QUALCOMM: UBIQUITOUS AI FOR 5G

MOBILE CHIP LEADER HOPES TO LEVERAGE PROVEN AI TECH AND POWER EFFICIENCY TO EXPAND FROM ON-DEVICE AI INTO PRODUCTS FOR THE DATA CENTER AND THE EDGE CLOUD.

INTRODUCTION: THE ROAD TO DISTRIBUTED INTELLIGENCE

Smartphones today are the pervasive interface for some three billion people to communicate, take photos and videos, and access personal data and applications—all of which are increasingly dependent on AI and Deep Learning. Consequently, mobile processors must accelerate a wide range of AI features in applications, including image processing, voice recognition, translation and gaming. It is critical that the mobile semiconductors are fast but also must be extremely power-efficient to help conserve battery life.

Some computing industry participants may be unaware of Qualcomm Technologies, Incorporated's (QTI) advances in Artificial Intelligence (AI). This is primarily because QTI AI currently consists of a collection of features and IP in the Snapdragon platform, not a separate product line. However, that perception is changing, as Qualcomm reaches out to a larger audience. QTI will soon roll out new solutions for AI in the data center and edge cloud, where many thousands of scaled-down data centers help push computing closer to edge devices to improve latency and reduce long-haul network traffic.

QTI believes it can build on its historic strengths in AI, research, power efficiency and advanced wireless networks to create an end-to-end platform for "Distributed Intelligence" or DI. The idea behind DI is that AI processing can and should be performed as close to the user as possible, while tightly integrating the workflow to offload to the cloud when needed. So, transparently to the user, the processing could occur on the device, in the cloud edge and/or in the data center, depending on power, data and latency requirements. By interconnecting these three processing tiers via 5G networking, each tier can collaborate to improve understanding and deliver advanced functionality for the user. This is a compelling strategy that positions QTI well in this rapidly evolving space.

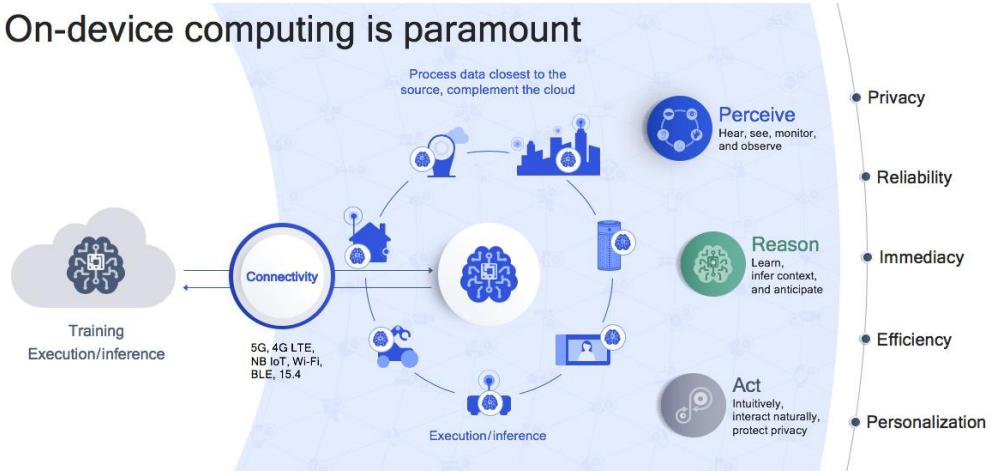
This research paper will explore QTI's strengths and challenges in AI and examine how QTI plans to leverage AI to optimize and deliver 5G network connectivity.

BACKGROUND

Machine Learning (ML), and specifically Deep Learning, will infuse practically every application in every industry with "intelligence," especially when processing unstructured data such as images and voice. AI centers on a new style of programming, wherein data and experience essentially create the program instead of the traditional step-by-step approach. While many CPUs are now enhanced to run AI, accelerators that utilize parallel processing are increasingly being deployed for complex, multi-model inference queries and have become the standard approach to train neural networks used.

QTI's Snapdragon designs now feature dedicated AI accelerators as well as CPU enhancements designed to handle inference workloads efficiently with low latency. Many use cases exist on advanced handsets that utilize Snapdragon's AI Engines, which consists of an on-die Adreno GPU, Kryo CPU, Hexagon Processor with vector extensions and the recent addition of a tensor accelerator).

FIGURE 1: THE IMPORTANCE OF ON-DEVICE COMPUTING



(Source: QTI)

WHY RUN AI ON THE DEVICE?

Mobile and edge devices are the foundations of QTI's business, so it should not come as a surprise that their Distributed Intelligence strategy starts there. Mobile AI apps that traditionally rely on centralized computing resources for complex tasks like natural language and image processing can now embrace partially distributed processing with as much work being done on the device as possible. By running AI on the mobile or edge device, the user can benefit from immediacy and advanced features, all while reducing risks to security and privacy since confidential data is not sent over the

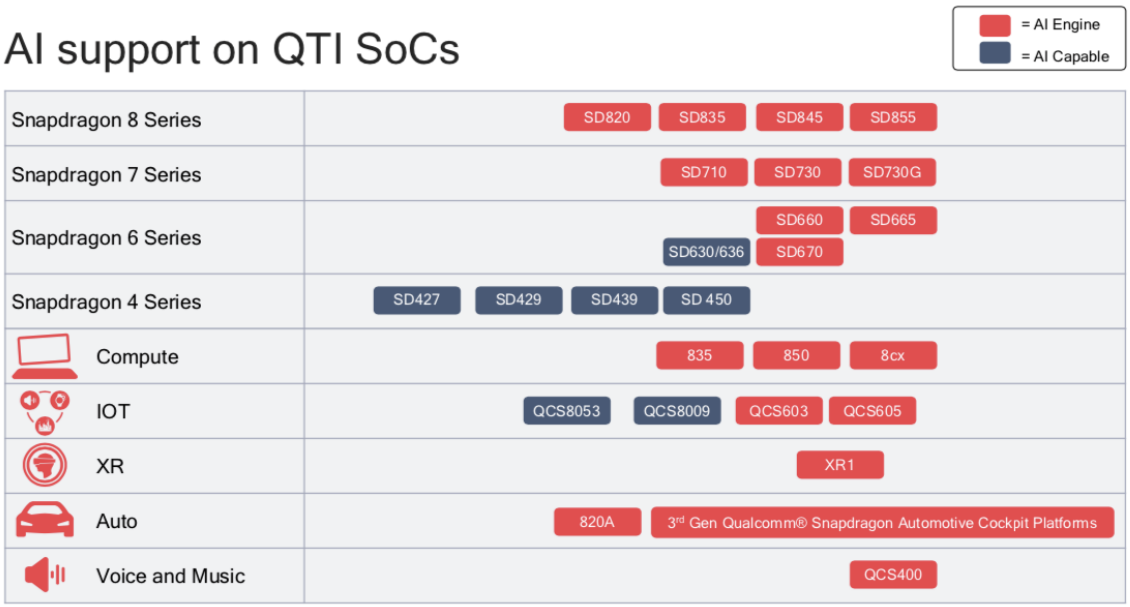
network back to the servers. Also, keeping the processing local whenever possible can substantially improve latencies and provide AI-powered features even when connectivity is marginal or absent.

QTI AI Technology Overview

Hardware: Kryo CPU, Adreno GPU & Hexagon AI Engine

QTI has now shipped over one billion devices for mobile handsets and has infused AI across its portfolio—from Snapdragon mobile devices to products for embedded edge applications and automotive autonomy. This demonstrates the company's belief that AI is no longer an optional, convenience feature; it is just as important to its business as 5G networking. Furthermore, QTI performance is excellent—the latest Snapdragon 855 touts over 7 Tera Ops Per Second (TOPS). QTI's SOC portfolio now supports AI across hundreds of handsets and embedded devices.

Figure 2: QTI SoCs with AI Support



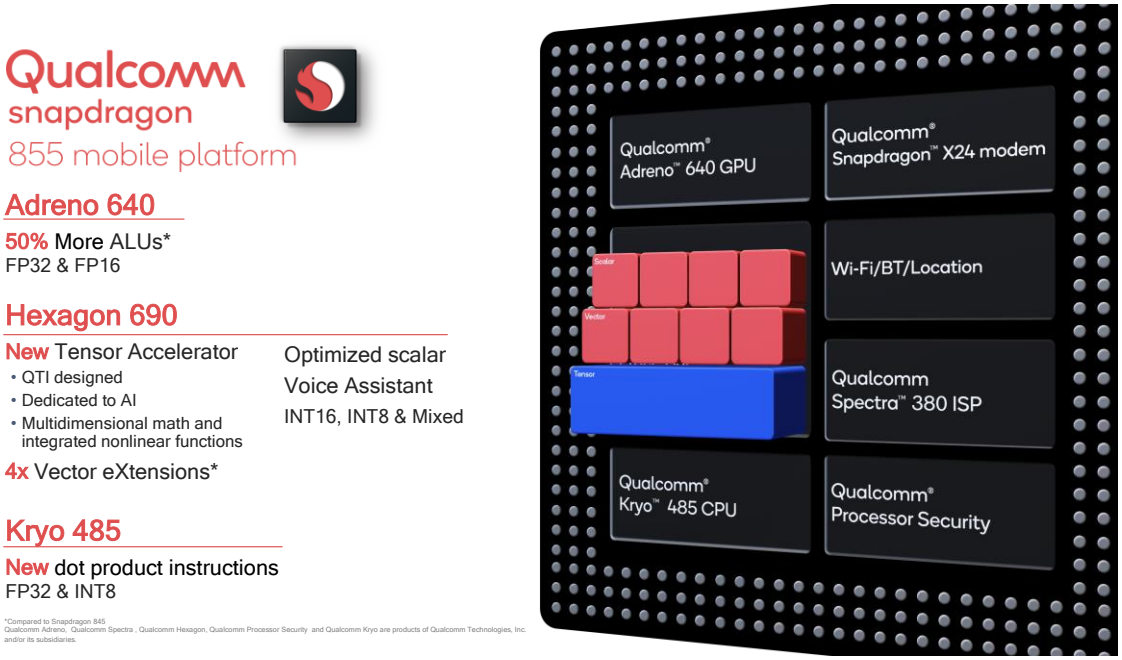
The QTI portfolio of SOCs pervasively supports AI features enabled in over 350 smart phones containing the Hexagon AI engine.

(Source: QTI)

QTI takes a heterogenous computing approach to deliver AI and application performance at low power and cost, placing its Hexagon Processor, Adreno GPU and

Kryo CPU, as well as the modem, security processor and other logic, on the same die. Note that the entire design has been updated to support faster and more efficient AI operations, including 8- and 16-bit instructions, needed for efficient inference processing. Benchmarks published by Anandtech and PCMag have validated the QTI approach, showing that the Snapdragon 855 outperforms the Kirin and Exynos mobile processors.[1]

FIGURE 3: THE SNAPDRAGON 855



The Qualcomm Snapdragon Mobile Platform has been continually enhanced to deliver faster AI capabilities. All processor blocks (CPU, GPU, AI engine and DSP) can be used by AI programmers.

(Source: QTI)

The latest update to Hexagon includes a dedicated Tensor accelerator, akin to the TensorCores found on NVIDIA's latest GPUs. A few AI-enabled applications on QTI-equipped mobile phones that make use of these cores include Dual and Single-camera Bokeh, Secure 3D Face Authentication, Scene Detection, Super Resolution and myriad

[1] https://www.anandtech.com/show/14794/snapdragon-855-phone-roundup-searching-for-the-best-implementations/5
https://www.pcmag.com/roundup/370066/the-best-chinese-phones

other computational photography enhancements. The Hexagon IP that enables these features scales across high- and mid-tier price-points and forms the technology basis for product line extensions into other segments, such as Cloud Edge and the Data Center. This helps to lay the foundation for the company's distributed intelligence vision.

QTI continues to roll out new platforms for edge devices with AI features. An interesting example here is the QCS60x, which landed design wins in applications, such as secure and smart imaging devices—perhaps most notably the Microsoft Azure Vision AI Kit that targets consumer, retail and enterprise security camera applications.

FIGURE 4: QUALCOMM'S VISION INTELLIGENCE PLATFORM

The range of QTI SOCs with the Hexagon AI Engine includes the QCS vision platforms, which is a complete AI platform for secure vision systems.

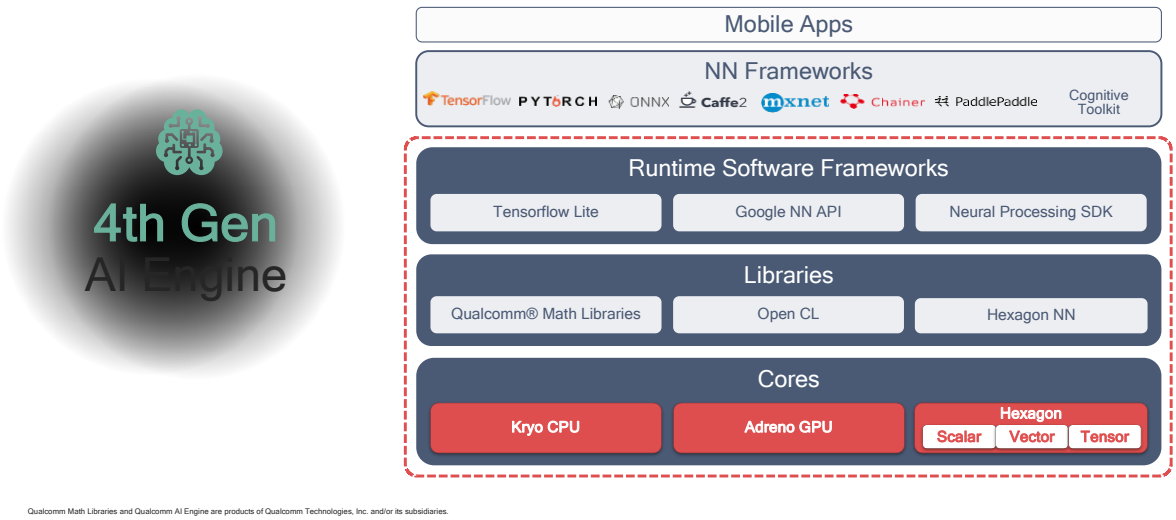
Source: QTI

SOFTWARE: FRAMEWORKS, TOOLS AND LIBRARIES

Of course, any chip needs software to be useful, and QTI built a robust stack for AI applications. This includes the popular Neural Network frameworks PyTorch and TensorFlow and also the frameworks developed by Microsoft, Amazon, Facebook and Baidu. QTI supports the Open Neural Network Exchange (ONNX), a common data format for importing neural networks, and runs them on the processors found in

Snapdragon, where the runtime frameworks and math libraries provide access to the various types of cores. This illustrates an important strategic thrust for QTI: the company wants to support practically every style of AI directly on Snapdragon and has the software to meet the needs of the diverse development community.

FIGURE 5: QUALCOMM'S SOFTWARE STACK FOR AI



QTI's software stack for AI on Snapdragon is quite robust, supporting a wide range of developers and DNN frameworks.

Source: QTI

QTI'S AI MARKET STRATEGY

We believe the Distributed Intelligence strategy makes sense. The AI processing market should continue to experience strong double-digit growth, particularly as the inference market begins to require more performance and power efficiency than is available with CPUs. Let's look at these markets and outline the challenges QTI will need to address to be successful.

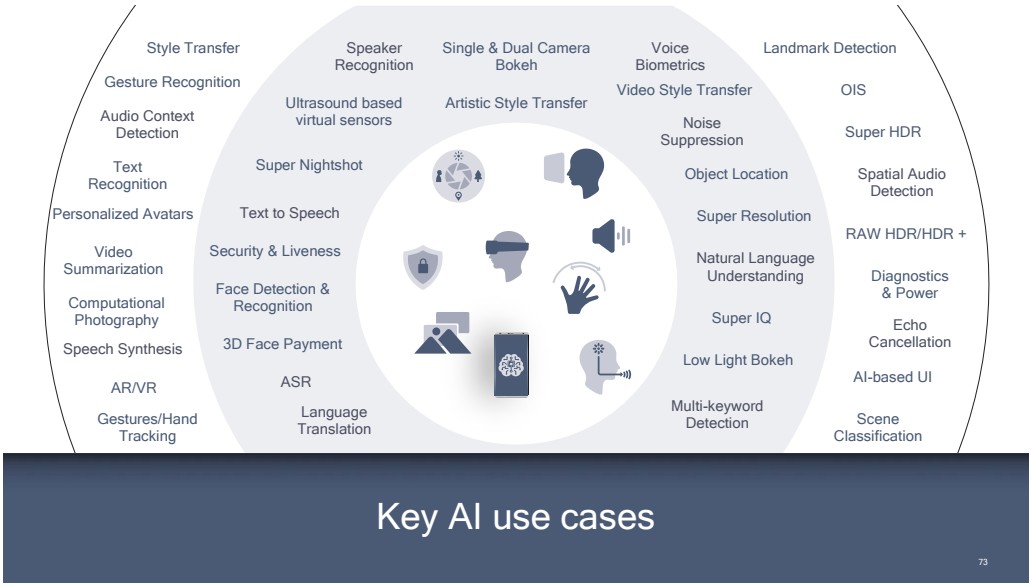
QTI TARGET MARKETS: OPPORTUNITIES AND CHALLENGES

Mobile Devices

Continued leadership in this market is critical for QTI since the company already has a leading global position in mobile devices with well over 1 billion devices in the field. The opportunity is to out-execute other mobile providers in supporting advanced AI features that users will value, such as increasing the accuracy of voice comprehension and improving image processing. Performing AI on the device is critical to provide lowest possible latency and improve privacy and security. Of course, all this on-device compute places tremendous demands on the efficiency of the device SOC.

QTI has created a solid foundation with four generations of AI Engines in the Snapdragon line, and the company will keep pushing the envelope as future use cases require faster chips. Importantly, the IP developed here will trickle down to mid-tier handsets, across to Automotive and IoT, and even up to the servers in the cloud. Figure 6 sheds some light on the vast array of applications that are emerging to enhance the mobile user's experience. Many of these are already in the field today on the AI Engine of Snapdragon mobile platforms.

FIGURE 6: SMART PHONE USE CASES FOR AI



While the vast majority of consumers have little knowledge of AI running on their handset, QTI has identified and helped develop over 30 use cases for AI in smart phones.

(Source: QTI)

Automotive & IOT

If strength in Mobile is a layup for QTI, Automotive and IoT will require increased market development and strong partnerships. QTI already has a presence in both markets but will need to carefully select its target markets since AI at the edge encompasses a very broad range of performance, latency, model types and power requirements. Automotive has a great deal of appeal since a few OEMs can deliver decent volume and relatively attractive revenue. However, some areas of IoT (such as smart surveillance cameras) represent far larger volume where QTI's own massive volumes will help them in a cost-competitive world.

AI in the Edge Cloud and the Data Center

Data Center AI is a highly concentrated market, primarily consisting of the world's seven largest data center operators (Alibaba, Amazon AWS, Baidu, Facebook, Google, Microsoft and Tencent). Some of these companies are experimenting with their own in-house inference chips, and Google has already deployed the Google TPU. Still, a large and respected semi-conductor company like QTI should find a receptive audience if it listens carefully and responds to specific needs related to the data center. This move to the data center is expected in the next year, as QTI previewed its Cloud AI 100 accelerator last April, where they were joined by Facebook and Microsoft on stage.

Facebook shared that its data centers process over 200 trillion inferences every day and that power consumption in its vast data centers has doubled every year for three years running. Clearly, the demand for a scalable, lower-cost and lower-power solution for inference is urgent and becoming business-critical for these massive data center operators. QTI's long-standing history of utilizing the latest process nodes, deep knowledge in signal processing, expertise in low-power computing, and global scale can give them an advantage in this space. The mere presence of Facebook and Microsoft at the AI 100 pre-launch event signaled that the companies are already working closely together.

Use Cases for AI in 5G

5G will utilize AI processing for network optimization and enable new AI-enabled applications, thanks to faster processors and 5G's 1ms or less latency, higher bandwidth at 1 Gbps, and massive connectivity. QTI believes the 5G future will both require and enable "distributed intelligence." This concept acknowledges that one cannot perform all the processing in any one place, be it cloud, edge cloud or edge devices. Ideally, data should be processed locally where it is collected and used, running AI on the mobile or edge device whenever possible. This is not always practical;

some jobs require the processing power and massive data sets that can only reside in central facilities and edge clouds.

Inherent in the concept of distributed intelligence is the need to transfer data to and from the edge device and the cloud. Fortunately, on-device AI can efficiently pre-process a lot of data types before transmission, improving latency and power consumption while reducing the required bandwidth. For example, instead of sending an HD video stream 24/7 from a smart home security camera, local processing can intelligently select images for further processing, such as an open door, a broken window or a stranger. Similarly, voice recognition typically requires full network connectivity back to the edge cloud or data center, but, with advances in neural networks and faster processors, these workflows will soon begin to harness more local processing.

AI Use Cases in Base Stations

The application of AI in wireless transmission and management is in its infancy but will become essential in optimizing 5G networks. AI will assist in transitioning the management of wireless networks from a human-centric model to an automated model, improving signal quality and service levels. Specific applications that reside in base stations include the following:

- Predicting base-station switching handoffs to minimize quality degradation and dropped calls,
- Planning and provisioning beams for frequently traveled paths to optimize signal strength and quality of service, and
- Enabling beamforming in massive MIMO arrays and millimeter wave antennae to optimize transmission quality by identifying the most efficient data delivery route for a particular user (this provides higher quality of service at lower power and bandwidth consumption).

As 5G becomes widely deployed, the availability of fast and power-efficient AI engines will enable continual evolution.

CONCLUSIONS

QTI believes that in order for AI to realize its potential as a transformative technology, AI must become pervasive and collaborative across mobile, edge and cloud computing resources. The company embraces this strategy as “Distributed Intelligence” and recognizes that this will require new products to extend its technology footprint far

beyond mobile and embedded devices. To accomplish this, the company must do the following:

- Continue to innovate in AI acceleration and software on mobile devices,
- Expand that technology in edge device markets, including smart IOT and self-guided devices such as autonomous vehicles, drones, etc., and
- Develop platforms for edge and cloud computing service providers to complement and extend the intelligence that requires more processing power than what is available on the device.

As we began to explore QTI's AI strategy and technology, we were continually impressed by the depth and breadth of its portfolio. The company's traditional strengths in research, power efficiency and connectivity create a foundation on which the company has already built out a formidable portfolio of products and a rich partner and software ecosystem. Going forward, the company is investing in the technologies and market development programs that will help it expand further into intelligent edge cloud and data center environments. In fact, the stated goal for the AI Cloud 100 of reaching 350 TOPS may well put it at or near the lead position when it comes to market in 2020, assuming the company can translate that performance potential into real application benefits (mlPerf benchmarks, for example).

Of course, QTI will face challenges in this market expansion into the data center and cloud edge—from established companies like Intel, NVIDIA and Xilinx, as well as from a vast array of startups. Moor Insights & Strategy believes that when QTI launches its cloud and next generation 5G products, the company will emerge as a significant player in the AI computing revolution. We look forward to hearing more details of their plans and products at the annual Tech Summit in December.

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

DISCLOSURES
This paper was commissioned by Qualcomm Technologies, Inc. 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.