Cases Require Increasing Power Reducing Everyday Phone Battery Life

- Pressure to increase SoC performance, DRAM, Camera Capabilities

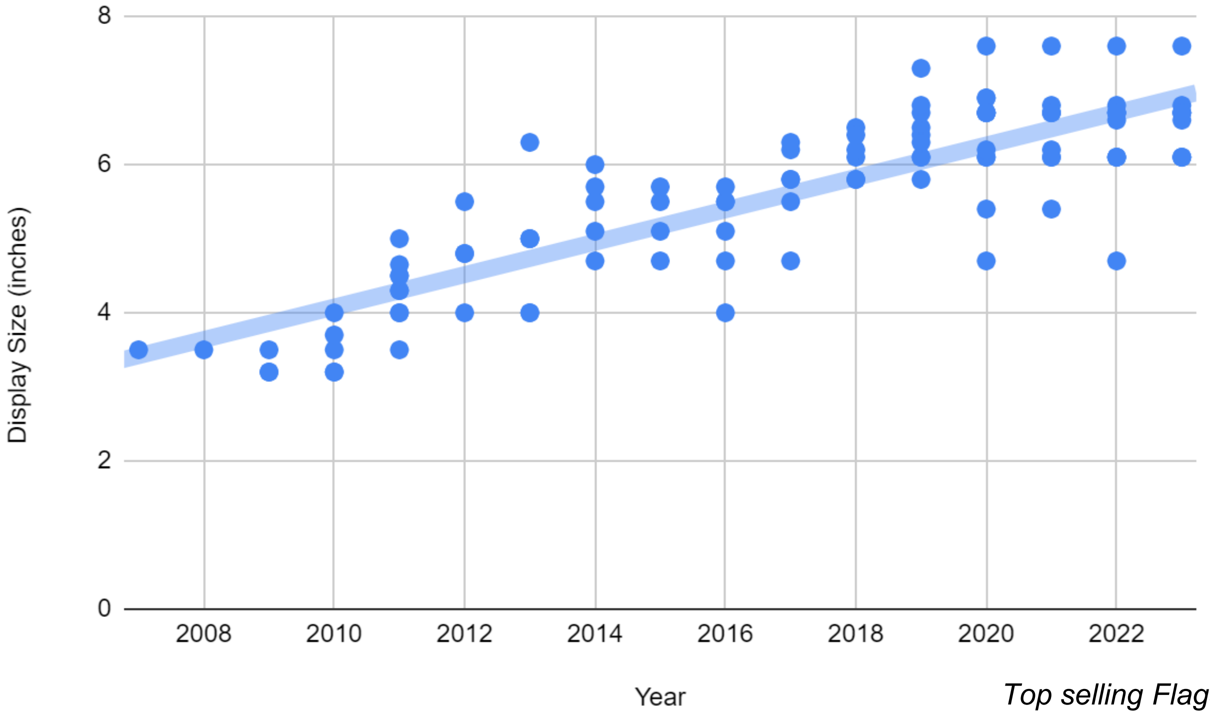
Generational Usage Shift & AI Driving the Most Demanding Use Cases

- Demanding tasks including copilot, content creation, filters, accessibility and safety
- Driving a permanent shift in global habits around streaming video, social media & video, and soon, AI

Smartphone Display & Dimension Growth Has Ended

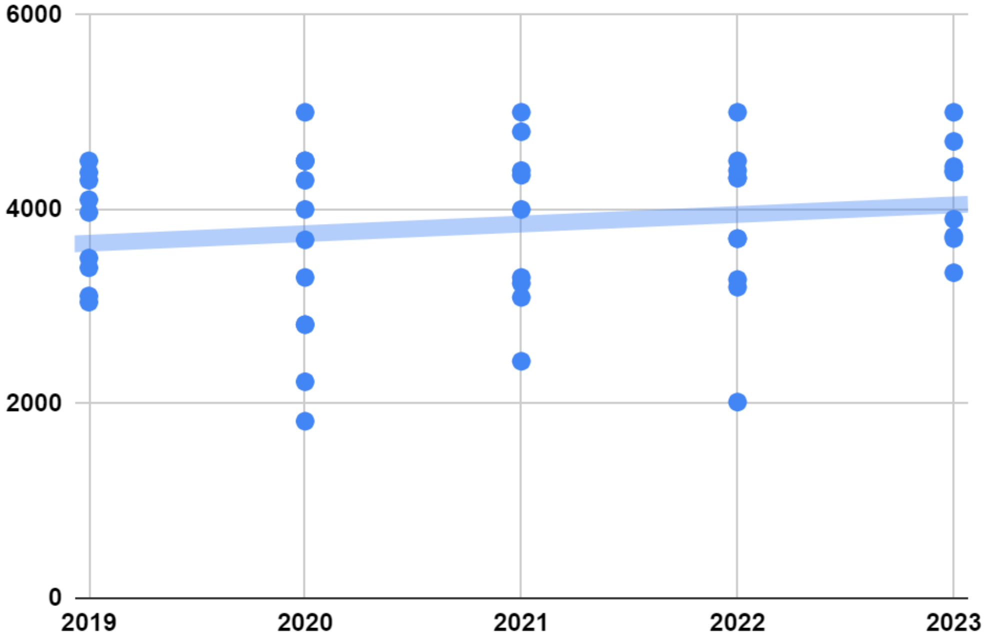
Smartphones Grew as Screen Size Increased 2007-2023

Linear Trendline



Plateau in 2019-2023 Screen Size

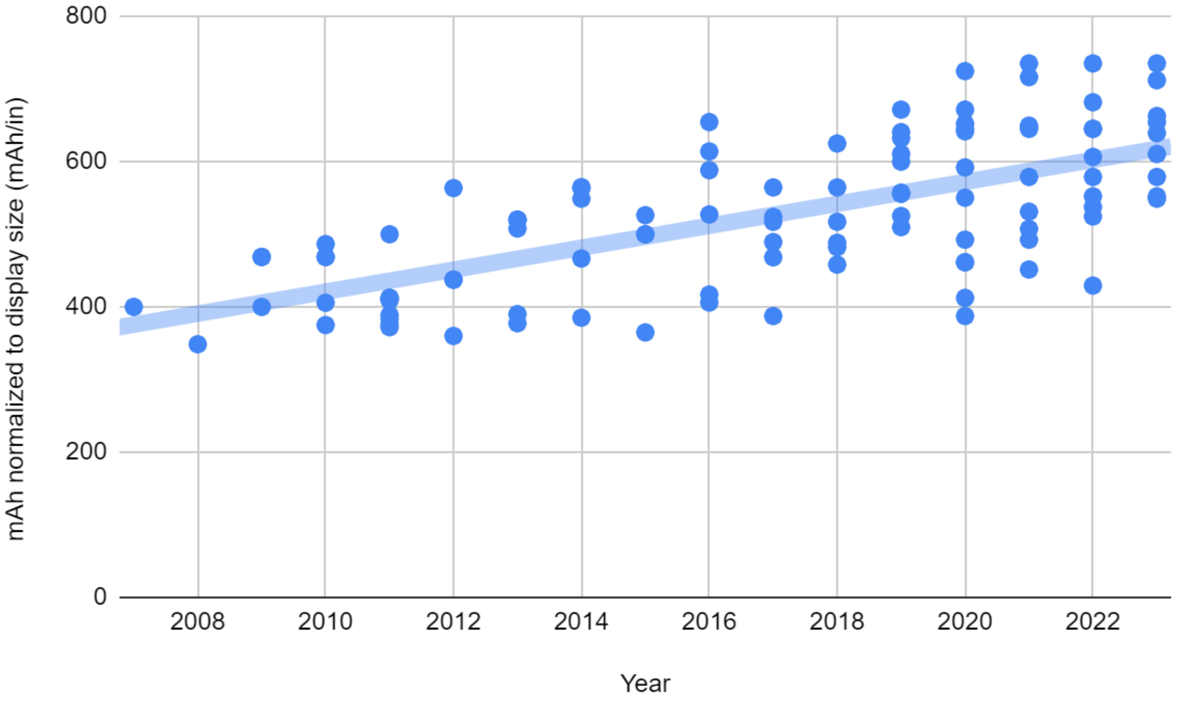
Linear Trendline



Top selling Flagship smartphones with linear trendlines for years shown. Source: Tirias Research

Battery Technology Not Keeping Up With Growth Demand

Battery Capacity Normalized to Display Size for Top Selling Flagship Smartphones with Linear Trendlines for Years Shown



Source: Tirias Research

Constrained form factors and low innovation slowing battery capacity growth

- Battery capacity growth primarily driven by device size which in turn is driven by display size
- Battery technology has remained constant over the smartphone era
- When normalized over display size, current battery technology has only delivered less than 5% growth

Mobile SoCs: Max SoC Clock Rate grew 14% in the Smartphone Era

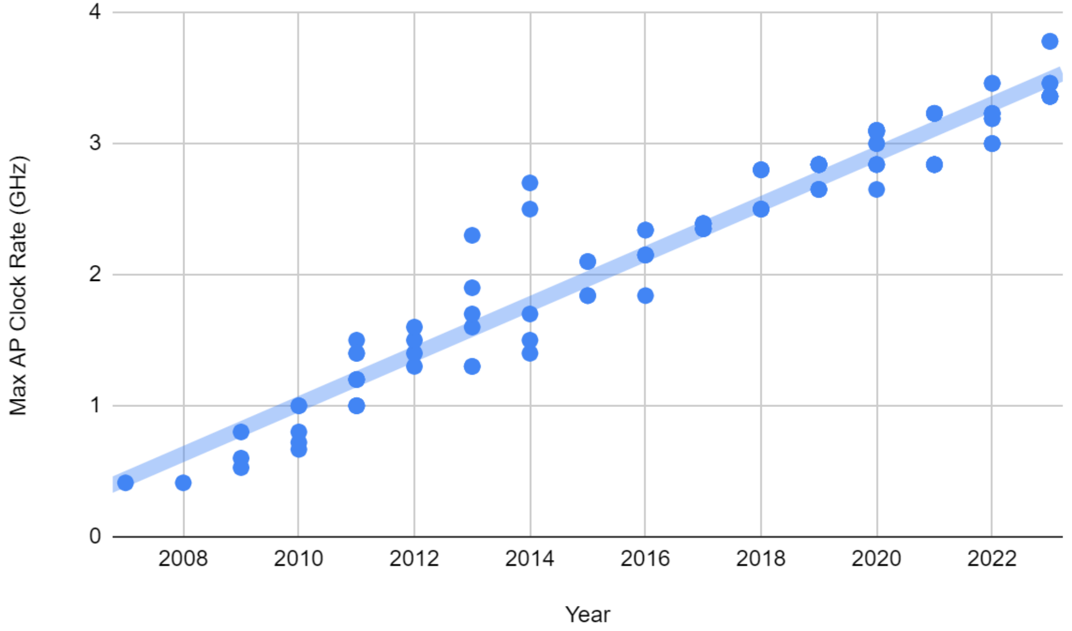
Demands for Increasing CPU Performance Continue to Push Core Count, Frequency, and Additional Hardware Accelerators

- SoCs have significantly increased core count from single-core to octa-core solutions both for the CPU as well as the GPU
- SoCs have added other heterogenous computing accelerators like NPUs, ISPs and sensor hubs

Moore's Law Unable to Keep Up with Compute Demands at Same Power or Capacity

- Increasing sustained power clearly creates opportunities for increasing use & SoC sophistication

Max Clock Rate for Top Selling Flagship Smartphones with Linear Trendlines for Years Shown



Top selling flagship smartphones with linear trendlines for years shown.
Source: Tirias Research

AI and GenAI on Smartphones Threatens All-Day Battery Life

As expected, there is a significant impact on battery life with the emergence of AI video enhancement, and GenAI image filters & chat, including:

- Video upscaling to 8K and enhancement
- Utilizing AI for video enhancement and image filters on local smartphones
- Adding AI to applications, even hybrid

As GenAI is optimized for mobile, users are likely to feel the pinch as they demand high performance on tasks that require maximal system resources

- Even the most efficient LLMs on the highest performance mobile platforms significantly underperform (50% token generation speed) the much larger cloud-based cousins (>50X model size)
- Its an unfair fight to provide users a great experience: multiple GPUs vs a single mobile SOC.

LLM Performance: Llama 2 7 Billion Parameters

	Response delay	Prefill	Decode
Apple (test 1)	1 sec	14.57 token/sec	9.0 tokens/sec
Apple (test 2)	3 sec	5.7 token/sec	7.8 tokens/sec
Apple (test 3)	1 sec	15.1 token/sec	6.9 tokens/sec
Apple (test 4)	1 sec	15.5 token/sec	8.2 tokens/sec
Samsung (test 1)	7 sec	5.4 token/sec	6.2 tokens/sec
Samsung (test 2)	8 sec	5.3 token/sec	5.8 tokens/sec
Samsung (test 3)	10 sec	5.4 token/sec	5.6 tokens/sec
Samsung (test 4)	10 sec	5.6 token/sec	4.7 tokens/sec

Llama 2 7B performance on flagship iPhone 15 Pro Max and Galaxy S23 Ultra is behind cloud-based modes in terms of token generation speed despite running significantly smaller models. ChatGPT 4 averages 13-15Tokens/Second throughput.

Source: Tirias Research

State of the Art Workloads

Analysis of Power Consumption for Emerging Applications on Today's Leading iOS and Android Smartphones

		Apple iPhone 15 Pro Max		Samsung Galaxy S23 Ultra	
		Battery Capacity - 4,441 mAh		Battery Capacity - 4855mAh	
Application	**Description**	**Capacity Used (Cap/hour)**	**% of Capacity Used**	**Capacity Used (Cap/hour)**	**% of Capacity Used**
		Without AI		**Without AI**	
4K Video / 30fps	**Video Capture**	666.2	6.7	825.4	5.9
Zoom (3)	**Video Conferencing**	266.5	16.7	534.1	9.1
YouTube (3)	**Video**	88.8	50	194.2	25
TikTok (3)	**Video**	133.2	33.3	291.3	16.7
Call of Duty (3)	**Mobile Gaming**	755	5.9	631.2	7.7
AdobeRush	**Photo & Video Editing**	133.2	33.3	291.3	16.7
		With AI		**With AI**	
8K Video / 60fps (2)	**Video Capture**	1154.7	3.8	1310.9	3.7
Llama 2 7B Version 1 (4)	**Chatbot**	932.6	4.8	776.8	6.3
ChatGPT3	**Chatbot**	310.9	14.3	339.9	14.3
Zoom w/AI Companion (3)	**Translation**	355.3	12.5	679.7	7.1
Otter AI (3)	**Transcription**	444.1	10	444.1	10
Stable Diffusion (5)	**Image Creation**	399.7	11.1	355.3	12.5
WOMBO (3)	**Image Creation**	355.3	12.5	266.5	16.7
Facetune (3)	**Photo & Video Editing**	177.6	25	88.8	50
Lensa	**Photo & Video Editing**	133.2	33.3	577.3	7.7

Source: TIRIAS Research

1 - Test Conditions - Only background tasks running, no external connectivity (when possible), display at 75%
2 - iPhone only capable of 4K
3 - Network connectivity required
4 - Llama 2 was the only LLM that was available for download through an APK.
5 - Stable Diffusion required the use of different apps between the iOS and Android platforms

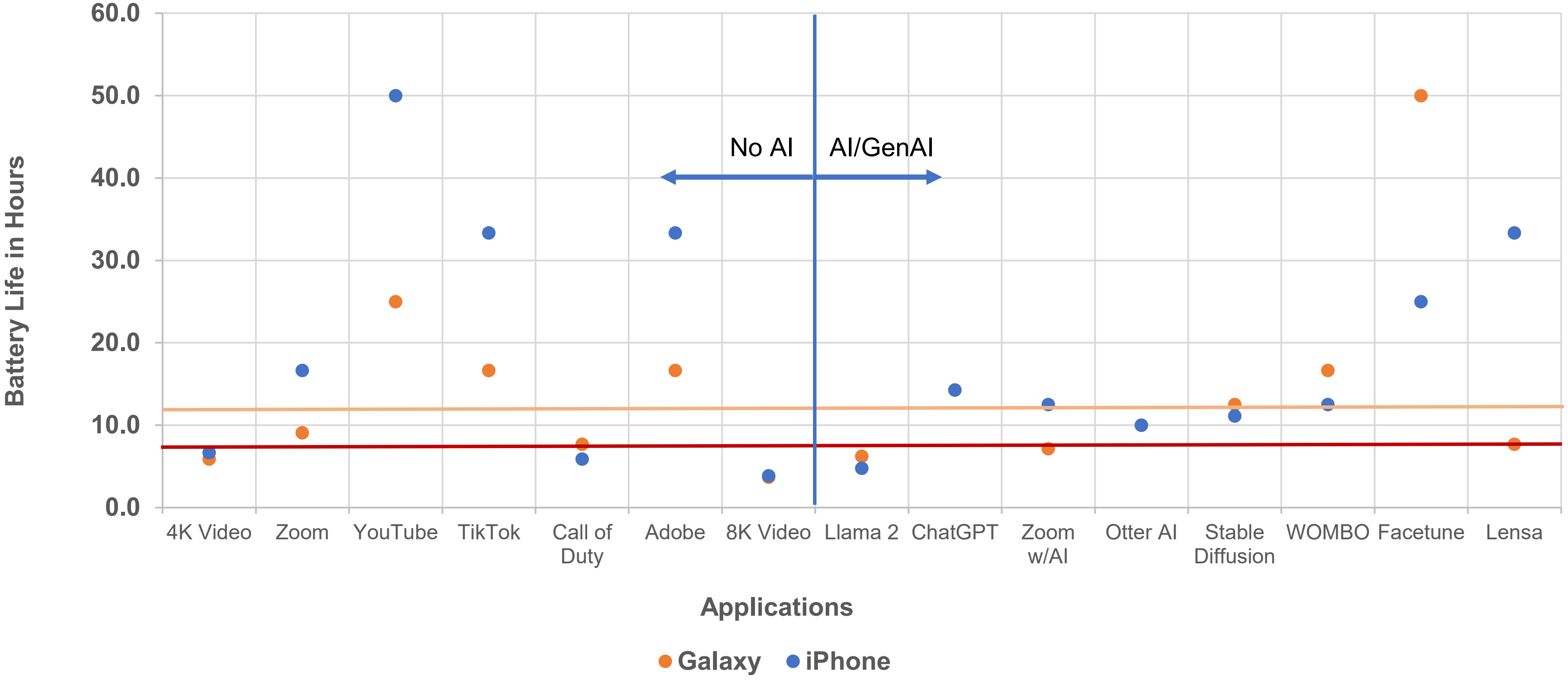
Emerging Workloads and AI Augmentation for Existing Workloads Will Drive Higher Power Consumption

- Most mobile AI applications are currently hybrid
- Adding AI to applications, even hybrid, has a negative impact on battery life
- Interactive video applications have the greatest impact on battery life
- Mobile devices have a difficult time supporting text-based GenAI today and will struggle to support video-based GenAI
- Media and gaming video at 4K and 8K with AI resolution enhancement kill battery life
- AI Test results vary depending on how the applications utilizes the SoC resources
- *Note: Cloud based applications like Facetune & Lensa show offload primary compute workloads to the cloud.*

As GenAI Applications are Optimized, Power Consumption will Increase and User Experience Will Improve

We are now able to benchmark AI and with preliminary data on GenAI application prior to full optimization to evaluate workloads on Smartphones relative to traditional video, gaming, and productivity usage. Among demanding applications, hardware video decode, and cloud offload create the outliers.*

Smartphone Battery Tests by Application



12-hour is more realistic for multi-app use

8-hour battery life is the baseline

Source: TIRIAS Research

*Current GenAI models built for smartphones are not yet optimized and lack support for onboard accelerators; as optimization improves, we expect power consumption to increase, however application performance should improve significantly.

TIRIAS RESEARCH

Today's Demanding Application Categories are the Strongest Candidates to Incorporate AI

Video applications run well when hardware video encoding is in place, however, when it is not, the impact to battery life is significant

- 8K video utilizing AI among the most demanding use cases stressing memory, processing, and storage
- VVC 4K video being adopted with scant hardware decoding support

Gaming remains among the most demanding use cases; the use of video enhancement or GenAI will push power consumption beyond feasibility without innovation and hybrid operation

- Demand to utilize the high-resolution, high-framerate smartphone screen to its fullest potential
- Opportunity to employ DLSS-like resolution upscaling enhancement vis AI

Interactive video also remains among the most demanding use cases, and similarly, use of GenAI to augment on-device will similarly push power consumption beyond feasibility

- Utilizing GenAI for collaboration seems natural, however, innovation is required to improve the feasibility
- Even today's in-reach tasks such as AI-based visual enhancement and backgrounds significantly impact battery life

GenAI Demand is Forecast to Increase over 150X 2023 to 2028

Demand for GenAI services is forecast to expand rapidly, with few obstacles to adoption and significant R&D investments in new service development and core GenAI technologies creating “double exponentials” in innovation

- By the end of 2023, user request for GenAI **tokens**, analogous to words or symbols, is forecast to exceed 6 trillion by with over 450 million monthly active users (overlapping) across services, plus over 15 billion GenAI **images & video frames**
- By the end of 2028, the forecast estimates over *one quadrillion* **tokens** and over *2.5 trillion* **images & video frames**
- Applications span companion, search, and automated/API with paid, free, and B2C monetization

Global GenAI Output (Billions)	2023	2024	2024 vs. 2023	2028	2028 vs. 2023
Images + Video Frames	15	59	4X	2,500	167X
Tokens	6,900	19,900	3X	1,034,000	151X
Web Searches (Reference)	2,000				
TikTok Videos Viewed (Reference)	365				
Digital Photos Taken (Reference)	1,800				

Source: Tirias Research GenAI FTCO Forecast

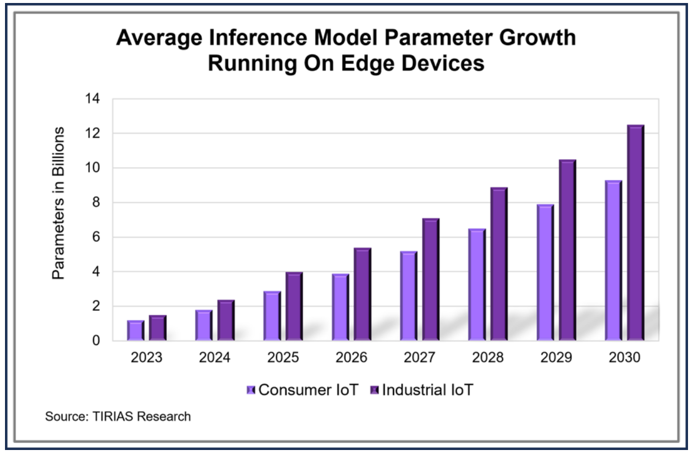
Hybrid Execution and Specialized, Optimized Models are the Best Path to GenAI on Smartphones

Edge devices, through advancements in AI performance, are expected to run local models in the 3 billion to 12 billion parameter range, including models optimized for specific tasks, reducing memory and processing requirements

- Usage Steering for Specialized Neural Networks: Edge devices can direct users towards specialized neural networks that are more efficient for specific tasks, using models with fewer parameters, reducing the burden on both cloud and client devices
- Model Size and Computational Efficiency: Optimization techniques aim to balance model size and computational efficiency; By reducing the number of parameters, the models become more feasible for execution on edge devices without significantly compromising accuracy
- Optimization Techniques: Techniques like quantization, pruning, knowledge distillation, and model specialization help in reducing the size and complexity of neural networks
- Advancements in Smartphone Capabilities: Modern smartphones, with their rapidly advancing processing, memory, and sensor technology, are increasingly capable of handling sophisticated AI tasks, including computational photography and AI-driven video processing
- Hybrid Computing Models: For tasks that are too demanding for edge devices alone, a hybrid computing model can be employed, processing initial layers or simpler parts of a task on the device, while more complex processing is offloaded to the cloud

Benefits of moving from the cloud to the edge include reducing service operating costs, lowering response delay time, and increasing privacy and security

- Reduction in Data Center Load and Costs: By offloading GenAI processing to edge devices, there is a significant reduction in data center infrastructure and operating costs, lowering power requirements and environmental impact.
- Increased Privacy and Security: Processing GenAI tasks on-device enhances data privacy and security, as it reduces the need for data transmission to and from the cloud



Source: Tirias Research GenAI FTCO Forecast

TCO, Environmental Impact Driving Motivation to Move GenAI to the Edge

GenAI 5 Year Forecast Highlighting Total Infrastructure Cost & Power, Showing Potential Savings by Moving Workloads to Edge Devices

Global GenAI TCO Forecast	2024	2025	2026	2027	2028
Forecast Total Data Center GenAI KWh - Power required to run GenAI servers at expected utilization with data center cooling & overhead	1,375,000,000 KWh	5,016,000,000 KWh	11,058,000,000 KWh	31,794,000,000 KWh	66,978,000,000 KWh
Forecast Total Operating Costs - Includes Amortized Servers, Power, Operations	$ 1,725,000,000	$ 6,285,000,000	$ 13,837,000,000	$ 39,859,000,000	$ 84,029,000,000
Potential $ Savings of Moving 20% of GenAI to distributed edge devices	$ 345,000,000	$ 1,257,000,000	$ 2,767,000,000	$ 7,972,000,000	$ 16,806,000,000
For Reference, the Number of Smartphones That Would Use the Same Power @ 3.5 KWh/year - Benchmarked Daily Charge Power Samsung Galaxy S23	392,859,000	1,433,145,000	3,159,350,000	9,084,080,000	19,136,815,000

Source: Tirias Research GenAI FTCO Forecast

GenAI inference will scale creating massive incentives to distribute workloads to edge devices

- Moving just 20% of the GenAI workload to the edge with smaller models and hybrid processing would save $16B in cloud TCO in 2028
- By 2028, Cloud GenAI power consumption is forecast to rise to over 66 Billion KWh, driving carbon impact concerns
- For perspective, Cloud based GenAI by 2028 anticipated to consume the same power annually as 19 Billion flagship smartphones

Usage and Resolution Growth Continue to Drive Growth in the Total Video Streams to Edge Devices

Social media video, video meetings, and streaming video on demand are all contributing to smartphone video usage; looking at leading streaming video providers we can see the scale of growth and video delivery

Streaming Social Video Global Forecast **Millions of Video Hours Viewed in Terms of 1080P 3Mbits Video Streams**	2024	2025	2026	2027	2028	CAGR
TikTok (Global)	43,760	54,845	61,613	68,525	78,303	12.2%
Instagram (Global)	346,211	383,910	400,975	414,609	411,993	3.5%
YouTube (Global)	103,065	120,078	127,244	133,490	129,960	4.6%
Total US Streaming on Demand (SVOD) Includes Netflix, Prime, Disney+, Paramount, Apple, and Rest of Market	143,892	153,661	159,395	163,690	166,602	3.0%

Source: Tirias Research Streaming Video FTCO Forecast

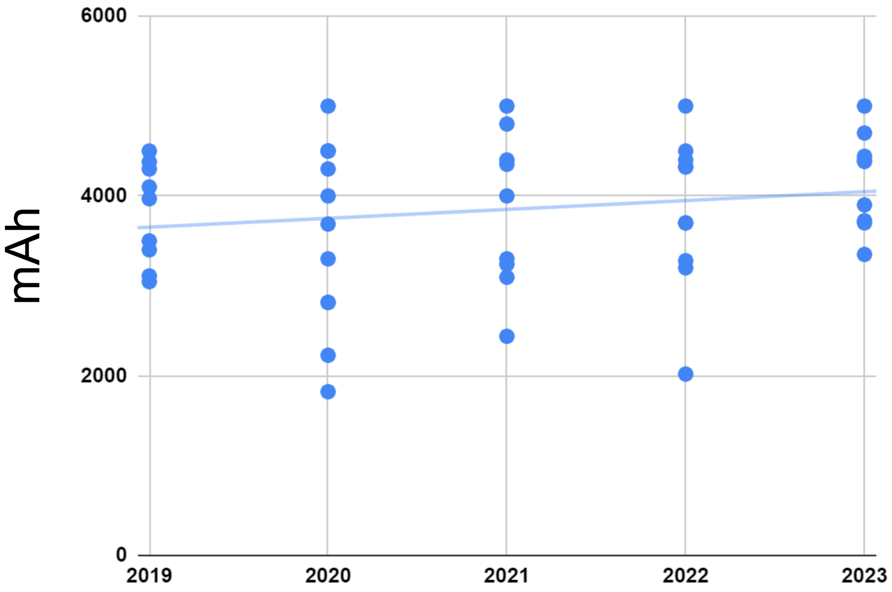
Streaming video is becoming a mainstay of social media, driving billions of video hours to Smartphones globally

- Social video is streamed continuously with preload for the next up video ensuring devices are highly utilized during everyday usage
- Usage is increasing at the same time resolution is increasing, leading to increased hours of usage and higher equivalent 1080P/3Mbits streams
- Video for social and meetings is largely smartphone encoded, creating heavy smartphone video encoding workloads over extended periods

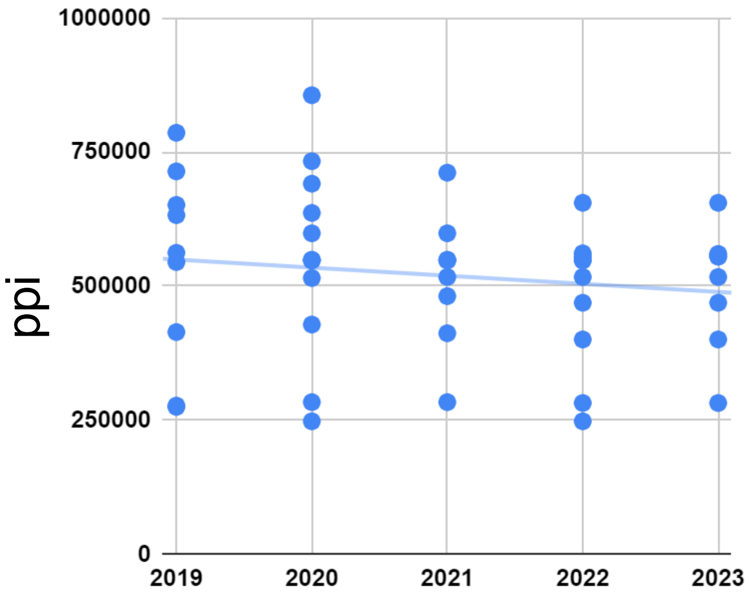
Innovation In Battery Power Required To Meet The Demands of Gaming, Video, AI, and Increasing Cloud Connectivity

We are reaching the end of bigger smartphone batteries; Augmenting current workloads with AI and the growth of GenAI is a tipping point driving the need for battery capacity innovation

Battery Capacity for Top Selling Flagship Smartphones with Linear Trendlines for Years Shown



Display Density for Top Selling Flagship Smartphones with Linear Trendlines for Years Shown



Source: Tirias Research

Battery capacity has only increased minimally due to slower battery technology innovation and internal volume competition from increasingly sophisticated components once screen sizes plateaued