Counterpoint Technology Market Research

Western Digital.

WHITEPAPER

How 5G, AI and IoT are Shaping Future of Storage

Executive Summary

The next phase of technological development will be characterized by the combination of three things: 5G, Internet of Things (IoT) and Artificial Intelligence (AI). Use cases driven by this intelligence-centric connectivity will catalyze computing at the edge and endpoints as they effectively become mini data centers and bring a completely new paradigm to storage at the edge.

Key considerations include the following:

- It will involve AI model storage and handling streaming data from sensors at the edge.
- Lower power consumption at battery-powered endpoints and greater capacity requirements will be a factor.
- Endurance and reliability of storage in harsher environments becomes critical.
- The storage capacity of the network edge must be scalable as more endpoints connect.
- Managing streaming data effectively with adequate buffering in a low-latency, intermittently connected environment will be challenging.
- Avoiding data duplications and conflicts will be important to enable quick and efficient decision-making.

5G and AI bring complexity as orders of greater magnitude of data are generated and handled at the edge than in today's traditional implementations. Planning storage early in the 5G IoT application design phase has become very important to adhere to the above requirements.

The key IoT applications to be driven by 5G and AI are **industrial IoT** endpoints, **edge gateways**, **robotic** automation across different verticals, and **mobility** applications from **vehicles** to **drones**. All these edge IoT applications have varying storage needs, which include performance, read/write speeds, data retention, capacity, temperature and endurance.

This report focuses on how these **key endpoint and edge applications** will evolve in the 5G era and, in particular, their impact on storage trends as designers and manufacturers move from today's IoT to a **5G-driven** intelligent IoT paradigm.

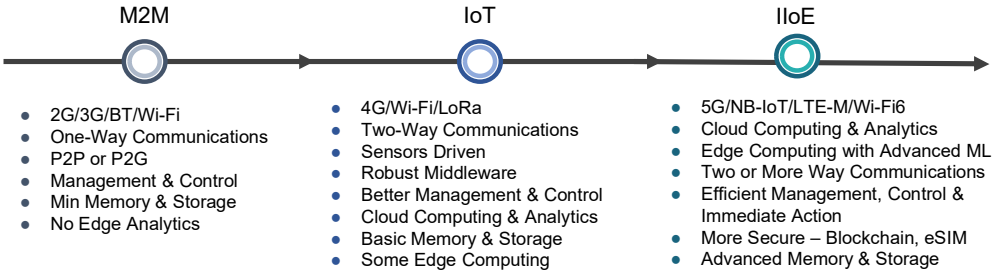
5G & AI Shaping IoT Endpoint and Edge Applications

The combination of wireless transmission speeds and computing power is enabling higher levels of intelligence to reside at the network edge, resulting in proliferation of applications not possible before.

5G Delivers Speed, Scale and Stability

The past couple of decades show how cellular technology has shaped IoT. 2G and 3G during the 2000s brought **Machine-to-Machine (M2M)** which, although revolutionary at the time, was comparatively simple in terms of what it enabled. 4G and LPWAN facilitated far more sophisticated applications in the **Internet of Things (IoT)**, with robust cloud connectivity allowing efficient data transfer and better device management, control and action.

Exhibit 1: Evolution from M2M to Intelligent Internet of Everything

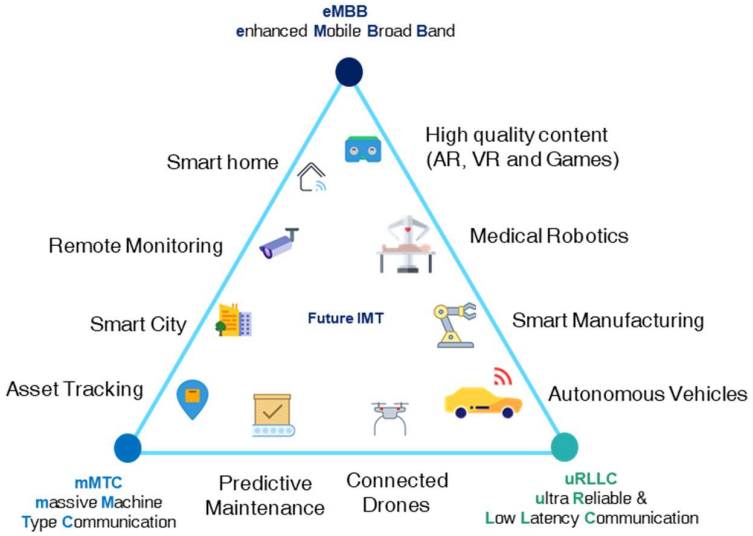


Source: Counterpoint Research

In the current phase of IoT evolution, 5G deployments are layering with AI and edge computing to enable Intelligent IoE – the **Intelligent Internet of Everything** – where ultimately anything powered can compute, and anything that computes can be connected.

Historically, these 'connections' were constrained by the network, as bandwidth, latency and capacity issues narrowed the breadth of Intelligent IoE applications possible. 5G's enhanced capabilities, highlighted by three key service categories (Exhibit 2), remove many of these constraints, bringing forth a level of AI integration that allows autonomous computing at the edge.

Exhibit 2: Key 5G Capabilities Enabling Intelligent IoT Applications



Source: Counterpoint Research

These new primary 5G New Radio (NR) use cases defined by the 3GPP bring order-of-magnitude improvements across network transmission speeds, capacity and latency:

- **Enhanced Mobile Broadband (eMBB):** Peak data rates will be in the tens of Gbps. Importantly, there are also three distinct attributes to be delivered by eMBB: 1) Higher capacity – availability in densely populated, indoor/outdoor areas; 2) Enhanced connectivity – availability everywhere; and 3) Higher user mobility – availability in moving vehicles from cars to planes. Typical IoT use cases include devices that require higher capacity and lower latency for video and data streaming, as well as industrial applications for AR/VR-based digital twins.

- **Massive Machine Type Communications (mMTC):** It enables massive network capacity to connect thousands of IoT endpoints and edge devices reliably without congestion issues. Typical endpoints would be low-cost, battery-powered devices periodically transmitting small amounts of stored data either to the core or other devices locally via mMTC IoT gateways.

- **Ultra-Reliable and Low-Latency Communications (URLLC):** It promises low latency and high reliability for mission-critical applications like autonomous driving, wireless control of industrial automation, and robotic surgery.

Early examples around these use cases highlight benefits for consumers and enterprises alike.

On the road, vehicles will use eMBB to access multimedia content for passengers, while autonomous driving will be made safer through URLLC by lowering reaction times to 10ms (vs 4G's 30-100ms). In the air, drones can leverage video and image capture via eMBB for remote monitoring, surveying, navigation and flight safety.

Industrial automation will benefit from real-time remote control, high-resolution video inspection, monitoring and highly autonomous operations. This would include applications around supply and inventory management, robotics, and operational control and positioning of general equipment and other items, many of which would have strict reliability and latency requirements (mMTC) and high data rate needs (eMBB).

Healthcare will be revolutionized, especially for patients in rural hospitals, with the advent of medical robotics and telesurgery. Enabled by 5G, orchestrated remotely by a doctor and performed locally by robots, telesurgery would allow a surgeon to perform an operation on a 'virtual copy' of the patient with a 'digital twin'. The life-or-death necessity for real-time feedback via 3D video and sensory data transmission will require networks to support both eMBB and URLLC.

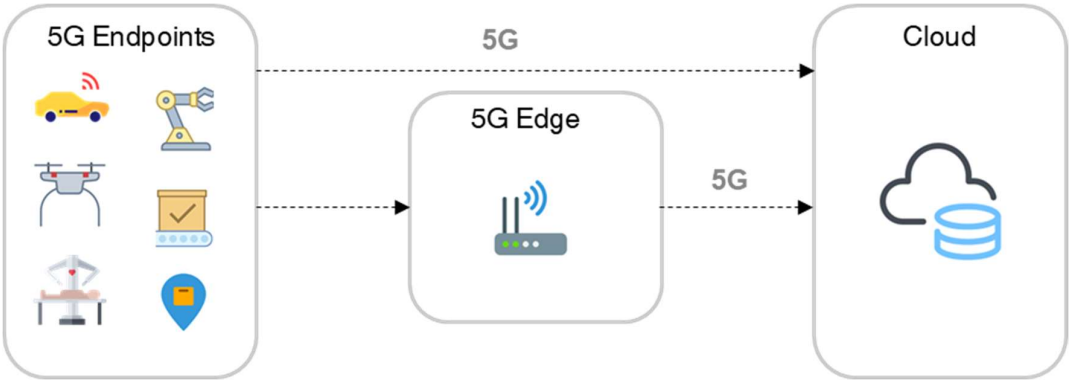
Other robotics growth areas include agriculture, where robots can help monitor crop condition and perform a broad range of actions from crop maintenance to harvesting and sorting. With robot-to-base station transmission times limited to 1ms (based on current ethernet or wired technology), 5G will be critical in the development of truly mobile, wireless and intelligent IoT robotics applications. The coming 5G-driven intelligent IoE era will enable a broad range of intelligent IoT applications, facilitating social and business digital transformation.

> Faster speed, ultra-lower latency and massive scale enabled by 5G network will drive the demand for advanced IoT edge and endpoints which are intelligent, mission critical and highly automated. These 5G-capable edge devices will drive advanced compute, memory and storage requirements for variety of industrial, content and mobility applications.

AI Brings Cloud Capabilities Closer

IoT devices are increasingly generating large amounts of data, much of which ultimately needs to go to the cloud. Although 5G solves issues around the physics of transmission, it is less helpful with the economics. AI helps in this respect, processing massive amounts of raw data at the edge to reduce data size. This can be easily seen in vision-driven applications such as autonomous vehicles and drones.

Exhibit 3: AI Enabling Intelligent IoT Endpoints and Edge



Source: Counterpoint Research

In many implementations, endpoints lack 5G due to cost and power constraints. Instead, they connect to intelligent 5G edge devices such as IoT gateways and CPEs enabled with mMTC. These devices collect endpoint data, filter and process it, and provide secure upstream and downstream channels. The devices can be potentially used in multiple scenarios, including for low-latency Untethered XR devices, industrial machines and others. Effectively, 5G IoT gateways will be edge servers loaded with AI capabilities to facilitate real-time devices, data management, control, analytics and action.

AI-empowered drones and robots are good examples of these types of endpoints. They show potential in various industrial settings, including routine inspections across extensive and dangerous environments. For example, direct human pipeline and dam inspections are known to be costly and frequently associated with life-threatening situations. AI-enabled drones and robots not only have the capability to replace human inspectors, but can also improve the process by plugging into a platform which identifies possible anomalies and shares this analysis across the entire system, thus improving it iteratively.

Opportunity for Storage Players

Endpoints and edge devices before 5G and AI required small storage footprints because their operating system, compute and analytical requirements were minimal. However, with 5G and AI advancing capabilities, storage needs will rise exponentially with the increase in demand for high-density code storage as well as for streaming and buffered data. Further, more reliable and higher-speed storage solutions will become critical in enabling optimal performance.

Code Storage – NOR to NAND Flash Transition

The two types of storage generally used in IoT applications are NOR and NAND flash memory. Each has its strengths and weaknesses, and selection criteria vary depending on usage.

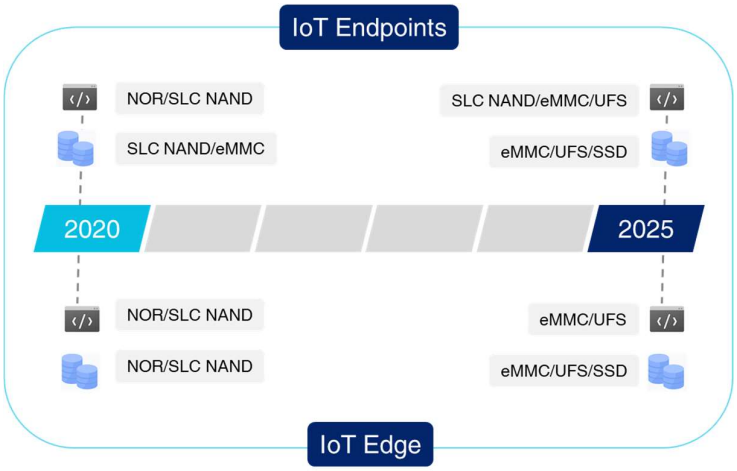
NOR is stable and reliable with high endurance, low power consumption and execute-in-place (XIP) capability that allows it to boot up and execute programs faster. The key downsides are unit cost and low density which limit its ability to store new OSes that are typically larger.

NAND does not have XIP capabilities but is faster in every other way, takes up less physical space and is able to achieve larger density. For 5G + AI/ML in edge applications, NAND offers better controllers to store ML code/algorithm for frequent updates.

Historically, the rule of thumb for choosing the flash memory type has been as follows:

- **NOR**: Normally used as code storage, which requires reliable performance, high retention and frequent reading.
- **NAND**: Often used where larger capacity and higher performance are required.

Exhibit 4: Storage Type for IoT Endpoints and Edge



Source: Counterpoint Research

Recently, a shift has been seen to NAND flash for code storage as well, especially across embedded applications requiring code capacity of 512MB and above. Beyond the cost advantages, which can be significant with NOR's unit cost double that of NAND, its smaller physical footprint and improvements in execute performance have helped it become more popular for code storage.

> Due to the advent of intelligent edge, the need for a memory footprint for both code and data storage is increasing rapidly and driving a paradigm design shift for future endpoints from NOR Flash to NAND Flash-based solutions across the board from industrial gateways, machines to intelligent manufacturing systems.

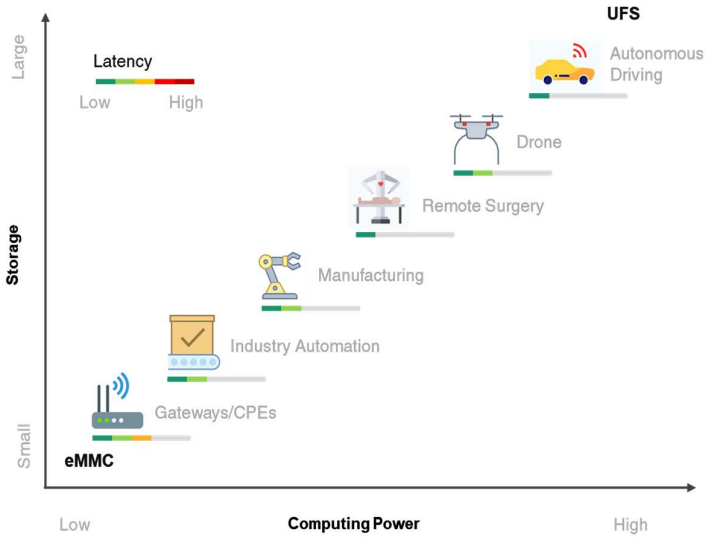
As IoT applications become more sophisticated, the trend towards more NAND usage will continue. Typical IoT devices run simple RTOS with low hardware specs, computing power and memory requirements. Next-generation intelligent IoT devices, such as those used in automotive, medical and industrial robotics, will have larger and more complex OSes (Linux/Windows), edge software agents, AI engines and memory for code storage – all of which will drive migration to more reliable, high-performance NAND flash.

Automotive and industrial NAND-based code storage solutions now exist commercially, and over the longer term, more NAND specifications (eMMC, UFS, and SSD) are expected to come to the fore.

Data Storage – Capability and Capacity Evolution

Today, the most common IoT data storage solution is eMMC. However, with the proliferation of 5G and AI, IoT applications will need to access and transmit data faster, all while facing ever-increasing demand for storage capacity.

Exhibit 5: Storage Parameters by Key IoT Endpoints and Edge



Source: Counterpoint Research

Key 5G IoT applications include the following:

- **Autonomous driving systems** require high performance and the capacity to efficiently manage large amounts of high-speed data from sensors, infotainment systems, operating systems and HD maps.
- Data storage solutions will be critical enablers in the evolution of autonomous vehicles, the next big leap for the automotive industry.
    - Sensor data will account for the largest share of onboard data storage, with most data coming from advanced driver assistance systems (ADAS) and vehicle-to-everything communication (V2X).
    - Additional space will need to be reserved for media, gaming, voice AI applications and other offline content. Also, 'black boxes' are expected to be made a legal and safety requirement by regulators.
    - Data generated from safety services, OTA updates and similar services will require some local storage space.
    - Unlike current 2D navigation maps, high-definition (HD) maps vary greatly in refresh rates, sampling methods and cost. Due to the high frequency of dynamic data updates, HD maps typically use real-time online updates via wireless transmission. The HD map includes a static layer, semi-static layer, semi-dynamic layer and a dynamic layer, with the base static layer updated monthly or as needed.
    - However, dynamic maps can be updated every hour, minute and second. Therefore, as the level of autonomy increases from Level 1 to Levels 4 and 5, vehicles will need an exponential increase in storage capacity and bandwidth.

    Non-volatile memory devices play a key role in autonomous systems, providing fast boot code storage and data storage for recording critical events, and storing AI models and HD maps. Further, the system will need to process more data at faster rates and with greater reliability as it becomes smarter.

Although NOR flash storage is an ideal storage technology for low capacity code storage, it is expensive and lacks the capacity to satisfy more complex systems. With improvement in reliability, eMMC, UFS and SSDs are set to become essential standards as favorable economics and generous capacity help drive the take-off. Counterpoint expects the requirement for in-vehicle storage to range from 1.2TB in Level 3-Level 4 passenger vehicles to 7TB in Level 4 robo-taxis by 2025.

- **IoT gateways** are becoming important edge devices for collecting, filtering and processing streams of data from multiple sensors based at IoT endpoints.
    - Edge gateways compute characteristics ranging from ARM Cortex-R or A to Intel-based processors and thus could act as mini data centers.
    - Further, Counterpoint estimates the 5G CPE/Gateway market for households as well as enterprise to grow significantly over the next decade to offer high-bandwidth and low-latency services to the consumers and enterprises.
    - As a result, storage requirements could range from a few GBs to several TBs depending on the deployment and application. Most household CPE designs will integrate eMMC solutions and UFS will be a choice for some higher bandwidth applications such as surveillance cameras.
    - However, the intelligent edge gateways for the enterprise or industrial applications will be designed with high density SSD solutions capable of storing millions of streaming datapoints from hundreds or thousands of sensor-based assets, along with meeting storage requirements for a higher-level OS, containerized software, AI/ML capabilities and the IoT platform.

- **Drones** have the potential to generate massive amounts of image and video data – from remote monitoring and surveillance to goods delivery.
    - MicroSD remains the major memory interface in today's drones. However, to provide higher capacity and higher-speed storage, some drones are implementing eMMC, UFS, and even SSD based on different drone applications.
    - With flight times of general commercial drones around 30 minutes, Counterpoint estimates a typical flight could generate at least 150GB of data, with high-resolution mapping, 3D modelling and other similar applications producing more.
    - Henceforth, for most of the near-to-mid-term 5G-centric connected drone designs, eMMC will continue to be the first choice for OEMs.
    - However, as drones move towards autonomy and longer flight-hauls in distant future, they may face intermittent connectivity and will thus require higher storage for the ML models, advanced HD maps navigation, higher-resolution (4K video, images) captured data and other system-level storage.

- **Robotics in Healthcare** is used for applications such as telesurgery, where a 3D video camera and multiple sensors on a robotic arm help maneuver and capture a variety of vital patient data and transmit it via a low-latency 5G network.
    - The high-res 3D video of patients' inner cavities helps remote surgeons identify organs and other structures for operation.
    - Large amounts of medical data is produced during surgery and must be saved for tracking or future machine learning training.
    - In order to provide 3D video visualization, two HD video streams from the surgical system must be processed and transferred to remote displays at the required resolution, frame rates and coding standards of an XR device or 3D display.
    - Under the following assumptions, a typical 60-minute operation will produce 67GB to 334GB of image data:
        - Resolution: FHD (1920x1080)
        - Bit depth: 32 bits per pixel
        - Frame rate: 30 or 60 frames per second (fps)
        - Compression ratio: 1:25 to 1:10

- **Industrial Robots** can be used in a variety of contexts and can significantly reduce manpower and occupational hazards by performing programmable actions. In situations where the environment is dangerous to humans, operators can control from a safe distance.
    - In industrial applications such as **manufacturing,** latency time requirements are less than 1ms between the robot and the base station, with probability of data packet loss being one out of 100,000 – which is based on current ethernet (wired) technology. Ultimately, the goal in many implementations would be to replace cables with wireless technology to achieve flexible configuration and predictable maintenance of industrial manufacturing equipment.
    - For non-stationary industrial robots, such as AGVs (Automatic Guided Vehicles) and delivery robots, Wi-Fi cannot meet all these requirements because transmission can be disrupted. As a result, 5G is considered the right wireless solution for such applications, especially when combined with edge compute solutions requiring lower latency, and stable and secure bandwidth. The rise in number of private 5G networks adopting Network Slicing techniques will drive even higher SLAs for reliability, moving forward.
    - For industrial applications such as **agriculture,** fruit-picking robotic arms have grown quite rapidly in recent years, for instance. Due to the shortage of farm workers caused by COVID-19, many farms do not have enough manpower to harvest, which results in the crop going waste. However, the trend to use robotics in agriculture will continue after COVID-19 as the AI technology is getting mature.

      Examples include, but not limited to:

        - Tomato harvesting robots in Japan from Panasonic.
        - Strawberry-picking robot in Belgium from Octinion.
        - Apple fruit harvester in the US from FFRobotics.
        - Kiwi-picker robots in New Zealand from Robotics Plus.

      These fruit-picker robots generally have several cameras (some even have the infrared camera) and AI chips are at the heart of the machine which uses a series of learning algorithms to recognize the features of fruit, including maturity, color, size and shape. The infrared camera can allow it to work in the evening and even at midnight.

      The memory requirement varies by application and development stage. Take apple- and tomato-picker robots as examples. The ideal working time per charge should be around 5-7 hours. A single image recognition requires three photos, and the entire fruit pick-up time should be controlled within 10 seconds. In the data collection stage, AI needs to store the images of fruits for subsequent data training. As a result, a typical fruit picking robot will require 16GB-32GB of storage space.

        - Photo size: 3MB
        - Number of photos taken each time: 3
        - Battery life per charge: 10

Edge Storage Interface Trends

The eMMC interface is today's mainstream data storage for IoT applications. However, the speed of eMMC is reaching a limit because of its parallel 8-bit transmission. As a result, UFS is poised to become the significant interface in IoT because its serial interface and full-duplex data transfer protocol offer two to four times the peak bandwidth of eMMC. Hence, the UFS interface aligns well with higher Gbps-level throughputs promised in 5G, compared to eMMC.

Also, UFS power consumption in standby mode is similar to eMMC, and although in active mode UFS is higher, it can rapidly transmit more data, allowing a return to idle state faster. Thus, UFS is significantly lower in terms of overall power consumption, and attractive for battery-powered IoT applications.

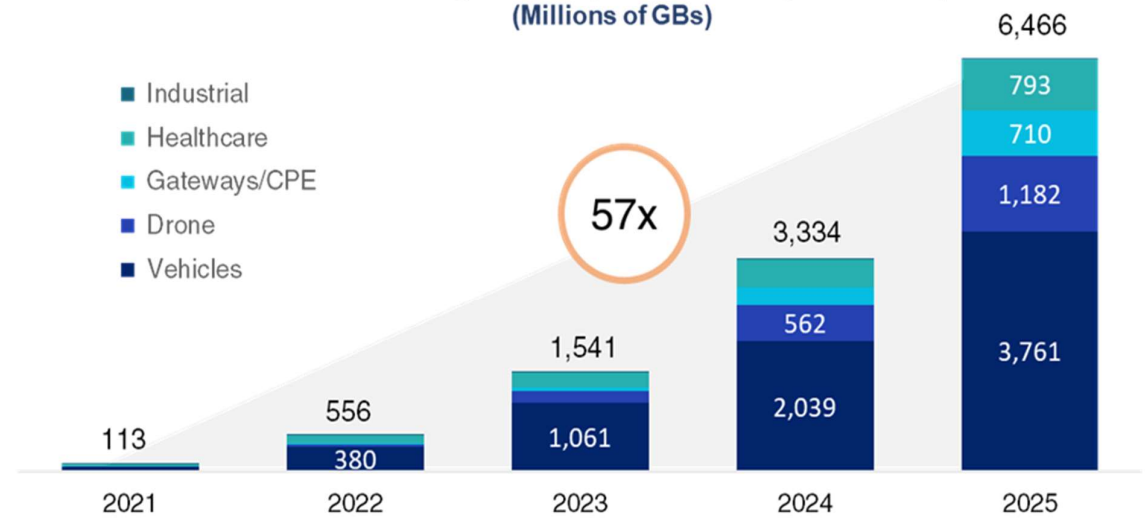
Further, UFS also delivers greater overall system performance and user experience, making it the preferred storage interface for many 5G IoT applications.

SSDs will also become essential for some 5G-driven applications over the next five years. Autonomous vehicles, for example, require several types of storage to accommodate data from multiple in-car sensors. This data needs to be extracted and processed quickly, requiring a shift from eMMC to UFS and embedded SSD – particularly for Level 3 to Level 5 autonomous driving.

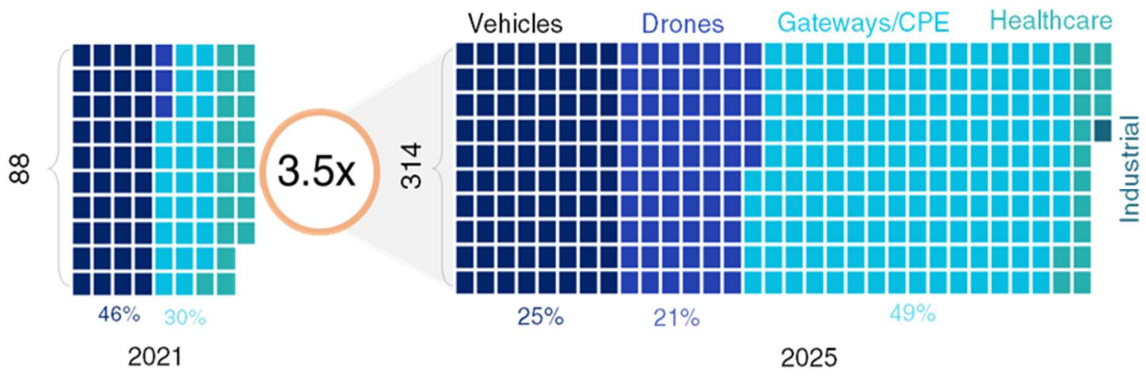
> “
>
> UFS interface will gain significance for some 5G IoT applications which warrant higher system performance to drive high-speed and low-latency 5G-centric experience. However, eMMC will dominate most of the other IoT applications in the 5G era. The selection of UFS vs. eMMC will be driven by the choice of the computing chipset being used in the 5G IoT application.
>
> ”

NAND Flash Consumption Forecast
by Key 5G IoT Endpoints Categories (in Million GBs)

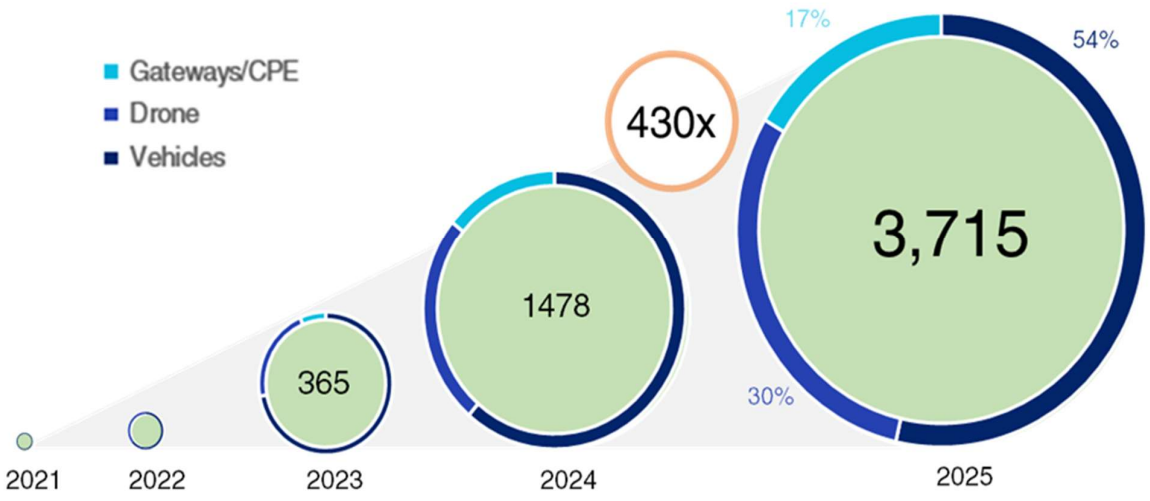
Growth of UFS Storage Consumption Across Key 5G IoT Endpoints (Millions of GBs)



Growth of eMMC Storage Consumption Across Key 5G IoT Endpoints (Millions of GBs)



Growth of SSD Storage Consumption Across Key 5G IoT Endpoints (Millions of GBs)



Key Takeaways

The intersection of 5G, AI and IoT is enabling more data to be captured, processed and actioned at the network edge, helping the industry enter the era of Intelligent Internet of Everything. This brings with it a need for advanced networking, computing and storage in edge devices and endpoints.

Most of the growth at the intersection of 5G, AI and IoT can be expected from **autonomous vehicles**, **medical robotics**, **industrial automation**, **drones**, and next-gen edge **gateways** and **CPEs** segments. The implications of this are especially big for edge storage – both for code and data storage – as it becomes key facilitator. Indeed, **eMMC, UFS and PCIe SSDs** are set to emerge as the dominant storage interfaces driving key 5G-grade use cases and experiences with cumulative demand of close to 19000 Petabytes of NAND-based flash storage over the next five years.

5G IoT developers will need to understand storage requirements and capabilities in order to optimize speed, capacity, reliability, scalability and, importantly, costs as 5G technology rolls out quickly and intelligent IoT implementations build upon this growth.

Authors, Copyright and Other General Information

Brady Wang

Associate Director, Components & Technologies

E: brady.wang@counterpointresearch.com

Neil Shah

Vice President, Research

E: neil@counterpointresearch.com

Graphics: Pavel Naiya

Editors: Charles Moon, Rajiv Gupta

COUNTERPOINT TECHNOLOGY MARKET RESEARCH

Hong Kong | USA | South Korea | India | UK | Argentina | China

20F Central Tower, 28 Queen's Road Central, Hong Kong

info@counterpointresearch.com