The Deloitte On Cloud Podcast

Gary Arora, Chief Architect: Cloud and AI Solutions, Deloitte Consulting LLP

Title: **Deloitte's Gary Arora unveils 2024's top tech advances: chiplets, quantum, AI, and beyond**

Description: In this Knowledge Short, Deloitte's Gary Arora reveals his candidates for the top five tech breakthroughs of 2024—exploring how chiplets are pushing the envelope on semiconductors, Generative AI is driving mainstream production, and quantum computing is accelerating innovation. Along with advances in solar, edge, and 5G tech, and the sophistication of large language models, these breakthroughs are setting the stage for an amazing future.

Duration: **00:11:55**

Gary Arora:
Welcome back to the On Cloud podcast. I'm your host, Gary Arora, chief architect at Deloitte, specializing in cloud and AI solutions. As we close out 2024, it's clear that this year has been incredible for technology. So, in today's knowledge short episode, we are counting down the top five technological breakthroughs of 2024, innovations that are not just shaping our present, but are set to redefine our future.

So, let's kick things off with number five: chiplets: Semiconductors power nearly every modern device. And in 2023 alone, nearly one trillion chips were sold globally. That's where 100 chips for every person on earth. However, we're starting to hit the physical limits of Moore's Law, which is this idea that the number of transistors on an integrated chip doubles every two years. And it's been true since 1965. Today, the smallest transistors in production measured just two nanometers, smaller than a strand of DNA, enabling 50 billion transistors to fit on a chip the size of a fingernail. So, how do we keep pushing these boundaries as demand continues to soar and shrinking transistors becomes increasingly difficult?

Enter Chiplets: These are small modular chips that can be combined like blocks to create more efficient and cost-effective systems on chip. Unlike traditional monolithic chips, chiplets allow designers to mix and match components such as processing, memory, AI engine, graphics, input output from a library of existing designs so that they can be integrated and optimized independently. Take the automotive industry for example. By integrating dedicated AI accelerators as chiplets, you can independently boost the performance of autonomous driving systems, while chiplets dedicated to battery cell management can optimize energy performance. This modular approach makes it easier to upgrade systems over time without redesigning entire chips. The global chiplet market is projected to hit 150 billion by 2028, growing at a CAGR of 86%. Chiplets are reshaping the semiconductor landscape and will influence everything from consumer electronics to AI systems.[1]

Coming in at number four are the advancements in solar panels and battery technology. Now, this might sound like an odd ball given our technology focus, but data centers are among the largest consumers of energy representing 2.5% of total U.S. power consumption in 2022 and projected to triple by 2030. AI alone consumed 8% of total data center power in 2023. This is expected to grow to 15 to 20% by 2028.[2] Over the past decade, we have witnessed a 900% increase in solar capacity.[3] More solar has been added to the grid in the past four years than any other form of energy generation. 2024 has been a watershed year for energy technology. In solar panels new materials have been developed, just over one micron thick that is 150 times thinner than your

[11] Chiplet Market Size, Share, Growth Drivers, Trends, Opportunities – 2032. Markets and Markets. Accessed November 1, 2024 https://www.marketsandmarkets.com/Market-Reports/chiplet-market-131809383.html#:~:text=The%20global%20chiplet%20market%20size%20was%20valued%20at,a%20CAGR%20of%2086.7%25%20during%20the%20forecast%20period.

[2] BCG on Energy. Unlocking solutions for uncertain times in energy. https://www.linkedin.com/showcase/bcg-on-energy/ Accessed November 1, 2024

[3] MIT Energy Initiative. Seizing solar's bright future.https://energy.mit.edu/news/seizing-solars-bright-future/#:~:text=Consider%20the%20dizzying%20ascent%20of,of%20solar%20installations%20coming%20online.

typical silicon wafer using perovskite.[4] These are so thin they can be worn into your clothing to charge devices on the go. And this material achieved over 28% energy efficiency, surpassing the current 22% from conventional panels.[5]

And this is not just R&D. I'm talking about a functional production line in Germany and within the battery domain, we saw mass production of sodium ion batteries capable of recharging 10 times faster than traditional alternatives and lasting 50,000 cycles. AI has been playing a pivotal role in optimizing these new battery technologies by continuously monitoring health, predicting failures and optimizing charging cycles. So, why is this matter to us? Well, these breakthroughs in solar and battery technologies remove one of the biggest hurdles in our industry - the energy barrier. By making clean, efficient and affordable energy more accessible, we can meet the rising energy demands of AI, cloud computing, and edge devices without being constrained by power consumption or operational costs.

Number three on the list is quantum computing, specifically quantum cryptography and encryption. Now quantum computing has been evolving from experimental into practical applications. But what's particularly exciting is how accessible this new technology is becoming. Multiple tech players are now offering quantum computing as a service, abstracting away the need for expensive quantum hardware. Now quantum computers are different than your current classical computers, and that quantum computing isn't restricted to processing data in bits—zeros and ones—and instead rely on principles of physics and quantum mechanics. This makes quantum computers incredibly powerful at simulating nature and complex processes like developing new materials and compounds.

Unfortunately, they're also exceptionally good at breaking the cryptographic systems we use today. Asymmetric encryption underpins the security of almost all software, billions of devices and most Internet communications. Our current RSA technology is around 40 years old. It's remarkable that it's still secure. It's within your organization’s strategic horizon, “the harvest now and decrypt later threat,” is real. That is when malicious actors collect encrypted data today planning to decrypt it later, when quantum capabilities become widely available. This poses a serious risk to sensitive data. Think emails, medical records, photo libraries, text messages, and even national security information.

So, what happened in 2024? Well, in August, the National Institute of Standards and Technology released its final set of post quantum encryption standards designed to withstand quantum attacks. The commercial sector is quickly adapting these new standards. Organizations must develop strategic plans, allocate resources, and stock the complex process of upgrading cryptographic systems now. This isn't just about staying ahead; it's about safeguarding the foundation of your digital infrastructure as we enter the quantum era.

Number two on our list is the powerful convergence of edge computing and 5G. Imagine a world where your smartphone, autonomous car, or home security device processes complex data instantly, right at the source. That's the promise of edge computing: moving computation closer to where the data is generated, reducing latency, lowering energy costs, and enhancing privacy by minimizing the transmission of sensitive information to a distant server. If data is the new oil, then with edge computing, we are refining it right where we find it. In 2024, we have witnessed significant breakthroughs, number one among them is the global expansion of 5G networks. 5Gs lower latency speeds—now that's the duration between a device receiving instruction and when the operation is completed—can be as low as 5 milliseconds. Compare that to at least 100 milliseconds in 4G. The 5G private network market was valued at $1.9 billion in 2023 and it is estimated to reach $70 billion by 2032 (That's a CAGR of 49%).[6]

The second development in this space is the advancements in edge AI, the rise of smaller, more efficient AI models allowing complex algorithms to run on less powerful edge devices like mobile devices, kiosk or autonomous vehicles. Companies are investing heavily in localized micro data centers. In early 2024, we saw the first commercial edge computing and AI enabled system on the International Space Station reducing dependency on mission control for data processing. So, why should this matter? Processing data at the edge lowers bandwidth requirements and energy consumption, leading to substantial cost savings. That's why autonomous vehicles have powerful computers that live in the vehicle itself. The same goes for medical equipment or devices used in dangerous types of manufacturing, where more computing needs to be done on the fly. This synergy of edge computing and 5G is unlocking new business models and services from smart cities to industrial IoT or requiring an inclusive strategy that integrates edge, AI, and cloud.

And that brings us to number one on our list, the monumental leap in large language models and Generative AI. Just look back to 2023 to appreciate how far we have come. In 2023, we saw the rise of LLMs. But they often left us wanting more; they were sometimes confidently incorrect or provided verbose answers that lacked substance. Prompt engineering was the secret to extracting useful responses from them. Fast forward to 2024 and we have witnessed LLMs evolve from curiosities into mainstream production tools reshaping entire industries. The Turing Test is becoming less of a test and more of a conversation starter. Key breakthroughs in 2024 include the rise of RAG: retrieval of augmented generation, which combines Generative AI with the real time contextual data by grounding responses in specific information sources, making LLMs far more dependable.

We also saw the mainstreaming of multimodal interfaces. In health care, for instance, models can now analyze medical images alongside patient history and genetic information, offering more accurate and comprehensive diagnosis. This fusion of data types allows for more comprehensive insights, but perhaps the most exciting breakthrough is Agentic AI. We have shifted from reactive to proactive AI systems. These AI agents now act autonomously, executing tasks without human intervention to achieve set goals. They understand their environment, make decisions, and get things done. If 2023 was the year we talked to AI, 2024 is the year AI agents started taking action and delivering results. For technology leaders, embracing these developments isn't optional, it's imperative. The era where AI was simply a back-end tool is over.

AI is now at the forefront. Actively shaping strategy, driving automation, customer experience, code generation, data analysis and enabling unprecedented innovation. The organizations that leverage these advanced LLMs today will not only lead the future but define it.

[4] Razor-thin solar panels could be ‘ink-jetted’ onto your backpack or phone for cheap clean energy. CNN. Accessed November 1, 2024. https://www.cnn.com/2024/08/09/climate/solar-panel-inkjet-renewable-energy/index.html#:~:text=At%20just%20over%20one%20micron,tools%20like%20an%20inkjet%20printer.

[5] Energy Box. Oxford PV Launches Commercial Production of Record-Breaking Tandem Solar Technology. Accessed November 1, 2024. Launches Commercial Production of Record-Breaking Tandem Solar Technology

[6] Allied Market Research. 5G Enterprise Private Network Market Size, Share – 2032. Accessed November 1, 2024. https://www.alliedmarketresearch.com/5g-enterprise-private-network-market-A12225

And with that thank you for joining me on this week's Knowledge Short. If you enjoyed this episode, please like, rate, and subscribe to the On Cloud podcast. For more insights and episodes, visit deloittecloudpodcast.com (all one word). Until next time, I'm Gary Arora.

Operator:
This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to Deloitte.com/about.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Visit the On Cloud library
www.deloitte.com/us/cloud-podcast

About Deloitte

As used in this podcast, “Deloitte” means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.