MOOR INSIGHTS & STRATEGY

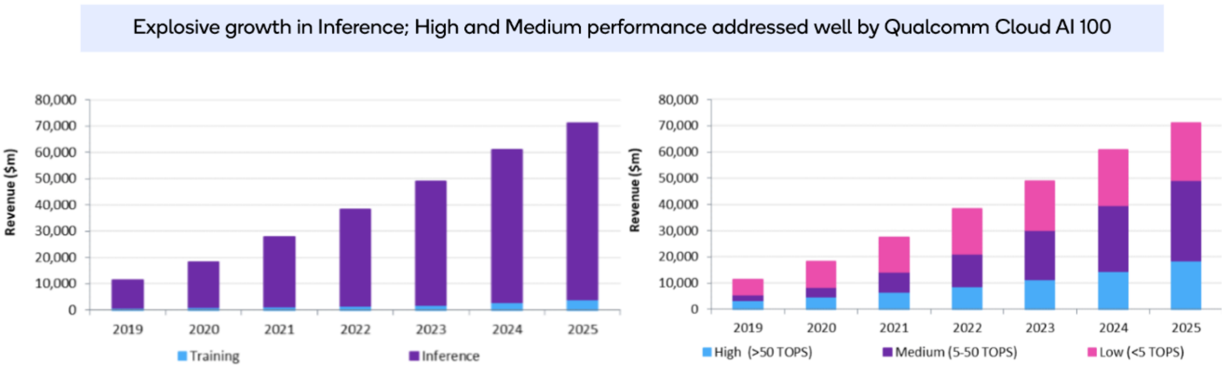
Qualcomm: A New Force In Cloud AI

Mobile Chip Leader begins delivering AI Accelerators with strong performance, a full suite of software and leading power efficiency

Summary

Qualcomm announced the first shipments of the Cloud AI 100 family of products, thus entering the data center, cloud edge, edge appliance and 5G infrastructure markets for accelerated artificial intelligence (AI) processing. Qualcomm has demonstrated its AI and power-efficiency capabilities for many years with the Snapdragon family of SoCs for mobile, networking and embedded markets. The company has now scaled up this technology to compete with NVIDIA, Intel, Xilinx, Amazon Web Services and scores of well-funded startups entering the current Cambrian Explosion of rapid advancements in AI technology. Like that geologic event - in which we find the sudden appearance and rapid diversification of almost all animal phyla emerging in the fossil record - in terms of the AI timeline, it will be interesting to see which solutions thrive and continue and which fall by the wayside as the technological evolution continues.

Figure 1: Qualcomm serviceable market (SAM) estimates


Explosive growth in Inference; High and Medium performance addressed well by Qualcomm Cloud AI 100

(Source: Qualcomm)

Qualcomm expects annual industry revenues for high- and medium-performance data center-class AI platforms to approach some $50B over the next four years. The company appears to have the right combination of hardware and software to capture a significant share of this market from what we have seen. Notably, Qualcomm is not pursuing the multibillion-dollar AI training market, where NVIDIA has what appears to be an insurmountable lead. Qualcomm will focus on the inference market, wherein the trained model classifies input data, which it expects will become more lucrative and accessible than the training market. We concur with this assessment and believe that the company's historical strengths and the new platform's head-turning performance and efficiency will likely position Qualcomm for success. This research paper examines those strengths and specs and outlines the challenges Qualcomm must address to realize its ambitious goals.

THE QUALCOMM GO-TO-MARKET (GTM) STRATEGY FOR ENTERING THE CLOUD

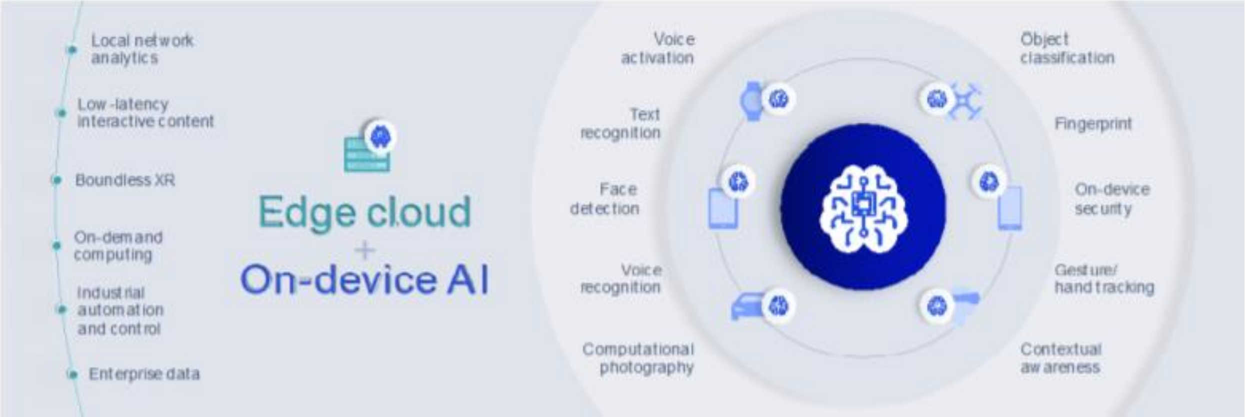
Before we dive into the speed and feeds of the technology and software, let's examine how Qualcomm intends to penetrate the data center platform market. The IT industry is comfortable doing business with the companies that created this market over the last three decades, and Qualcomm is historically a provider of mobile, communications and embedded computing devices, not leading-edge data center gear. Demand for AI accelerators, however, is changing the IT landscape, and consequently, Qualcomm has a significant opportunity to execute a thoughtful market entry strategy.

INITIAL TARGET MARKETS

Qualcomm plans to gain a foothold in the AI server market with one or two as-yet-unnamed hyperscale clients and then expand into edge cloud and 5G infrastructure. Going with a limited set of lighthouse accounts allows the company to focus on a small group of applications where prospective clients believe it may have a highly differentiated offering and are willing to invest the engineering time to thoroughly vet the technology. However, approaching tier-one hyperscale cloud companies with an inference platform is difficult. Many cloud providers, such as Google, Amazon AWS and Alibaba, are developing proprietary SoCs to accelerate internal applications.

The expansion applications are relatively greenfield markets not currently dominated by any single vendor. Interestingly, the edge cloud market is ripe for Qualcomm products as tight integration with a mobile platform such as Snapdragon can create the "distributed intelligence" capabilities Qualcomm has long espoused. Let’s suppose network service providers building this infrastructure buy into that vision. In that case, Qualcomm is one of the only companies that can provide edge servers, appliances and mobile platforms that use compatible hardware and software. It is unclear how this differentiator may materialize, but it is certainly a card Qualcomm intends to play to achieve its vision for distributed intelligence.

FIGURE 2: QUALCOMM'S VISION FOR DISTRIBUTED INTELLIGENCE



(Source: Qualcomm)

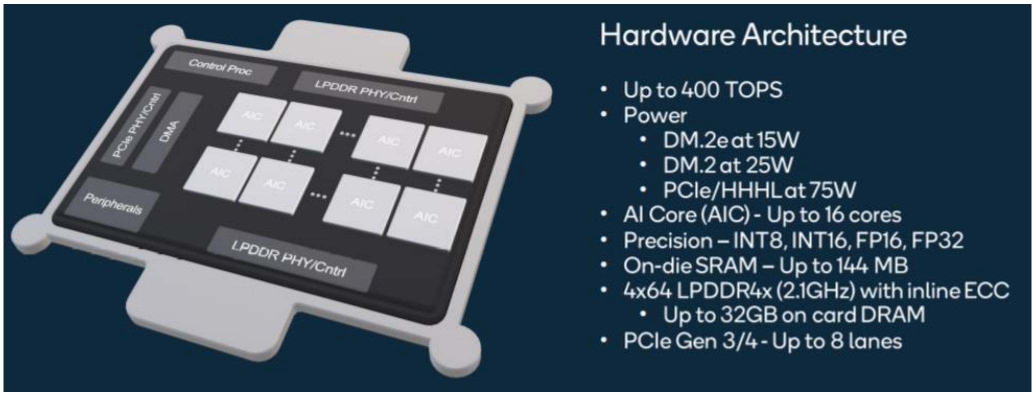
VALUE PROPOSITION

The Cloud AI 100 plans to deliver leadership AI performance and low power consumption, supported by a wide range of software and models initially developed for the Snapdragon ecosystem. Qualcomm built a heritage of power efficiency by engineering milliwatts out of mobile circuits for years, and the Snapdragon AI Engine demonstrates Qualcomm's AI prowess. Today, Snapdragon remains one of the fastest and most efficient mobile AI engines on the market, delivering up to 26 trillion operations per second (TOPS) of AI performance in the newly announced Snapdragon 888. The Cloud AI 100 is a data center implementation built on that wealth of AI research and development, energy efficiency and software. Consequently, the Cloud AI 100 could significantly lower total cost of ownership for data center inference processing.

THE CLOUD AI 100 FAMILY OF AI SOCS AND CARDS

The new Qualcomm AI platform seems to be a thoughtful design, likely assisted by noteworthy hyper scalers; both Facebook and Microsoft presented at the April 2019 unveiling of Qualcomm's cloud strategy. The Qualcomm platform's heart is a new data center-class AI Core (AIC) that delivers up to 400 8-bit integer TOPS. Additionally, some AI models or layers within models run best using floating-point math. Consequently, the AIC supports 16- and 32-bit floating-point tensor operations.

FIGURE 3: THE CLOUD AI 100 BLOCK DIAGRAM



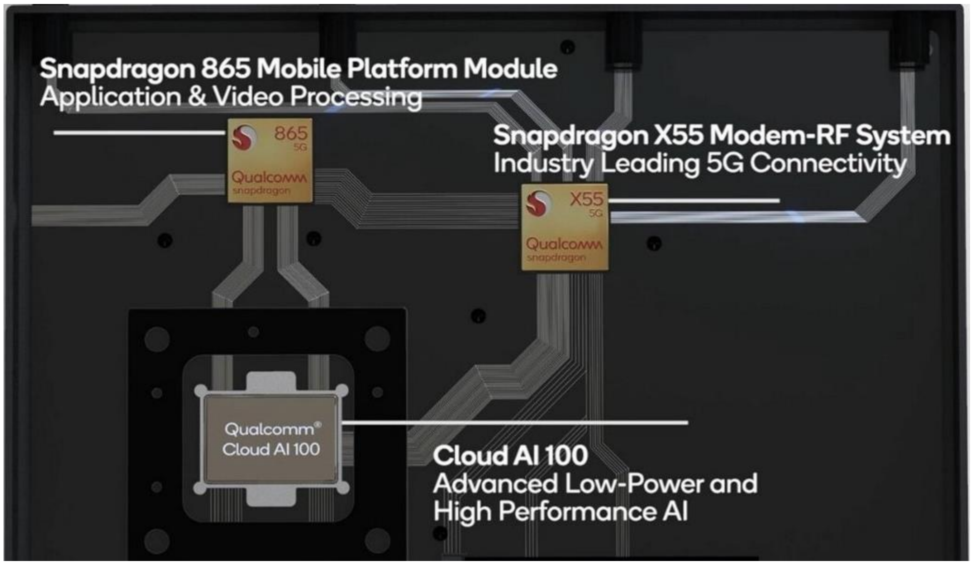
Hardware Architecture

- Up to 400 TOPS
- Power
  - DM.2e at 15W
  - DM.2 at 25W
  - PCIe/HHHL at 75W
- AI Core (AIC) - Up to 16 cores
- Precision – INT8, INT16, FP16, FP32
- On-die SRAM – Up to 144 MB
- 4x64 LPDDR4x (2.1GHz) with inline ECC
  - Up to 32GB on card DRAM
- PCIe Gen 3/4 - Up to 8 lanes

(Source: Qualcomm)

Two form factors are available: Dual M.2 (15W and 25W variants) and PCIe, with up to 16 AICs on the PCIe card for data center throughput, delivering 70, 200 and 400 TOPS at 15-, 25- and 75-watts thermal design power (TDP), respectively. These chips are some of the most efficient mid- and high-performance AI designs we have seen at roughly five to eight TOPS per watt. We would note that several startups intend to deliver comparable efficiency but are not as far along in the development cycle as Qualcomm. Each AIC includes up to 144MB of on-die SRAM, while two DRAM controllers access up to 32GB of LPDDR4 on the cards.

FIGURE 4: THE CLOUD AI 100 EDGE DEVELOPMENT KIT



Snapdragon 865 Mobile Platform Module
Application & Video Processing

Snapdragon X55 Modem-RF System
Industry Leading 5G Connectivity

Cloud AI 100
Advanced Low-Power and
High Performance AI

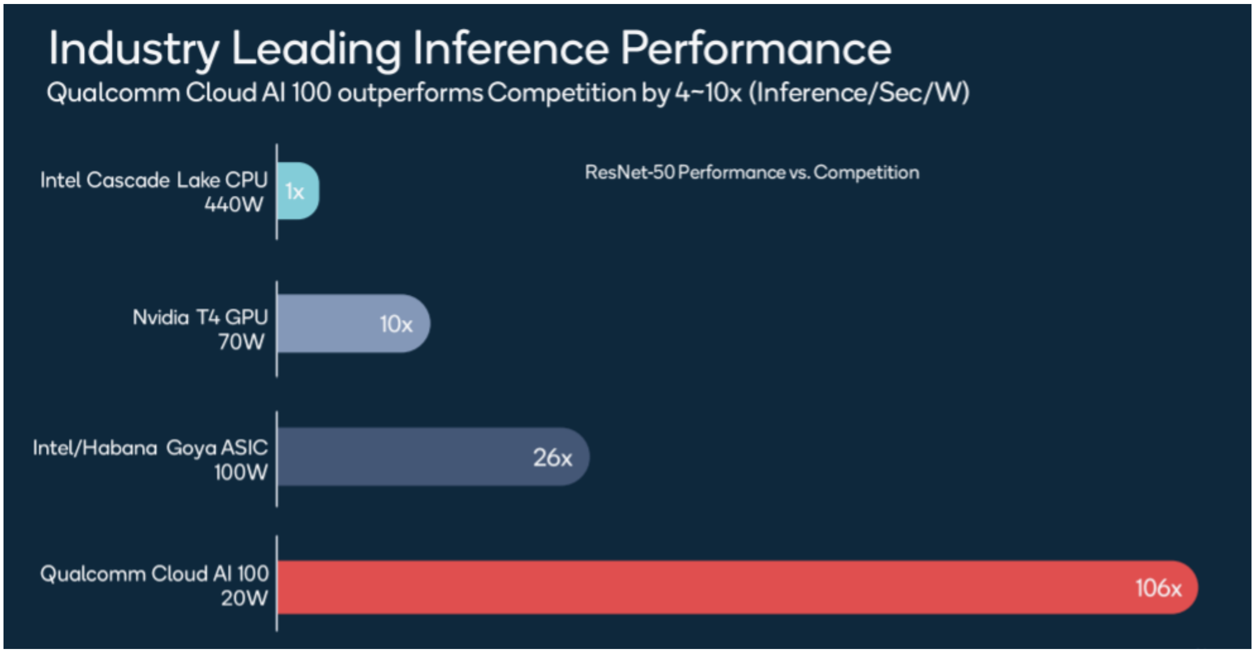
(Source: Qualcomm)

Qualcomm also introduced a new 5G edge development kit to enable ecosystem partners and prospective clients to optimize codes and models for the new platform. The Cloud Edge AI Development Kit workstation comes complete with development software, the M2.e Cloud AI 100 edge card and the Snapdragon X55 5G modem-RF system and 865 processor.

Performance, Latency and Efficiency Comparisons

While it remains too early to expect a broad array of benchmarks on the new platform, Qualcomm has provided a few indicative results. Looking at ResNet-50, a popular image recognition, we see excellent performance and efficiency on the Cloud AI 100.

Figure 5: Qualcomm initial benchmarks for image recognition



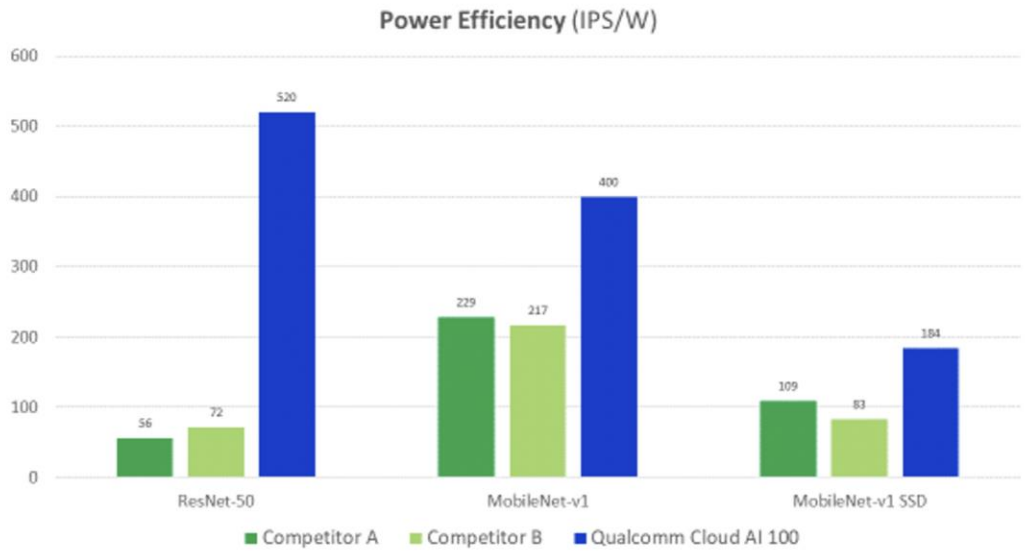
Industry Leading Inference Performance

Qualcomm Cloud AI 100 outperforms Competition by 4~10x (Inference/Sec/W)

(Source: Qualcomm)

I would point out that ResNet-50 and Mobilenet are limited benchmarks by today's standards, especially when compared to NLP models like Google's BERT and OpenAI's GPT-3, which demand significant memory capacity. But this initial data shows an impressive result: The 20-watt Cloud AI 100 M2 delivers roughly 10 times the performance as the 2-year-old NVIDIA T4 and consumes less than a third as much power. Intel Goya is four times slower and requires five times more energy. We anxiously await performance data with larger models, such as the MLPerf benchmark suite.

FIGURE 6: CLOUD AI 100 POWER EFFICIENCY



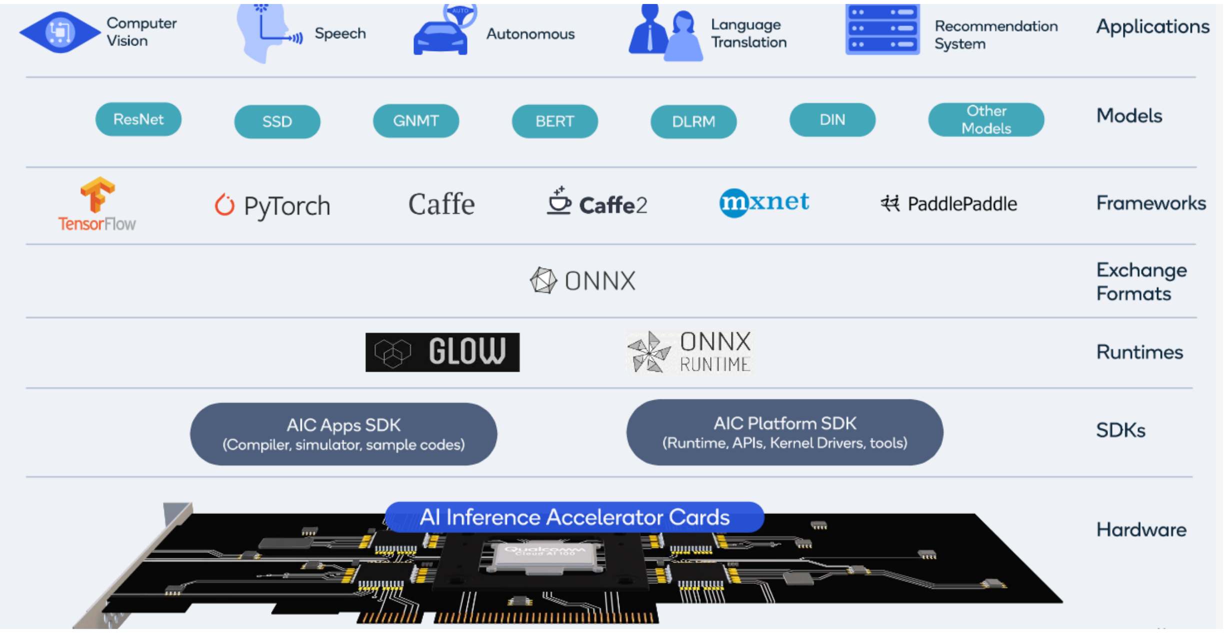
(Source: Qualcomm)

QUALCOMM'S AI SOFTWARE STACK

Most AI silicon entrants fail to provide a complete optimized development environment concurrent with their first production chips launch. Qualcomm is a notable exception, having spent years building software and models for the Snapdragon AI Engine. It takes much more than TensorFlow and a compiler to make it easy to port and develop AI on a new chip.

We would assert that Qualcomm's AI software story is one of the industry's most comprehensive, trailing only NVIDIA in breadth and depth. The figure below shows that the stack now includes two software development kits (SDKs) explicitly targeting the Cloud AI 100: an AIC Apps SDK, which provides an optimizing compiler, and an AIC Platform SDK containing the requisite runtime libraries, APIs, kernels and tools. A natural language model, BERT, and recommendation engine, DLRM, are new additions, likely indicators of early customer evaluations now underway for the platform.

FIGURE 7: THE SOFTWARE DEVELOPMENT STACK FOR THE CLOUD AI 100.



(Source: Qualcomm)

COMPETITIVE LANDSCAPE

Graphics company NVIDIA is by far the leader in the data center AI space, building from its early and aggressive multibillion-dollar presence in AI training. NVIDIA's new Ampere-based A100 supports multi-instancing for the inference space, dynamically creating up to seven powerful AI accelerators. This multi-instance GPU (MIG) capability will likely sell well to cloud service providers seeking to streamline AI infrastructure and will probably be Qualcomm's most serious competitor.

While NVIDIA dominates AI training, Intel Xeon still dominates AI inference since cloud service providers and clients can easily use installed CPUs for many other workloads. Importantly, server CPUs provide vast amounts of memory compared to AI GPUs and ASICs, so memory-intensive applications such as recommendation engines typically run on x86 CPUs. Intel's Habana Labs Goya, introduced three years ago as the fastest inference chip, has not seen any traction publicly; Intel has not yet disclosed its replacement roadmap.

These dynamics may soon evolve as AI models become more complex, demanding far more computational capacity. Last year, the industry group OpenAI estimated that AI model sizes double every 3.5 months, creating new entrant opportunities. Startups like Groq, SambaNova, Blaize, Tenstorrent, Flex Logix and Gyrfalcon are launching inference platforms for solving huge AI models. Meanwhile, giants like Google, Amazon, Baidu, Huawei, Alibaba and Tencent all have proprietary chips to accelerate internal and hosted customer applications. Many of these firms will encounter the software development ecosystem challenges mentioned previously.

We conclude that the Qualcomm Cloud AI 100 has a window of opportunity to become a leader in this space if the company can quickly gain initial traction over the next six to nine months, after which Qualcomm will face increasingly challenging, although perhaps not overwhelming, competition.

Overall Assessment and Conclusions

Beyond the specs, we believe that many customers would choose to do business with an established semiconductor company like NVIDIA, Intel or Qualcomm over a startup—unless the youngster can deliver dramatically better performance and efficiency. Given the impressive, albeit initial, specs of the Cloud AI 100, that threat looks unlikely any time soon. Qualcomm provides rock-solid quality, performance, efficiency and support, as well as a complete software ecosystem for AI inference processing, born from years of experience with Snapdragon. It's a powerful premise, and we look forward to seeing more benchmarks and customer testimonials.

To create a multibillion-dollar data center business, Qualcomm must gain initial success in the GTM strategy discussed earlier. Specifically, even a single design win in one of the Super Seven data center giants would establish the company and the product's credibility in a meaningful way. Barring that, Qualcomm must build preference and deploy the edge cloud data centers, leveraging the company's relationships with network service providers. That will take a lot of work and more time, but remains a viable path for Qualcomm’s success.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

This document, including any supporting materials, is owned by Moor Insights & Strategy. This publication may not be reproduced, distributed, or shared in any form without Moor Insights & Strategy's prior written permission.

DISCLOSURES

This paper was commissioned by Qualcomm Technologies, Incorporated. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.