

SPECIAL REPORT

EMPOWERING THE EDGE

Table of Contents

CHAPTER 1

2 **The Antenna is the Edge: Why Connectivity Plays a Critical Role**

CHAPTER 2

6 **From Chip to Cloud: Building Trust into Physical AI at the Edge**

CHAPTER 3

9 **Achieve Megawatt-Scale AI with SiC JFET Technology**

CHAPTER 4

14 **Edge AI Does Much More for Manufacturers Beyond Production- Line Automation**

CHAPTER 5

19 **Why Sensor Fusion is the Real Differentiator in Physical AI**

CHAPTER 6

21 **Performance Scaling— with Balance—in Edge AI**

CHAPTER 7

26 **Evolving Antenna Requirement for LEO and MEO**

Empowering the Edge

By **ALIX PAULTRE**,
Editor-at-Large, Microwaves & RF



Society has been transformed

by cloud computing, as it enables people and devices to communicate seamlessly with one another and address challenging applications from medical devices to military systems. The capabilities of wireless connectivity are being further pushed by the ongoing deployment of edge computing, which enhances functional capabilities while reducing the wireless bandwidth and latency in an IoT-oriented, cloud-based, edge-computing environment. Today, edge computing is being complemented by AI-based solutions to maximize the computing power of the devices.

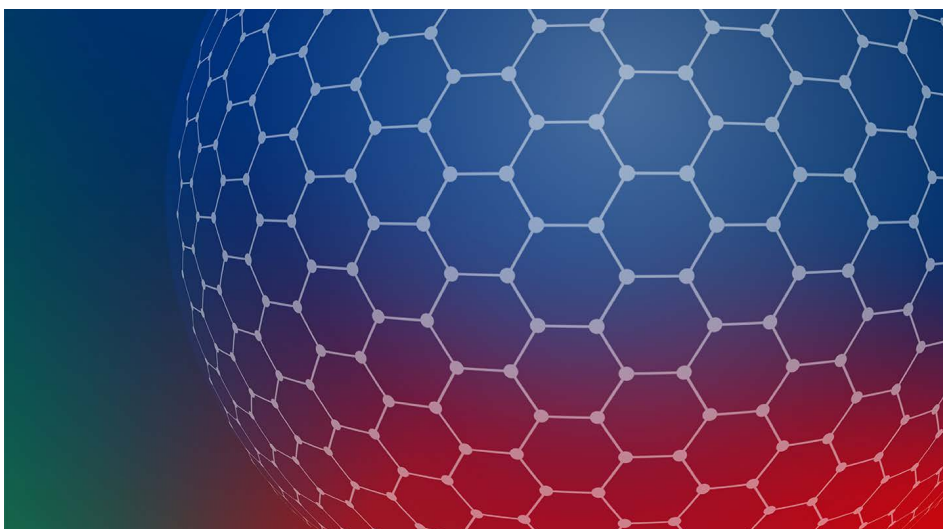
Wireless communications and networking have exploded in their functionality and capabilities as the core technologies involved continue to evolve, providing better components and subsystems that can uplift the finished product. This eBook collects together articles that can help you make good selection choices when developing your edge-computing solutions.

CHAPTER 1

The Antenna is the Edge: Why Connectivity Plays a Critical Role

Intelligence at the edge is one of the defining trends in IoT today. But intelligence must travel somewhere to matter, and the radio link carrying it rarely gets the same scrutiny as the silicon.

DERMOT O'SHEA, CEO, Taoglas, ADAM COUSIN, Sales and Business Development Manager, IoT Solutions EMEA, TELIT CINTERION, and PAVESH GOVENDER, Head of Northern Europe IoT Sales, Vodafone



A three-decibel reduction in antenna efficiency doesn't just weaken a signal — on NB-IoT or LTE-M networks, it can force the network to command up to 128 retransmissions of the same message, multiplying energy use. It's a detail most edge AI conversations never consider, because they're too busy talking about the chip.

That's the blind spot the industry keeps stepping into. Intelligence at the edge — devices that sense, decide, and act without waiting on the cloud — is one of the defining trends in

IoT today. But intelligence must travel somewhere to matter, and the radio link carrying it rarely gets the same scrutiny as the silicon making the decisions. As Dermot O'Shea, CEO and co-founder of Taoglas, put it: "An Edge AI device that can't reliably move data is just an expensive way to make a decision that nobody hears."

The Physics that Compute Can't Escape

Processors keep shrinking, cheapening, and sipping less power with every product cycle. Antennas don't get that

gift. They're bound by the physics of the wavelength they need to resonate with, and no amount of clever firmware changes that.

O'Shea, whose company designs antennas for demanding industrial and outdoor deployments, pointed to real-world examples where the toughest requirements come down to environment: "If there are extremes in temperature and if it's outdoors, it's probably going to be subject to impact." He cited metro and traffic-light systems in Toronto, where even a bird strike on an antenna enclosure in freezing conditions can crack it open, forcing far more rugged designs than most teams have budgeted.

Shrink an antenna and efficiency, range, and bandwidth all suffer. Get the impedance match wrong, and a device can lose half its transmitted power before it ever leaves the board. Wrap the same design in a metal enclosure — standard for industrial or automotive housings — and range can collapse by around 30% because this screens the antenna.

CHAPTER 1: The Antenna is the Edge: Why Connectivity Plays a Critical Role

Since battery life on constrained networks degrades in steps rather than a smooth curve, a mediocre antenna doesn't cost a business a little efficiency. Rather, it costs a multiple of the energy per message, because the network compensates for a weak link by repeating transmissions rather than boosting them gently.

There's a second, quieter problem sitting on top of this: self-interference. The second harmonic of a common cellular uplink band sits almost exactly on the GPS L1 frequency, alongside Wi-Fi, Bluetooth, and switching noise from nearby digital circuitry. The field symptom is usually described simply as "GNSS is flaky sometimes." Teams can spend months chasing it as a firmware or coverage bug before realizing it's an RF layout issue from the start.

When Position Data Can't Be Trusted

Positioning brings its own emerging concern, and it's shifting from an accuracy question to a security one. GNSS jamming is more mundane than it sounds. In trucking, the dominant source is often cheap, driver-installed devices meant to defeat fleet tracking, which end up jamming every other receiver in range as collateral damage. Spoofing is the more dangerous cousin. Rather than knocking a signal out, it feeds a device in a plausible, precise, entirely wrong position that the cloud trusts without question.

This is squarely O'Shea's territory as an antenna specialist, and the framing worth adopting industry-wide is straightforward: Wherever position data drives a decision or a payment,

GNSS should be treated as an untrusted input rather than simply an accuracy problem.

Real mitigations exist at the antenna level, from shaping radiation patterns so that they favor satellites overhead rather than jammers near the horizon, to using polarization discrimination and multi-band reception, since many low-cost jammers only target a single frequency.

The Power Budget Nobody Wants to Own

If antenna performance sets the ceiling, power budgeting is where most projects break, and this is where Adam Cousin, sales and business development manager at Telit Cinterion, speaks from two decades of watching the same mistakes repeat.



CHAPTER 1: The Antenna is the Edge: Why Connectivity Plays a Critical Role

“People are doing essentially today what they’ve been doing for the last 20, 25 years,” said Cousin, “and that is switching the system on, switching the modem on, transmitting a small amount of data, and then switching it off again.”

Layering AI onto that pattern only works if it reduces what needs to be sent, not if it adds another power-hungry process competing with the radio for the same battery. Cousin warned, “If the AI system uses more power than cellular communication, then you’ve defeated the object of using AI in the system.”

That tension plays out clearly in large-fleet deployments — utility meters, smart-city sensors, asset trackers — where shaving a couple of euros off a bill of materials by pushing compute to the cloud can quietly add tens of milliamp-hours of daily radio activity. At scale, across tens of thousands of units, that tradeoff doesn’t save money; it just moves the cost from procurement budgets to truck-rolls for battery replacements years later.

The smarter approach, borrowed from recent cloud-positioning techniques, is to send a small raw signal snapshot uplink-only and let a server resolve the computation, rather than requesting a computed result back. That’s because a round trip on a constrained network is almost always more expensive than one-way traffic.

The same logic applies to network selection. A marginal impedance

match forces a power amplifier to work harder for the same effective output, draining a battery that better RF matching would have preserved.

Making the Case for Judged Adoption

None of this is an argument against edge AI. It’s an argument against adopting it uncritically. Cousin is candid about the hype surrounding the space: “The problem with the hype curve of any technology is that the people who are out there evangelizing about it will make you believe this technology A exists, B is available, and C is completely infallible,” when in reality, none of that is yet true for every use case.

His advice to teams starting out is blunt: “Ask yourself why. Ask yourself what you’re trying to achieve, what you’re trying to fix, improve, or monetize” before committing hardware and budget to an AI-enabled redesign.

The clearest wins so far sit in industries already comfortable with technical complexity. Pavesh Govender, head of Northern Europe sales at Vodafone IoT, points to fleet management, using AI cameras for driver-behavior monitoring and predictive maintenance, and remote patient monitoring in healthcare, as two applications furthest along. Both run largely on mains or rechargeable power rather than pure battery deployments.

Crucially, Govender noted, these sectors “are not deploying AI because

it’s fashionable,” but because it solves a defined operational problem. Battery-only use cases remain the harder frontier, and progress there depends as much on RF efficiency as on model design.

Connectivity Once Data Leaves the Device

Getting a signal off the device is only half the job. What happens to that data afterwards matters just as much, particularly as connected products are increasingly expected to remain in the field for a decade or more. Security for a device with a 10- to 15-year lifespan can’t rely on a single shared key across a fleet, since one extracted key would compromise every unit. Per-device cryptographic identity, encrypted transport, and hardware-protected keys are becoming baseline expectations.

Supply-chain posture is following the same trajectory. Certifications like ISO 27001 and defense-grade frameworks such as CMMC Level 2 are increasingly used as a proxy for whether a connectivity partner can be trusted with sensitive data at all.

That same interconnectedness is what unlocks the actual business value of edge intelligence, and it’s the point Govender returns to repeatedly. The real payoff of a device making decisions locally, he argued, is “only really felt when this is interconnected into enterprise systems.”

A smart sensor that never reports

CHAPTER 1: The Antenna is the Edge: Why Connectivity Plays a Critical Role

back into an operational workflow is a technical achievement without an outcome attached to it. Scaling that across borders adds its own complications. Network capabilities and regulatory environments can vary significantly between markets, making it difficult to deliver a consistent experience across geographies. Addressing those differences remains one of the most challenging aspects of global IoT deployment.

Designing the Nervous System, Not Just the Brain

The most forward-looking developments — federated learning across device fleets, swarm coordination between robots or drones, split compute between a device and a nearby edge node — all share one buried assumption: That a wireless link exists whereby it's good enough to carry model updates and shared observations reliably, within a usable time window. Take that link away, and the entire architecture collapses back into a single disconnected device running on a stale model.

That's the shift worth internalizing, and it's one all three specialists converge on from their own vantage points. Govender stated the underlying philosophy: "I would not look at the components of an edge AI solution as discrete components, but again, as an entire system...the sum of the parts would always be greater than each component."

O'Shea agreed, echoing that same conclusion from the antenna side: "I like that saying — it's one system." And Cousin, closing the loop from the module and connectivity side, brought it back to discipline over hype: "Get your ducks in a row when it comes to this stuff, and don't just leap, because it could be quite a costly mistake."

As intelligence moves from the chip to the network of devices around it, the antenna stops being a component chosen late in a design cycle and becomes the substrate that the whole system depends on. The teams that treat connectivity as a core design input from day one are the ones likely to get furthest past the pilot stage. Everyone else will keep discovering, expensively, that a brilliant decision made at the edge is worthless if nothing ever hears the result.

This article draws on insights shared during a [Taoglas-hosted panel discussion on edge AI and IoT connectivity](#), featuring Dermot O'Shea (Taoglas), Adam Cousin (Telit Cinterion), and Pavesh Govender (Vodafone).

[view this article online](#)

 [BACK TO TABLE OF CONTENTS](#)

CHAPTER 2

From Chip to Cloud: Building Trust into Physical AI at the Edge

AI-generated content may be incorrect, which means embedded trust must begin at the silicon level.

CHRISTOPHE NICOLAS, Senior Vice President GTM, Kudelski Labs



ID 44924818 © Andose24 | Dreamstime.com

The next generation of AI won't live in the cloud. It will live inside vehicles, robots, industrial equipment, and critical infrastructure, making real-time decisions without waiting for cloud connectivity. That shift fundamentally changes how engineers must think about cybersecurity. In physical AI systems, security failures become safety failures.

When AI was limited to chatbots or recommendation engines, a mistake was embarrassing or costly. If that the same technology steers an auton-

omous vehicle through the streets, interprets sensor data on a factory floor, or controls pumps in a water treatment plant, a spoofed input or hacked model can become lethal. Security stops being an IT problem and becomes a safety problem.

To make physical AI safe for widespread deployment, we must rely on data integrity, system resilience, and the ability to maintain security over product lifetimes measured in decades, not software release cycles.

This article explains why embedded

trust must begin at the silicon level, examining the role of hardware roots of trust, secure enclaves, device identity, and lifecycle management in protecting autonomous systems. It also explores how engineers can design resilient systems that continue operating securely in unpredictable real-world environments.

The Model T Moment

In 1908, Henry Ford put the world on wheels. He later observed that if he had asked people what they wanted, they would have said faster horses. Ford didn't build a faster horse. He replaced the entire system of how people moved. Real innovation doesn't merely improve the existing thing — it reframes the problem.

Physical and autonomous AI is our Model T moment. Intelligence is migrating out of the data center and into remote devices, vehicles, and industrial systems. These systems fuse sensors, software, hardware, and machine learning into a single decision-making unit that must operate in the real world, but not without risk.

CHAPTER 2: From Chip to Cloud: Building Trust into Physical AI at the Edge

The Model T arrived before safety regulations fully caught up. Innovation came first, and governance and security followed. We're at a similar point with autonomous systems. Capability is racing ahead of the trust and security infrastructure required to deploy it safely at scale.

Why Cloud-Centric Security Falls Short

Cloud-only security assumes that decisions can be made centrally and pushed out when needed. But physical AI systems don't work like that. They operate in real-time, in unpredictable environments, often without a reliable connection.

A self-driving vehicle can't phone home to decide whether to brake. An industrial robot can't wait for a cloud service to validate a sensor reading when a human is nearby. Trying to secure these systems exclusively from the cloud is like servicing a racecar from the factory while it's traveling at 300 km/h. By the time information travels out, is processed, and returns, the moment has already passed.

Centralized, cloud-only architectures introduce latency, expand the attack surface, and create dependence on connectivity that may not exist when it's most needed. Security therefore has to be embedded where the decisions are made, on the device itself. Security defenses must ensure that a compromise doesn't hand over control of the system. Risks occur when physical

devices, running in real-time, leave the network unexpectedly.

Building Trust from the Silicon Up

Securing physical edge AI at scale requires treating security as a foundational property of the system rather than an add-on layer. Several elements are essential. Hardware roots of trust and secure enclaves provide the strongest starting point. Cryptographic keys and sensitive operations must be isolated from the application layer so that even a compromised operating system or AI model can't freely extract or misuse them.

Furthermore, device identity must be established at provisioning and maintained throughout the product's life, including through deployment, field updates, and eventual decommissioning. Lifecycle management isn't optional when devices are expected to remain in service for 15 years or more.

Security evaluation, design practices, certification, and compliance frameworks must keep pace with the technology and regulatory environments. End-to-end system security should span the individual device, the surrounding infrastructure, and the connected ecosystems in which these systems operate.

Equally important is cryptographic agility. Algorithms that are considered strong today will face new threats tomorrow, including those arising from advances in quantum comput-

ing. Hardware and firmware should be structured so that security mechanisms remain updatable rather than frozen deep in silicon, where a change would require a new tape-out.

Key management systems need the ability to provision, rotate, and revoke credentials, including quantum-resistant ones, over the full operational lifespan of the product. And in-field updates to cryptographic algorithms must be possible as standards evolve and new vulnerabilities appear.

Many are looking at a hybrid approach as a practical way to secure products now entering the market. A hybrid approach combines classical algorithms running alongside quantum-resistant alternatives during the transition period. This preserves interoperability with existing infrastructure while establishing longer-term resilience. As the broader ecosystem migrates and classical algorithms are deprecated, systems designed with crypto-agility can adapt without wholesale hardware replacement.

Hardware alone isn't sufficient. The software and firmware layers must be designed to manage these capabilities across years or decades of deployment. The goal isn't a one-time transition to a new set of algorithms, but a sustained capacity to adapt as threats, standards, and computing capabilities continue to change.

Designing for the Real World

Physical AI will transform vehicles,

CHAPTER 2: From Chip to Cloud: Building Trust into Physical AI at the Edge

industrial systems, and critical infrastructure in ways comparable to the arrival of the automobile. The capabilities and innovations are impressive, but true usage and deployment will come down to whether it can be trusted. Trust depends on engineering decisions made at the silicon level and carried through the entire lifecycle of the device.

Engineers who treat security as an architectural requirement from the start by embedding roots of trust, maintaining strong device identity, designing for updatability, and planning for cryptographic change will build systems that can operate securely even when the network is unavailable, the models are tampered with, and the environment is hostile. Those who treat security as a later addition will discover, as the automotive industry once did, that catching up after the fact is far more difficult and far more costly.

The Model T changed how the world moved because the engineering underneath was solid. Physical AI will change how machines decide and act in the physical world only if the same discipline is applied to trust.

[*view this article online*](#)

 **BACK TO TABLE OF CONTENTS**

CHAPTER 3

Achieve Megawatt-Scale AI with SiC JFET Technology

With the proliferation of generative AI models such as ChatGPT, Llama, Gemini, and DeepSeek, data centers are under pressure to deliver unprecedented computational power, energy efficiency, and thermal management.

SALLY FENG, Technical Marketing Analyst, onsemi



The rapid evolution of artificial intelligence (AI) is reshaping the architecture and demands of modern data centers. With the proliferation of generative AI (GenAI) models such as ChatGPT, Llama, Gemini, and DeepSeek, data centers are under pressure to deliver unprecedented computational power, energy efficiency, and thermal management. These demands are driving a shift toward high-voltage, high-density power architectures, where silicon-carbide (SiC) technology is emerging as a key enabler.

As AI data centers scale to meet these requirements, the need for efficient, reliable, and cost-effective power delivery becomes increasingly critical. SiC Cascode JFETs (CJFETs) are engineered to enhance power-conversion efficiency in server power-supply units (PSUs), while normally-on SiC JFETs and SiC Combo JFETs provide robust protection for high-voltage hot-swaps and eFuse/ORing applications.

This article explores the latest market trends and system requirements of AI data centers, along with how to leverage the technical advantages of

SiC JFET families in enabling efficient, reliable, and scalable AI data center infrastructure.

AI Data Center Market Trends and System Purpose

AI data centers are specialized facilities designed to support the intensive computational demands of AI workloads. These centers host thousands of servers equipped with GPUs, AI-specific ASICs, and high-speed networking components. The [International Energy Agency \(IEA\)](#) estimates that data centers consumed about 415 terawatt-hours (TWh) of electricity in 2024, accounting for around 1.5% of global electricity use. The consumption is expected to more than double to around 945 TWh, increasing to about a 3% share by 2030.

As AI models grow in complexity and size, the power requirements of these systems have surged. Current server-rack power levels range from 30 to 120 kW. However, leading computing companies are actively developing

CHAPTER 3: Achieve Megawatt-Scale AI with SiC JFET Technology

technologies to push this threshold to 1 MW, which is 10X the present capacity, driving the need for significant innovations in power conversion and distribution (Fig. 1). Traditional 48- to 54-V rack-level power distribution can no longer meet the needs of today's megawatt-scale AI infrastructure.

Most AI racks still rely on 48 V DC delivered via heavy copper bus-bars from power shelves to compute trays. However, once rack power surpasses 200 kW, 48-V DC distribution encounters critical limitations: It consumes excessive space — up to 64U per rack for power shelves in systems like NVIDIA's GB200/GB300 NVL72 — leaving no room for computing hardware.

It also demands unsustainable amounts of copper, with a 1-MW rack requiring over 200 kg, scaling to 500,000 tons for a 1-GW data center. In addition, repeated AC-DC and DC-DC conversions introduce significant energy losses and increase system complexity.

A more efficient alternative is to convert 13.8-kV AC directly to 800-V high-voltage DC (HVDC) at the data center edge by introducing a solid-state transformer (SST). This reduces conversion stages, minimizes energy loss, and simplifies the power chain. Meanwhile, fewer PSUs and fans also improve reliability, reduce thermal load, and enhance energy efficiency, making HVDC a scal-

able, high-performance solution for next-generation data centers.

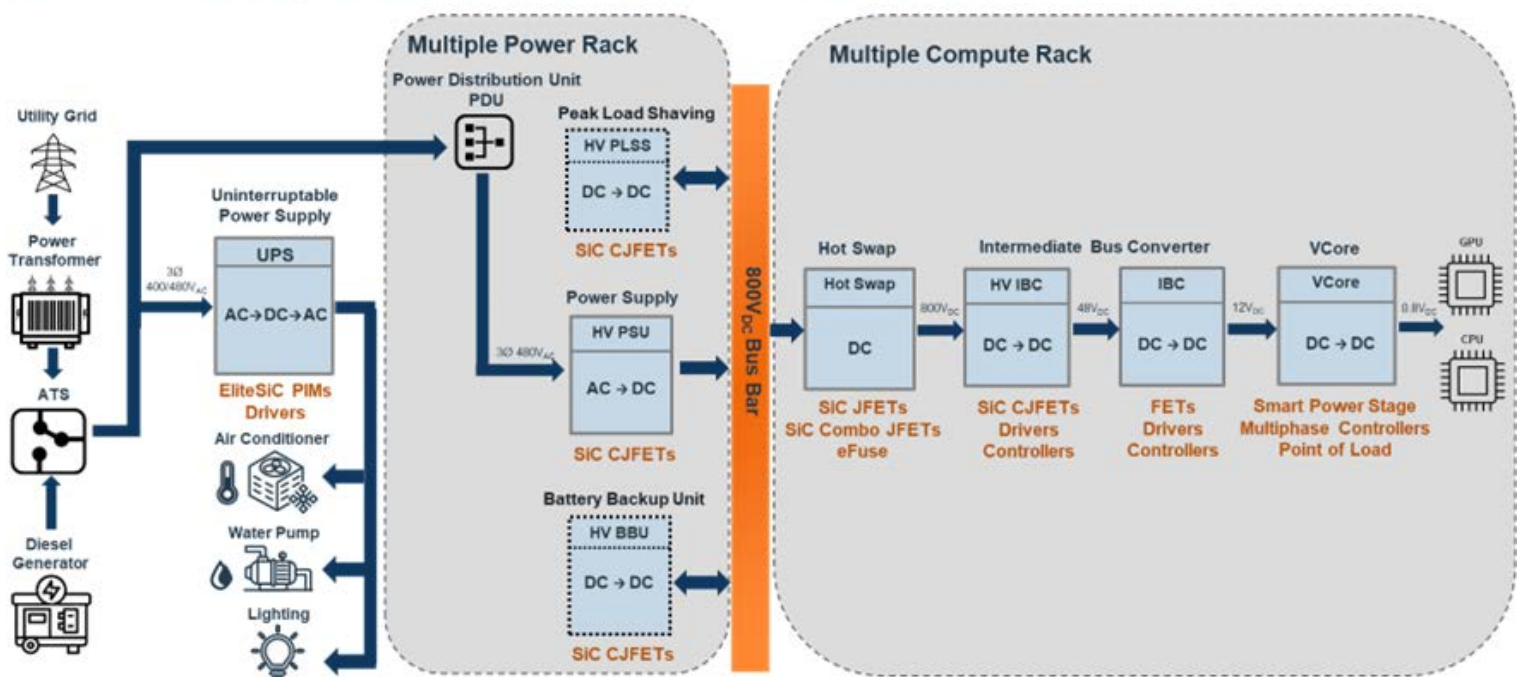
In summary, here are the key market trends in data center innovations:

- **High-voltage DC bus** (400 V DC and 800 V DC) to support dense server racks
- **Increased power density** to minimize footprint and cooling costs
- **Fast switching and low conduction loss** for energy efficiency
- **Robust protection mechanisms** to safeguard sensitive components during power events

Server PSU Requirements in AI Data Centers

AI data centers utilize advanced

Power Delivery from Grid to Processor in AI Data Center



1. (±)400-V DC / 800-V DC bus system architecture (up to 1 MW).

CHAPTER 3: Achieve Megawatt-Scale AI with SiC JFET Technology

server racks designed to support high-density computing, enhanced cooling solutions, and efficient power management. They often use high-density configurations (e.g., the PSU should meet or exceed the 80 Plus Titanium or even Ruby certification), with each rack potentially supporting up to 100 kW or more. The

total number of racks can range from a few dozen in smaller setups to several hundred in larger facilities. Open Rack V3 (ORV3) is the latest version of the Open Rack standard developed by the Open Compute Project ([OCP](#)).

Server PSUs in AI data centers must meet stringent requirements:

- **High efficiency** across wide load

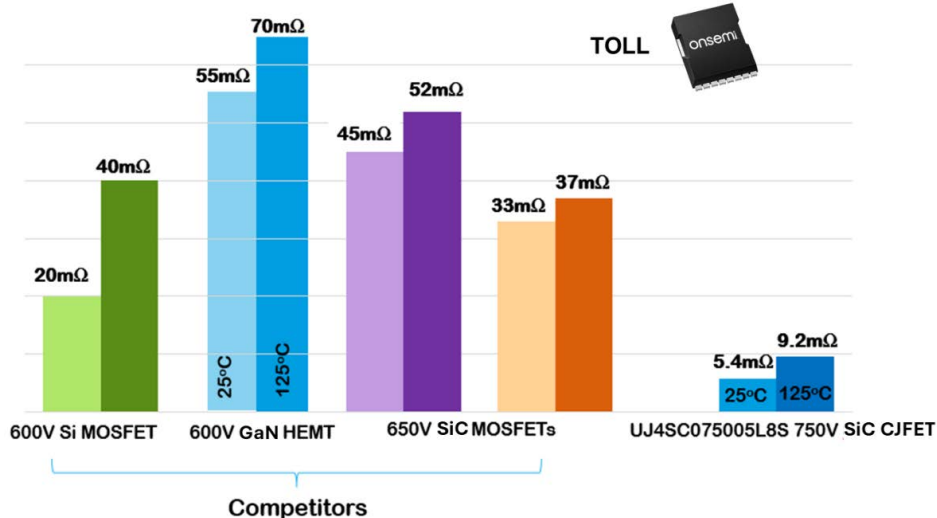
ranges

- **High power density** to reduce system costs
- **Compact form factors** for space-constrained environments
- **Compatibility with standard gate drivers** to simplify design

Traditional solutions using silicon superjunction (SJ) MOSFETs often fall short in meeting these demands due to limitations in switching speed, on-state resistance, and thermal performance.

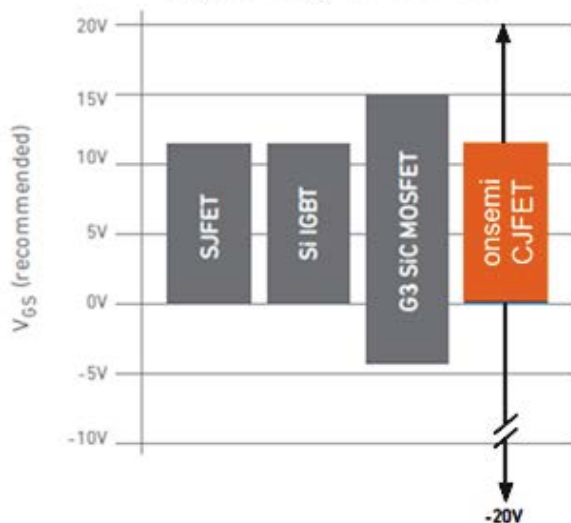
Optimizing Power Efficiency for Server PSUs: SiC Cascode JFETs

CJFETs, like those from [onsemi](#), integrate a normally-on SiC JFET with a low-voltage silicon MOSFET to create a normally-off device. This cascode configuration offers several advantages:

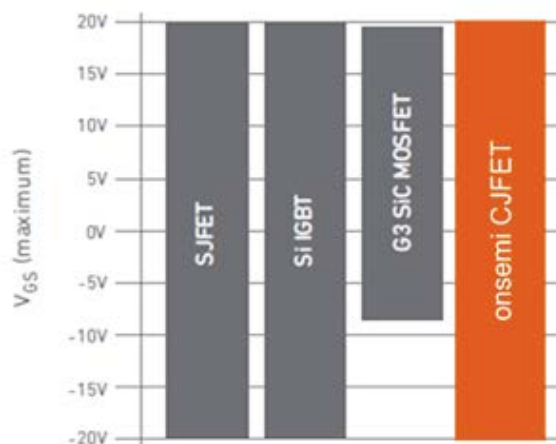


2. Lowest $R_{DS(on)} \times \text{area}$ compared to leading competitors.

12V/0V Operation simplifies upgrading to CJFET



Superior gate and ESD protection



3. Comparison of power-device technologies' gate-drive voltages

CHAPTER 3: Achieve Megawatt-Scale AI with SiC JFET Technology

- **Lowest $R_{DS(on)}$ per unit area:** Enables higher current handling and reduced conduction losses (*Fig. 2*).
- **Fast switching speeds:** Improves efficiency in both hard- and soft-switching topologies.
- **Gate-driver compatibility:** Works with standard silicon gate drivers, simplifying integration (*Fig. 3*).
- **Smaller die size:** Reduces cost and improves switching loss and conduction loss.

In an AI server rack, SiC CJFETs enable high-efficiency AC-DC and DC-DC conversion across key power subsystems, including the PSU, battery-backup unit (BBU), peak load shaving shelf (PLSS), and intermediate bus converters (IBCs). Their high efficiency and compact footprint support the demanding power profiles of AI workloads (*see table*).

Compared to SiC MOSFETs and silicon SJ MOSFETs:

- **SiC CJFETs offer significantly lower on-resistance per unit area**, delivering 10X improvement over silicon SJ devices and up to 50% improvement over SiC MOSFETs.
- **Switching losses are significantly reduced**, enabling higher-frequency operation and smaller passive components.
- **Thermal management is simplified**, reducing the need for bulky heatsinks.

COMPARATIVE ANALYSIS OF DIFFERENT DISCRETE TECHNOLOGIES			
Feature	SiC CJFET	SiC MOSFET	Si SJ MOSFET
$R_{DS(on)} \times \text{Area (Rsp)}$	Lowest	Moderate	High
Switching Speed	Fastest	Moderate	Slow
Thermal Performance	Excellent	Good	Moderate
Cost Efficiency	High	Moderate	Low

These benefits of SiC CJFETs translate into lower system costs, higher efficiency, and reliability for hyper-scale data centers.

Hot-Swap Function in AI Data Centers

In AI data centers, hot-swap circuits play a pivotal role in ensuring seamless power module insertion and removal without interrupting the system's operation. These functions are engineered to prevent high inrush currents during module engagement, deliver rapid fault response to safeguard sensitive electronics, and maintain compact, thermally optimized designs suitable for densely packed server environments.

Such a combination of robustness and efficiency is essential for maintaining reliability and uptime in high-performance computing infrastructures.

SiC JFETs and Combo JFETs: Enabling Robust HV Hot-swap Solutions

onsemi's 650- to 1,700-V SiC JFETs and SiC Combo JFETs integrate advanced features that make them

well-suited for high-voltage hot-swap and eFuse applications. Their ultra-low on-resistance ensures efficient power delivery with minimal thermal losses during conduction.

Furthermore, their superior thermal stability in linear-mode operation provides a safe current-limiting function. And controllable turn-off speed provides robust protection against inrush currents and short-circuit conditions, contributing to enhanced system reliability and operational safety.

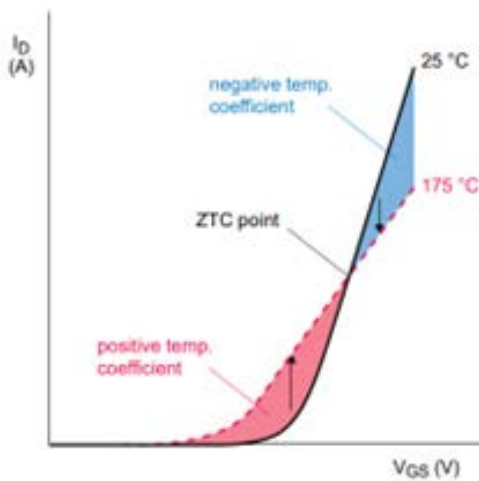
SiC JFET Key Features

- Normally-on devices with ultra-low $R_{DS(on)}$
- Overdrive with $V_{GS} = 2 \times V_{R_{DS(on)}}$ is 10% to 15% lower than $R_{DS(on)}$ with $V_{GS} = 0 \text{ V}$
- Wide SOA and thermal stability zone in linear mode for current-limiting applications
- V_{GS} in the on-state is a direct measurement of T_j , ideal for self-monitoring circuits
- Provide fast response and high thermal conductivity

SiC Combo JFET Additional Key Features

- Integrate a SiC JFET and a silicon MOSFET in a single package

CHAPTER 3: Achieve Megawatt-Scale AI with SiC JFET Technology



4. A SiC JFET has a very small thermal instability zone in linear mode.

- Deliver a normally-off switch compatible with gate drive for Si/SiC MOSFETs
- Enable switching speed control through adjusting JFET gate resistance
- Enable paralleling of multiple JFETs

Superior Linear-Mode Thermal Stability in Hot-swap

In hot-swap circuits, power devices often operate in linear mode to control inrush current and ensure a smooth, safe startup under high capacitive loads. The adoption of 400- and 800-V HVDC busbar architectures in AI data center systems has significantly elevated the importance of SiC-based hot-swap solutions. In these high-voltage environments, SiC Combo JFETs and SiC JFETs enhance system safety and operational efficiency by delivering the following key functions:

- Prevent inrush currents during

module insertion.

- Protect against short circuits and overcurrent events.
- Enable compact, efficient designs with fewer components.

Unlike conventional silicon or SiC MOSFETs, SiC JFETs exhibit a much wider thermal stability zone above the zero temperature coefficient (ZTC) point with a negative temperature coefficient of current. As a result, they can conduct more current at elevated junction temperatures without entering thermal runaway. This behavior ensures predictable performance and enhances system reliability under demanding thermal conditions (Fig. 4).

Moreover, the gate-source voltage (V_{GS}) of a JFET in the on-state serves as a direct indicator of junction temperature, enabling real-time thermal monitoring and protection as a temperature sensor. SiC Combo JFETs further integrate a low-voltage Si MOSFET with a normally-on SiC JFET in a compact package, combining thermal robustness with simplified gate-drive (normally-off) and control flexibility.

Given these advantages, SiC Combo JFETs and SiC JFETs are particularly well-suited for linear-mode applications such as hot-swap, eFuse, and ORing, where inrush current limiting, overcurrent protection, and thermal management are critical to system reliability.

Conclusion

As AI data centers continue to evolve, the need for high-performance, energy-efficient power solutions is rapidly growing. onsemi's [EliteSiC](#) portfolio of 650- to 1,700-V SiC CJFETs, SiC JFETs, and SiC Combo JFETs offers comprehensive solutions for server PSUs and hot-swap/ORing applications in AI data centers. These devices combine high efficiency, thermal performance, and design flexibility, enabling data centers to meet AI workloads.

For onsemi's detailed technical insights and complete set of recommended solutions for AI data centers, visit onsemi.com and download the [AI Data Center System Solution Guide](#).

[view this article online](#)

[BACK TO TABLE OF CONTENTS](#)

CHAPTER 4

Edge AI Does Much More for Manufacturers Beyond Production-Line Automation

By deploying intelligent solutions in manufacturing environments, companies can also improve operations beyond the manufacturing line.

MATTHIAS HUBER, Sr. Director, Solutions Manager, IoT and Embedded and Edge Computing, *Supermicro*



dreamstime_Sompong-Sriphet_126723440.jpg

The world's attention is being grabbed by artificial-intelligence (AI) technology becoming pervasive in consumer electronics, but that's only half of the edge AI story. Edge AI can also be scaled up to be as computationally powerful as needed.

Powerful edge AI servers have become indispensable tools in the manufacturing sector, improving production processes, increasing manufacturing line efficiency, providing

superior inspection, and monitoring of products, plus more.

Advantages of Edge AI

All AI models require some training. The process is particularly elaborate for the most comprehensive models, which require repetitively running through enormous datasets until they can reliably produce expected results in response to queries. Training complex models such as large language models (LLMs) and vision language

models (VLMs) almost always take the formidable processing capabilities and memory resources resident only in giant data centers.

After being trained, most of the work that AIs perform is called inference. This includes parsing new inputs in the context of what the AI has already learned. Inference doesn't need quite the processing power of training. Improving the capabilities of an application or adding features typically requires AI models that are more complex, though. As products become more sophisticated, some inference workloads turn quite compute-intensive.

Many of the innovations in AI in the last few years have involved getting processors to run inference workloads far more efficiently in terms of both processing power and the energy consumed. This has made it eminently practical to run more AI workloads — and more sophisticated workloads — in servers located at the network edge, and even in smaller consumer electronics such as smart speakers and nanny cams.

CHAPTER 4: Edge AI Does Much More for Manufacturers Beyond Production-Line Automation

Edge AI is faster and more efficient than relying on the cloud. “Edge computing provides notable energy efficiency benefits compared to large data centers by reducing data transport, lowering latency, and enabling dynamic resource allocation,” according to “[Edge AI for Internet of Energy: Challenges and Perspectives](#),” published in the journal *Engineering Cyber Physical Human Systems Internet of Things; Engineering Cyber Physical Human Systems*.

Latency can be significantly lowered by not sending data on a round trip to and from some distant computing center, which results in achieving fast-

er results. For example, GE Healthcare reported notably faster results after embedding AI in its X-ray systems.

Edge AI can be more efficient in several ways. It minimizes network bandwidth usage, again by not shipping data to and from a data center. This can reduce telecom service bills, and the energy not expended on data transmission is saved.

Generally speaking, edge AI applications tend to run smaller, more specialized AI models, which means an AI processor at the edge is likely to consume less power than a server running AI workloads in a data center.

There are applications at the edge, including some in manufacturing, that don’t need to run continuously. In these instances, edge AI can operate in standby and/or sleep modes to minimize energy consumption. While engineers must balance the factors involved in any specific application, it’s possible to achieve significant efficiencies with edge AI.

Finally, data generated and used at the edge stays at the edge, serving the interests of both data security and data privacy. This is often a decisive consideration for companies seeking to protect their intellectual property (IP), or companies operating with



1. Machine-vision inspection systems can be backed by edge AI to improve quality control of production lines.

CHAPTER 4: Edge AI Does Much More for Manufacturers Beyond Production-Line Automation

legal obligations to protect consumer information, e.g., medical facilities.

Edge AI in Manufacturing

Edge AI is a powerful means for meeting a variety of manufacturers' needs and improving a range of their business operations and practices. Edge AI can perform data analysis in real-time, which facilitates rapid decision-making, enhances operational agility, and streamlines workflows, ultimately driving additional efficiency gains, cost savings, and improved product quality. By analyzing real-time data, AIs are able to identify conditions and trends. Manufacturers can enable AI systems to trigger automatic responses in production lines, trip safety alarms, and more.

Machine vision has become integral for any number of manufacturing applications, from quality assurance to process optimization to waste reduction. Production lines vary: Some might inspect only a sample of parts for defects, perhaps only 1 out of every 20, or 1 out of every 5. The choice not to inspect every part can sometimes depend on limitations inherent in the inspection of equipment.

However, by using computer vision backed by edge AI, manufacturers can inspect every part coming off the production line. Even in instances where every part is already being inspected, AI-based systems have proven to be better at detecting defects

Machine-vision systems backed by

AI quickly identify defects (**Fig. 1**). The AI will immediately evaluate the number and severity of defects that appear and can be programmed to recommend taking remedial action to minimize waste, or to automatically take action.

In fact, it's possible for machine-vision systems backed by edge AI to analyze variations in product quality that indicate an incipient problem. In other words, AI can be trained to identify variations in quality and alert line managers to take remedial action before the appearance of actual defects. All of this ultimately leads to minimized production line downtime.

By some estimates, leveraging AI in manufacturing may reduce machine downtime by 30% to 50%, and quality-related costs can be lowered by 10% to 20%.

Such predictive capabilities could also be profitably used in the maintenance of production equipment. Without AI, manufacturers must schedule regular maintenance checks. If the maintenance check reveals no problems, then the scheduled shutdown has wasted production time.

Meanwhile, if maintenance is required between scheduled checks, without AI monitoring, the need may go unidentified, leading to downtime and possibly expensive equipment damage. AI is singularly capable of continuously monitoring equipment and learning what performance anomalies are indicative of pending equip-

ment failure. Edge AIs can be relied on to recommend maintenance checks when necessary, and only if necessary. This minimizes costly downtime and increases operational reliability and efficiency.

One of the manufacturing industry's most important and expensive challenges is workplace safety. Edge AI is invaluable for analyzing data from security cameras to identify potentially dangerous behaviors or situations. Companies also use edge AI to connect data feeds from CCTV, machinery, and operational software to provide real-time predictive analysis of potential hazards. AI models could, for example, identify when an employee isn't wearing protective gear or is moving too close to dangerous machinery.

Furthermore, as autonomous mobile systems are increasingly being introduced in workspaces, AI will be instrumental in avoiding hazardous situations. Real-time processing of such large amounts of data is practical only with edge AI. Since most manufacturing facilities already employ video monitoring systems, this safety boost is a matter of utilizing that data most effectively.

Optimized AI Solutions

There's a thriving market for edge AI servers. Manufacturers should be careful to evaluate servers engineered for long-term reliability. Companies commonly maintain the computing

CHAPTER 4: Edge AI Does Much More for Manufacturers Beyond Production-Line Automation

2. [Supermicro's SYS-212B-FN2T](#) is an edge server with a short-depth form factor. This 2 RU model features an Intel Xeon series processor and can be configured with up to two double-width NVIDIA GPUs, such as the NVIDIA RTX PRO 6000 Max-Q Workstation Edition or L40S.



infrastructure in manufacturing operations for far longer than, say, desktop computers. While processing capabilities and operational efficiencies are key factors when considering the installation of edge AI systems, there are others.

One is the mean time between failure (MTBF), which speaks to system reliability. It's also important to evaluate if the physical environment where the edge AI server(s) will be installed requires additional ruggedization. This might include the ability to operate across wider temperature ranges, or the ability to withstand shock and vibration. Fanless servers are indicated for factories and other work environments where dust or other particulate matter is commonly present.

The main challenge for manufac-

turers is specifying a configuration of edge AI infrastructure that's optimized for the mix of workloads that must be handled. Manufacturers can now choose from among a wide selection of standard servers integrating any of a range of processors and AI accelerators to perform AI workloads.

Some of the more established suppliers of edge AI servers have modular architectures, which give manufacturers the flexibility to build semi-custom edge AI server systems to fulfill specific requirements. Such requirements can include anything from processing capabilities to level of ruggedization to the physical dimensions of the system being installed. Importantly, a modular approach also leaves open paths for upgradability over the course of years.

Manufacturers planning to develop

their own applications should also inquire about the availability of software development kits (SDKs). A useful SDK will already contain modules that manufacturers can use so that they don't have to build applications from scratch. Examples include applications that can visually inspect products and need only to be modified to monitor printed circuit boards, the seals on containers, or anything else. This makes AI development easier and reduces time-to-market.

For instance, [Supermicro](#) has a modular approach to building servers that's resulted in a very broad portfolio, including a wide selection of standard products for both the cloud and the network edge (**Fig. 2**). This modular approach also makes it easier for the company to devise optimized

CHAPTER 4: Edge AI Does Much More for Manufacturers Beyond Production-Line Automation

solutions. Supermicro's edge AI solutions are able to bring the capabilities of AI accelerators to ruggedized, remote environments such as factory floors, where they can be integrated into key business processes.

Successfully deploying and running AI applications at the edge often requires specialized hardware. AI accelerators are the workhorses in this model, processing large amounts of data based on a pre-trained model.

NVIDIA is the global market leader for these accelerators, with a broad range of AI-optimized CPU and GPU modules for workloads of any type and scale. The NVIDIA Jetson Orin NX platform for the embedded edge, NVIDIA RTX PRO 6000 Blackwell Server Edition, and NVIDIA H200 NVL GPUs for enterprise edge all have distinctive use cases based on the type and volume of data being processed, as well as deployment factors such as size and power consumption. Selecting the right hardware, therefore, includes identifying which AI accelerator is required and ensuring that the server platform is compatible.

Developing AI applications can be a daunting and costly process. That's where a platform like [NVIDIA Metropolis](#) comes in. This end-to-end platform helps businesses fast-track the creation and deployment of intelligent video analytics by offering pre-trained models, SDKs, and optimized infrastructure. For factories, this means faster rollout of applications

that save costs, improve production throughput, and make factories safer, without starting from scratch.

By simplifying AI development and reducing the time-to-market, NVIDIA Metropolis helps company's lower costs while unlocking powerful insights from their video data, turning everyday surveillance into a strategic business asset.

[view this article online](#)

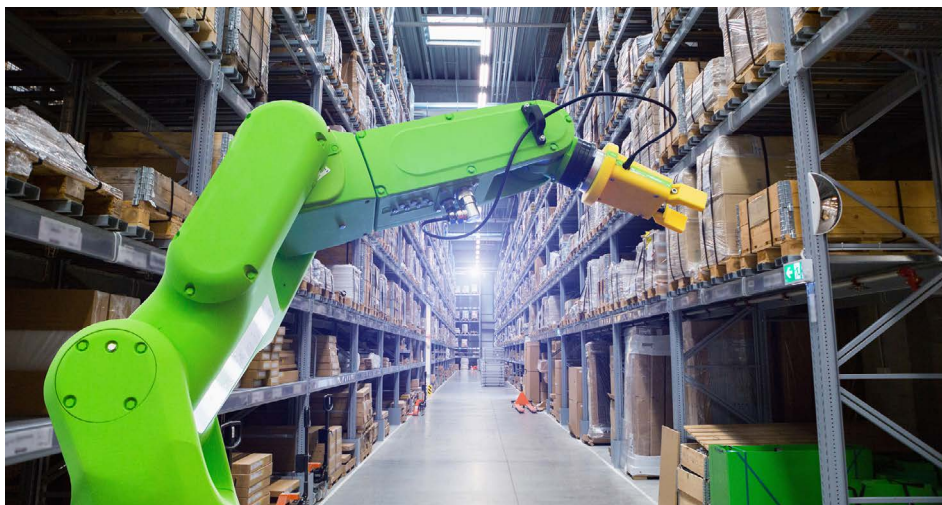
 [BACK TO TABLE OF CONTENTS](#)

CHAPTER 5

Why Sensor Fusion is the Real Differentiator in Physical AI

It's no longer enough to simply bolt AI onto an existing mechanical form — now the shape, movements, operation, and senses of modern robots are built around AI.

MADHURA TAPSE, Senior Product Marketing Manager, NXP Semiconductors



dreamstime_ Scharfsinn86_115451269.jpg

AI is breaking free from the chat box. As physical AI increasingly becomes a part of our workplaces and the world around us, it's driving an evolution in robotics. It's no longer enough to simply bolt AI onto an existing mechanical form. Instead, the equation is flipping: the shape, movements, operation, and, most importantly, senses of modern robots are built around AI.

The importance of sensing in the era of physical AI is driven by a simple fact: A robot that can't feel will eventually fail. And feeling doesn't take place

in a large language model (LLM) or a GPU in a data center. It takes place in the real world, at the edge.

No matter how powerful a GPU or advanced an AI model is, they can't underpin effective robotics without trustworthy data on the world around them. In other words, physical AI needs real-world inputs — from temperature and pressure to voltage, weight, and acoustics. Providing that additional sensory layer means that there must be an evolution in how we design and build robots.

Creating a New Architecture for Physical AI

Physical AI calls for a fundamentally different architecture from the robots of yesterday, one split into multiple different layers. Multimodal sensors act as the nervous system and reflex layer. These give the robot awareness of everything, including force, sound, and temperature.

High-quality data capture is essential here, because such data is fed into the perception and AI layer as raw material. Raw inputs are transformed into meaningful information that can be used for features like object detection, pose estimation (calculating the position of the robot in space), and more.

All of this sensory input is captured before the robot acts. Movement, based on guidance from the AI that's grounded in trusted sensory data, is controlled by a separate layer. At this point, high-performance compute handles the planning and motion goals, which are then sent to real-time motor controllers that ensure every

CHAPTER 5: Why Sensor Fusion is the Real Differentiator in Physical AI

movement is precise.

Finally, edge gateways and cloud services capture multi-sensor trends, using them to predict maintenance and monitor use — all of which can be used to optimize performance. In other words, in physical AI architecture, multimodal sensor fusion is vital at both the start and the end of the loop.

Why Sensor Fusion Matters Now

Crucially, in effective physical AI architecture, sensors don't work in isolation. Sensor fusion — the combining of data from multiple sensors — is used to create a clearer picture of the real world. It's rapidly replacing legacy systems that rely on single sensor monitoring.

Picture a large robotic arm working in the automotive industry. Deep within its joint, a tiny 2- to 3-mA current offset pairs with a minor thermal change in torque; a minuscule hint that the bearings are starting to wear. With the right sensor fusion that combines temperature and torque, those tiny warning signs can be caught early by an AI. As a result, an engineer repairs the arm, and it keeps operating without failure.

The same logic applies beyond automotive arms. Moving from single sensors to multi-sensor fusion creates dramatic improvements in autonomy and reliability across different robotics use cases. By fusing temperature,

current/voltage, position, and other sensory input, anomaly detection improves and unexpected emergencies are reduced.

This only becomes more important as new forms of robotics, e.g., humanoids, enter the workforce. For humanoid platforms to operate effectively, internal measurements like those focused on temperature and torque are fused with environmental sensing of the world around them.

The Importance of Precision

Achieving reliable sensor fusion requires more than simply connecting multiple sensors to an AI processor. The analog signal chain must preserve information across a wide range of sensor outputs, from microvolt-level bridge sensors used in force and torque measurements to industrial voltage and current signals employed for system monitoring.

High-performance analog front ends (AFE) are essential here. They integrate precision excitation sources, programmable gain amplification, low-drift references, and high-resolution data conversion into a common measurement platform, ensuring consistent accuracy across different sensing modalities.

Devices like [NXP's NAFE43388](#) do just this, enabling a single hardware platform to acquire force, temperature, current, voltage, and resistance-based measurements. Such an architecture allows subtle correlations

between mechanical, thermal, and electrical parameters to be detected with confidence, helping AI models identify early indicators of wear, efficiency degradation, or abnormal operating conditions before they develop into system-level failures.

With the right hardware, fusion becomes practical at scale, leading to efficiency and safety gains across fleets of robots. That might mean assembly robots or humanoid robots using vibration and motor currents to detect bearing wear. Or in another use case, autonomous mobile robots (AMRs) could capture wheel motor currents, encoder drift (the accumulation of small errors over time), and thermal signatures to build a live health profile across a whole fleet.

[view this article online](#)

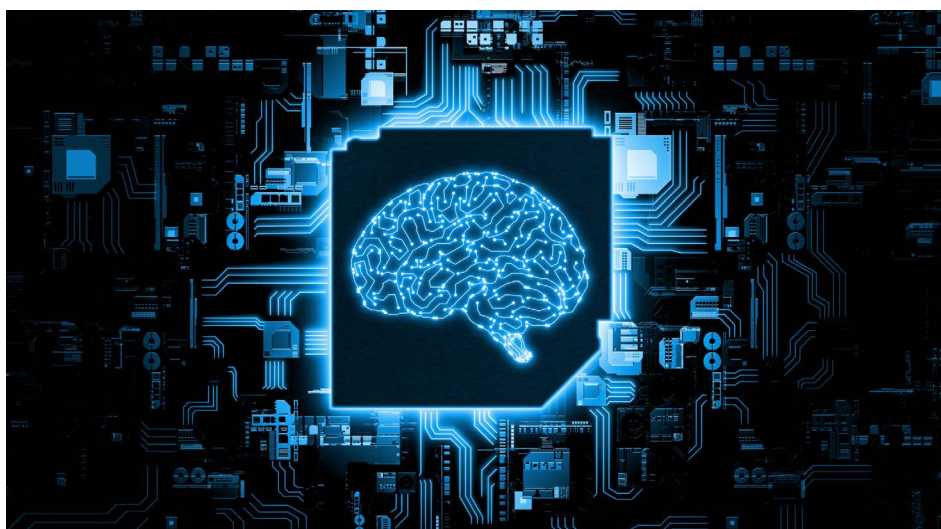
 [BACK TO TABLE OF CONTENTS](#)

CHAPTER 6

Performance Scaling — with Balance — in Edge AI

Despite the different computational principles at the heart of neural processing, the underlying hardware behavior is the same as that of traditional systems.

EDWIN HILKENS, Product Marketing Manager, and François de Rochebouët, Head of AI Solution Marketing, STMicroelectronics



dreamstime_kkssr_142816692.jpg

As microcontrollers leverage embedded neural processing units (NPUs) to accelerate AI inference at the edge, data-movement overheads present a barrier to further progress. In-memory compute is an effective accelerator that demands close engagement with the underlying technology.

Accelerating AI at the Edge

Edge AI has reached the plateau of productivity in Gartner's hype cycle, with influential new products reaching the market to transform life and work. The edge has been described as the place where physical signals become data. AI now brings unprec-

edented ability to interpret observations instantly and determine accurate and timely responses.

Bringing neural processing to edge platforms has allowed product developers to implement capabilities like efficient computer vision, keyword detection, pattern matching, anomaly detection, and activity tracking in resource-constrained microcontroller-based systems. It's driven powerful changes in equipment such as smart protective clothing, industrial safety systems, predictive-maintenance sensors, quality-inspection systems, and many others.

Today's quest to embed AI skills in

tiny edge platforms began by adapting neural workloads to fit the constraints of ordinary embedded CPUs like Cortex-M cores. Processors like these can handle repeated multiply-accumulate patterns and memory accesses, although they're clearly not well adapted for such tasks.

With quantization to reduce data footprint and movement overheads, and with carefully optimized handling of operators and memory layout, embedded platforms can host basic AI like 2D convolutions, keyword spotting, and anomaly detection. Toolchains began to adapt accordingly: For example, STMicroelectronics introduced the AI Studio as part of the STM32Cube ecosystem to help developers select and tune suitable models to run on their selected hardware.

However, any scope for handling more powerful and complex workloads this way is severely limited, especially within the tight size, power, and cost constraints that face embedded platforms. Practically, measurements show the maximum possible compute efficiency to be only about 0.1 teraoperations per second per watt (TOPS/W).

CHAPTER 6: Performance Scaling — with Balance — in Edge AI

Embedded Neural Processing

To accelerate inference and deliver more TOPS/W for AI workloads, the leading MCUs now integrate a purpose-designed NPU in dedicated hardware.

ST's Neural-ART Accelerator (**Fig. 1**) is one example of the embedded NPUs emerging to drive higher-performing edge AI, built around

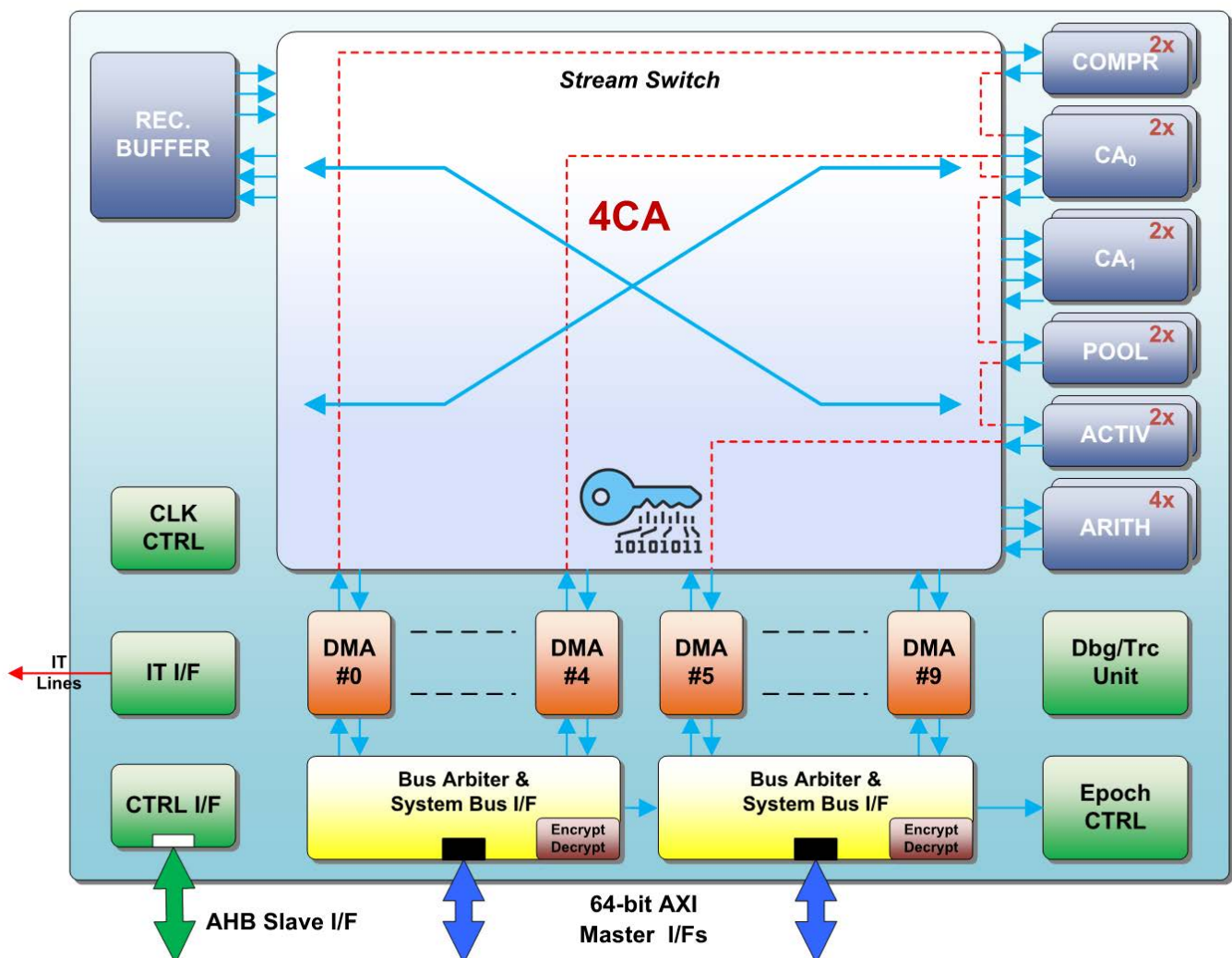
a stream-processing non-blocking switch fabric with hardware including dataflow control and routing. Also in the mix are dedicated AI accelerators, DMA engines, and interrupt generators, with host-decoupled execution.

The NPU architecture is designed to enable scaling from a few GOPS to multiple TOPS, and it allows multi-island configurations for extensive optimization, including voltage

and frequency scaling to optimize power consumption. It also permits multi-model computing that assigns models such as movement recognition or low-demand networking tasks flexibly according to workload and criticality.

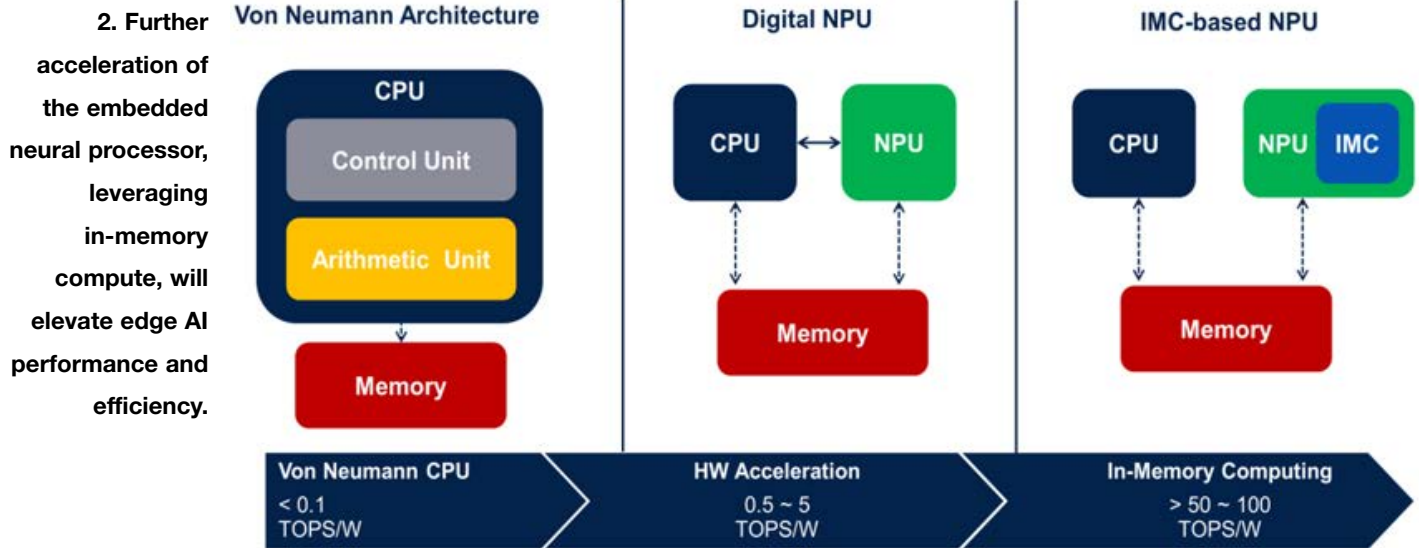
Although more powerful, and well adapted to tiny edge AI, NPUs integrated in microcontrollers must work with extremely limited resources.

Gen1 0.6 TOPs 16FF



1. The Neural-ART Accelerator NPU architected for dataflow processing.

CHAPTER 6: Performance Scaling — with Balance — in Edge AI



Maximizing local data reuse is a priority and maintaining a balance between compute performance, routing, and interface bandwidth is vital. A deficiency in any one of these, relative to the others, creates a bottleneck that limits system performance.

Based on currently available 40-, 28-, and 18-nm FDSOI, and 16-nm FinFET fabrication technologies, neural processing benchmarks are projected to reach 1 to 5 TOPS/W and 0.1 to 2 TOPS/mm² with NPU acceleration alone. These metrics also depend on NPU architecture and application operating modes.

However, anticipated market demands call for performance and efficiency rising to 50 to 200 TOPS/W over the next five years. Scaling NPU performance requires adding hardware, though, such as extra processing blocks and broader interconnects that increase silicon area and power. A different approach is needed.

Embedded In-Memory Compute

Despite the different computational principles at the heart of neural processing, the underlying hardware behavior is the same as that of traditional systems. Although the NPU establishes parallelism through many small sequential units, memory remains passive and executing instructions in the processing unit requires bringing data across an interface of strictly limited bandwidth.

In-memory compute (IMC) is a promising technique in high-performance computing and is extremely well-suited to linear operations like vector multiplications in neural processing. Developing IMC for edge platforms can potentially deliver the large gains in power efficiency and computing density in terms of TOPS/W and TOPS/mm² of silicon needed to meet future expectations for edge AI performance (Fig. 2).

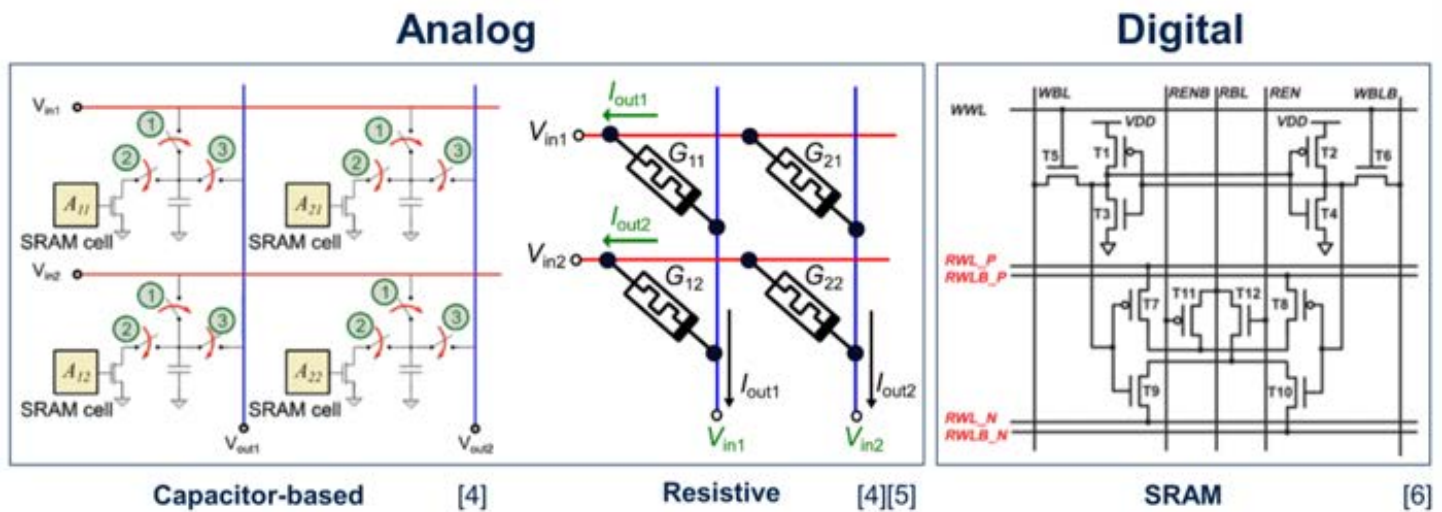
There's more than one way to realize IMC and the technique is at the beginning of its development cycle. Digital IMC (DIMC) implements multipliers inside memory. In analog IMC, the network weights are encoded in memory cell states and multiplications are performed in situ based on analog phenomena such as Kirchoff's law or Ohm's law.

Digital IMC

Digital IMC is typically implemented with SRAM arrays, holding weights, or small tensors that are tightly coupled with fixed-function logic such as MAC units fabricated locally. DIMC tiles are the internal combined temporary storage and computing workhorses of this new type of NPU. The compiler handles assigning operations to the NPU and its internal IMC block orchestrates data movement between them.

The internal DIMC tiles are small,

CHAPTER 6: Performance Scaling — with Balance — in Edge AI



3. Analog IMC generates a dot product from the memory's capacitive or resistive elements. Digital IMC implements multipliers in the memory.

and weights are reloaded for each computation. By making it possible to handle arbitrary complex neural networks like transformers and recurrent neural networks, or RNNs (including GRUs, LSTMs), as well as enhancing the performance of convolutional neural networks (CNNs) typically used for computer vision, DIMC can extend edge AI use cases to encompass new opportunities such as time-series analysis.

However, digital IMC still calls for moving data and is subject to memory bandwidth limits and clock-related throughput limitations. The area cost of digital multipliers must also be considered.

Analog IMC

Analog IMC is appropriate to SRAM or embedded non-volatile memory (eNVM), including charge-based flash or resistance-based types such

as phase-change (PCM) and ferro-electric (FRAM) memories (Fig. 3). Weights are stored in the charge state or conductance state of the cells; outputs are summed as currents; and a matrix-vector multiply can be performed in one shot when the cells are activated.

As this is analog electronics, signals need time to settle and the results can be affected by factors such as noise and temperature drift, which requires compensation.

IMC eliminates the memory-interface bottleneck and delivers parallelism without area cost. In analog IMC, the memory array is effectively the compute unit — no digital engine is needed. Reducing compute energy by orders of magnitude compared to conventional processor-memory interaction makes power efficiency a strong point, and high-density multi-level storage is possible. Meanwhile, tile

size determines application-mapping efficiency as larger tile sizes can boost TOPS/W at the expense of lower area utilization. Though digital IMC is less dense, very large networks can be handled as weights and quickly reloaded in SRAM.

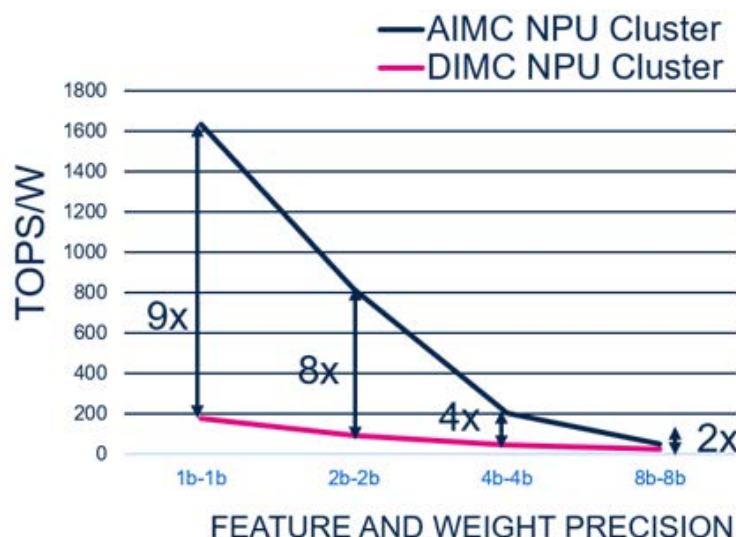
Each technique has its strengths, and committing to develop both gives edge-AI developers the flexibility they need to create properly optimized solutions. Figure 4 compares the acceleration, efficiency, and precision of digital and analog IMC.

Performance Enhancement Potential

Application developers and chip vendors understand that software alone, running on conventional MCU architectures or even MCUs enhanced with neural accelerators, can't deliver the performance progression needed to meet future edge AI processing

CHAPTER 6: Performance Scaling — with Balance — in Edge AI

5. IMC can deliver the acceleration needed to meet future edge AI performance demands.



Conclusion

The embedded NPU is already becoming an essential tool driving the performance of edge AI forward to meet the needs of future applications. However, further acceleration is needed, beyond the potential of today's NPUs alone. In addition to raw compute speed, both energy efficiency, in TOPS/W, and area efficiency in TOPS/mm² are the dominant performance metrics.

requirements.

Digital IMC and, ultimately, analog IMC that relieves both data movement and the propagation delay associated with digital processing, show the way forward for neural processing on resource-constrained embedded MCUs. Ultimately, edge AI energy efficiency could improve up to 100X (Fig. 5).

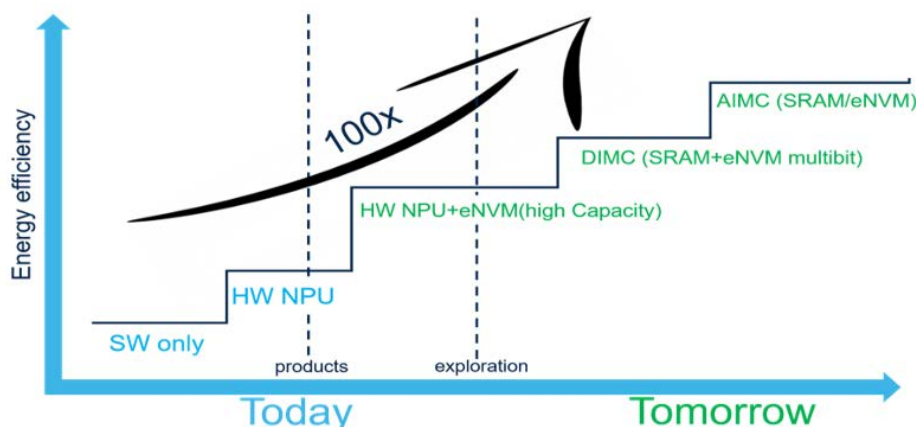
IMC design is implicitly connected with array size and shape, memory technology, routing, and hardware placement. By effectively migrating

compute into device physics, IMC connects with the silicon layout in a way unlike previous hardware accelerators. The silicon defines the computation. The compiler can handle mapping the model onto the available fabric and orchestrating the available resources.

Viewed from this standpoint, ST's proprietary Neural-ART Accelerator is an example of the process-level control and flexibility needed to embed effective IMC acceleration in future generations of edge AI microcontrollers.

IMC overcomes memory bandwidth limits by performing the computation inside the array at the location of the data, and as the data is accessed. Only the computation result leaves the array.

Both digital and analog IMC are viable and effective strategies to continue accelerating neural processing at the edge. For MCU vendors, control over the NPU architecture as well as the fabrication technology is critical to realize these accelerators and deliver the tools that developers will need to fulfill their potential.



[view this article online](#)

[BACK TO TABLE OF CONTENTS](#)

4. Digital and analog IMC in an embedded NPU.

CHAPTER 7

Evolving Antenna Requirements for LEO and MEO

Explore how LEO and MEO constellations are driving new approaches to antenna design, from phased arrays and hybrid beamforming to distributed digital architectures.

IAN LANKSHEAR, CEO, EnSilica



dreamstime_Andreyi Armiagov_43333060.jpg

Satellite communication is changing quickly as large constellations are deployed to deliver global broadband. In low-Earth-orbit (LEO) and medium-Earth-orbit (MEO) systems, antennas must track fast-moving satellites, handover often, and cope with a much tougher interference environment than GEO links ever faced.

Unlike geostationary-Earth-orbit (GEO) satellites, which appear fixed relative to the Earth, LEO satellites move rapidly across the sky at alti-

tudes typically between 500 km and 1,200 km. A satellite may cross the visible sky in a few minutes, with apparent angular velocity approaching one degree per second near zenith, so antennas must hold accurate pointing through repeated transitions from one satellite to the next.

Constellation density also matters. Some systems put several satellites in view at once; others force the terminal to search a much larger area of sky to acquire one. As scan angles widen, designers face scan loss, beam distor-

tion, and higher sidelobes, while interference grows harder to ignore where satellites sit only a few degrees apart or where constellations share spectrum. Narrow beams and low sidelobes are essential, not merely desirable.

Reflector Antennas and Array-Fed Designs

For decades, parabolic reflectors have been the standard antenna for satellite communications. Peak gain approximates to $G \approx \eta (\pi D / \lambda)^2$ and half-power beamwidth to $\theta_{BW} \approx 70 \lambda / D$, where D is the diameter, λ the wavelength, and η the aperture efficiency. That's why gateway earth stations still use dishes several meters across.

However, mechanical steering, adequate for geostationary links and gateways working a few satellites, limits beam agility and adds maintenance, making it a poor fit for user terminals following many LEO satellites. Array-fed reflectors, in which a phased-array feed illuminates the dish, offer a middle path where high throughput and agile steering are both needed.

CHAPTER 7: Evolving Antenna Requirements for LEO and MEO

Phased Arrays and Hybrid Beamforming Architectures

The stronger momentum, though, is behind electronically steered phased arrays. By controlling phase and amplitude across many elements, they steer beams rapidly with no moving parts, supporting fast acquisition, agile tracking, simultaneous multi-beam operation, and spatial processing that actively suppresses interference.

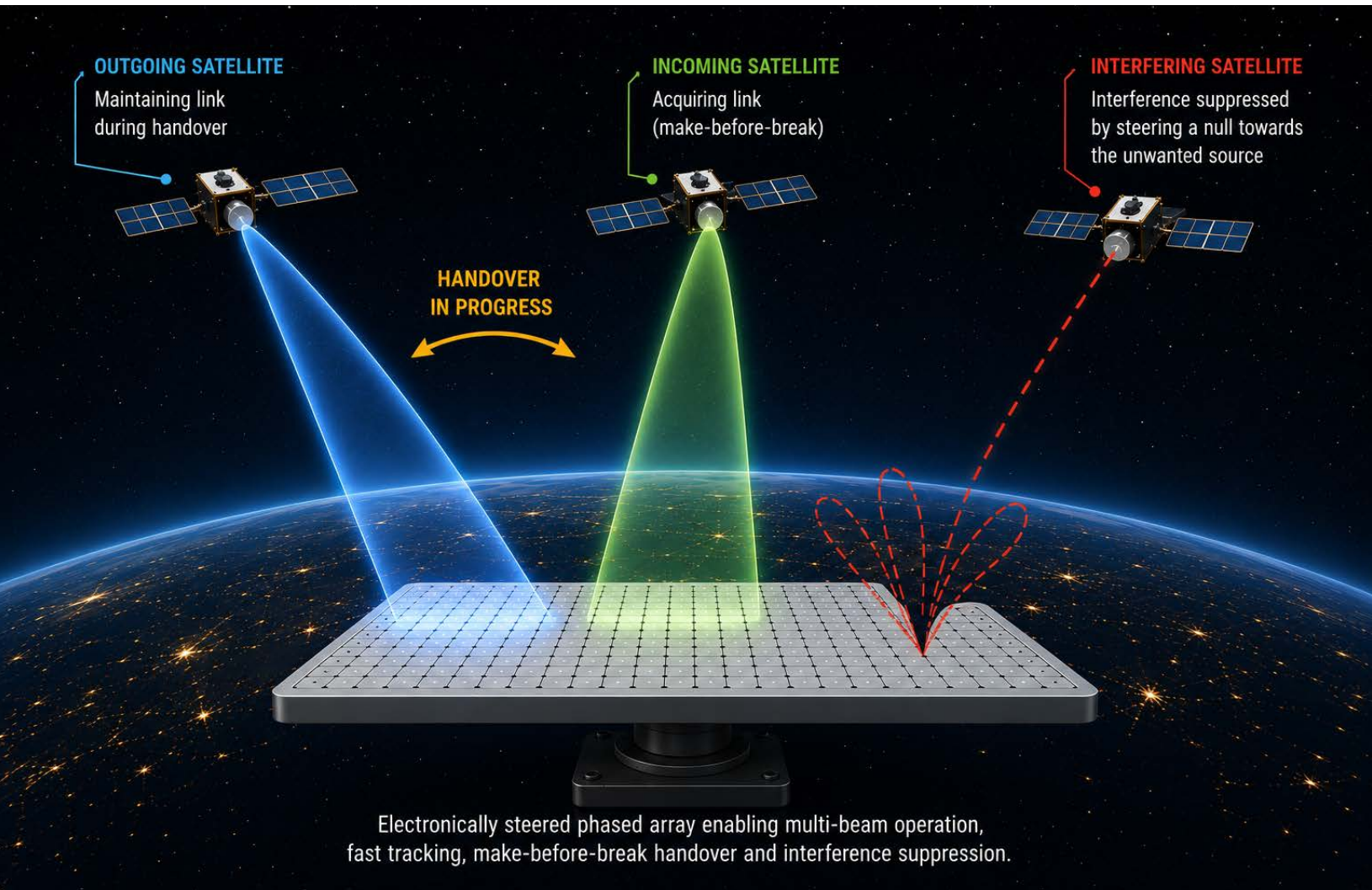
For an array with element spacing

d , the phase shift between adjacent elements needed to steer the beam to an angle θ from broadside is $\varphi = (2 \pi d / \lambda) \sin(\theta)$, and the resulting pattern follows the array factor $AF(\theta) = \sum_{n=0}^{N-1} e^{j n \varphi}$, where N is the number of antenna elements. The theory is well understood; scaling it efficiently is the challenge.

As arrays grow to hundreds or thousands of elements, complexity rises quickly. Many RF channels, high-

speed data converters, and substantial digital signal processing are needed, and power, calibration, silicon area, thermal management, and internal bandwidth begin to dominate implementation choices. That's why satcom designs usually mix analog and digital beamforming.

Analog beamforming phase-shifts and combines in the RF domain before digitization. It is low power and simple but generally gives only one beam per



1. Multi-beam operation supports make-before-break handover and spatial interference suppression.

CHAPTER 7: Evolving Antenna Requirements for LEO and MEO

RF chain and little spatial flexibility. Fully digital beamforming digitizes every element independently – the most flexible option for multi-beam operation and adaptive interference suppression. However, it scales poorly because RF chains and ADCs grow with element count.

Hybrid beamforming is usually the practical compromise: Elements are grouped into subarrays, combined locally in analog, then digitally beamformed at the subarray outputs. This cuts the number of digital channels

while retaining most of the benefits of digital control – a 1,024-element array using 16-element subarrays needs only 64. The ratio of elements to digital channels (the hybrid ratio) sets flexibility, complexity, and scalability.

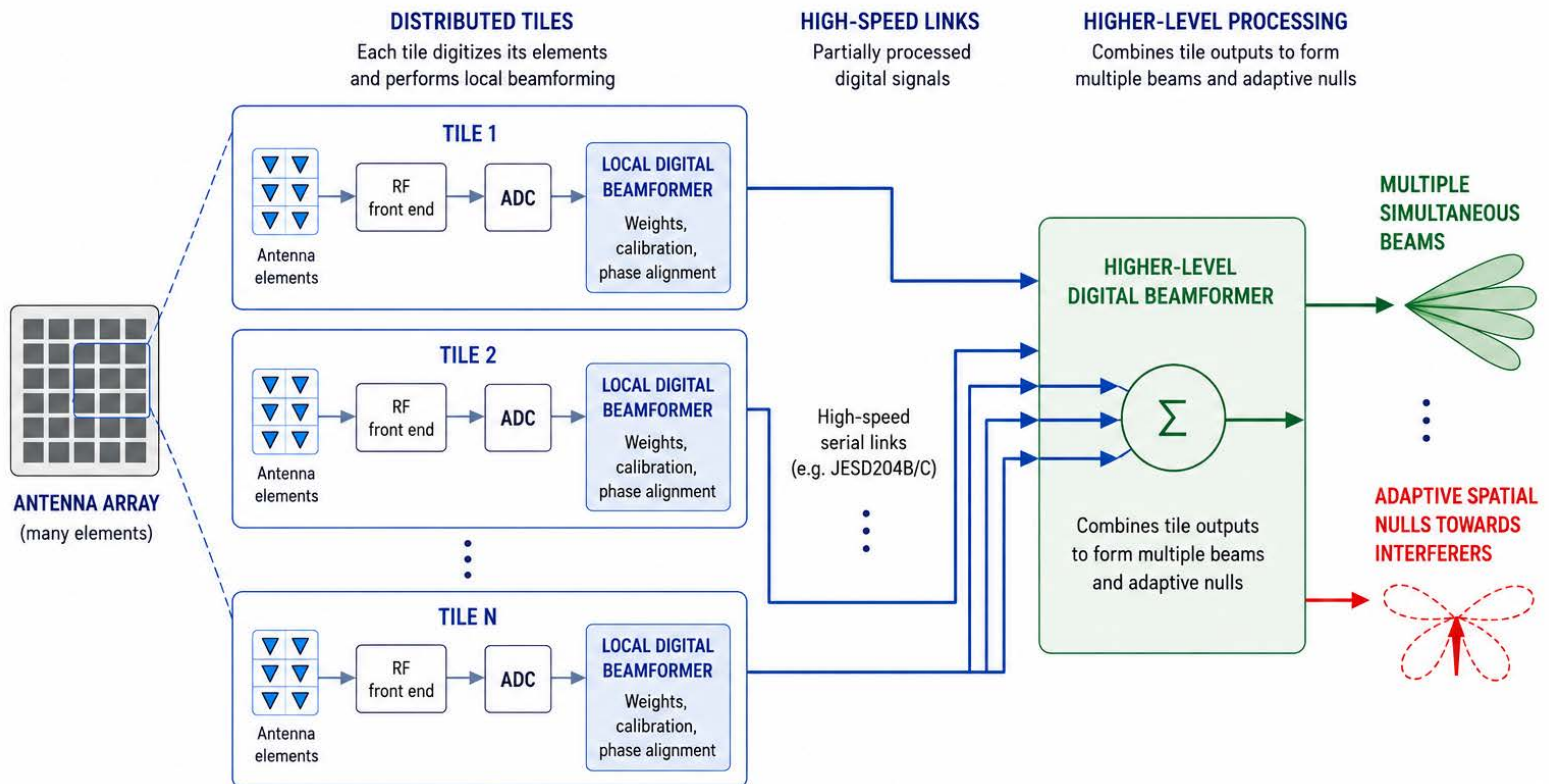
Handover behavior is a further consequence of these architectural choices (Fig. 1). Depending on the constellation, a terminal can hand over several times an hour. Break-before-make releases the beam from the outgoing satellite and re-points it in a few microseconds, but reacquisition and

reentry into the network scheduler can leave a gap of tens to hundreds of milliseconds.

Make-before-break removes that gap by forming a second beam toward the incoming satellite and synchronizing on it while the first still carries traffic. That beam can come from a second panel, a second beamforming chain on the same aperture, or time-interleaved beam hopping at the cost of duty cycle. Splitting one aperture between two beams costs roughly 3 dB of gain per beam. Conventionally, concurrent

DISTRIBUTED DIGITAL BEAMFORMING (DDBF) ARCHITECTURE

The array is built from identical tiles. Each tile performs local beamforming; tile outputs are combined digitally to create multiple beams and adaptive nulls.



2. Distributed digital beamforming combines locally processed tile outputs to form multiple beams and adaptive nulls.

CHAPTER 7: Evolving Antenna Requirements for LEO and MEO

beams must be paid for in hardware.

Accurate steering also depends on knowing the satellite's actual location, not just where the ephemeris says it should be. Attitude sensor drift, mounting misalignment, thermal distortion, and residual calibration error all leave a pointing offset, and with narrow beams that offset is expensive: An error of half the half-power beamwidth costs 3 dB by definition.

Monopulse tracking using the subarrays closes the loop: the aperture forms a sum beam and one or two difference beams at the same time, and the ratio of difference to sum gives a signed angle-error estimate that's largely independent of received signal level. In a hybrid array, those patterns are just two weight sets applied to subarray outputs, so fine pointing costs digital combining rather than RF hardware.

As arrays grow further, distributed digital beamforming (DDBF) becomes increasingly attractive. Rather than sending raw data to a central processor, processing is done locally within antenna tiles, each integrating antenna elements, RF front-end circuitry, data converters, and local digital processing, with partially processed signals combined by higher-level beamformers over high-speed serial links (**Fig. 2**).

In practice, those are JESD204-class interfaces (JESD204B or JESD204C), the established standard between data converters and digital logic, which add deterministic latency and multi-lane

synchronization. Therefore, tile outputs arrive with a known, repeatable sample alignment rather than merely a known bit rate; coherent combining depends on the tiles agreeing on time. A 64-element Ka-band receive panel, for example, can be built as an 8x8 patch aperture served by four analog subarrays, each with its own digital beamformer, interconnected over JESD204B.

Because each tile sends out an already-combined signal instead of raw element data, the interconnect carries far less traffic than a centralized digital array would, and extra tiles can be added without widening a central bus. Clocking benefits, too: Instead of routing a high-frequency Ka-band local oscillator across the panel, a low-frequency reference is distributed and multiplied up by phase-locked loops (PLLs) inside each tile. LO paths stay short; only digitized signals and a reference clock cross the array, and the array grows by replicating identical tiles.

This structure also helps with handover. Because each tile has its own digital processing, a second beam is created by applying another set of weights to the tile outputs rather than by adding RF chains or a second panel. Thus, concurrent beams become a software choice rather than a build-time one. The aperture can be shared for the few seconds of an overlap window and returned to a single full-gain beam immediately afterwards, making

the 3-dB penalty temporary rather than designed in.

Each tile also keeps its own calibration and timing, so a newly assigned beam is coherent as soon as it's formed. The incoming link can be verified before the outgoing one is released at the cost of a little extra digital logic inside the tile rather than a second aperture and a second set of front ends.

System-Level Benefits of Distributed Architectures

Distributed architectures also bring advantages beyond data movement. Coherent combining across elements improves signal-to-noise ratio (SNR) by the array gain, $SNR_{array} \approx N \cdot SNR_{element}$, where N is the number of antenna elements, so each doubling of coherently combined elements adds about 3 dB.

Because ADC quantization noise is largely uncorrelated between channels, combining suppresses it relative to the wanted signal, allowing lower-resolution ADCs at the element level, saving power and cost. Phase noise behaves similarly: Contributions from individual PLLs are partially uncorrelated, so the wanted signal adds coherently while phase noise averages, improving beam phase stability.

Spatial Processing and Interference Mitigation

Another major strength of subarrays using DDBF is spatial interference

CHAPTER 7: Evolving Antenna Requirements for LEO and MEO

mitigation. Adjacent satellite interference is increasingly common, and large arrays can steer nulls toward unwanted sources while holding gain on the wanted one.

In hybrid architectures, this depends on the number of independent digital channels, which correspond to subarrays rather than elements. With N_{DBF} digital channels, roughly $M \approx N_{DBF} - 1$ interference sources can be suppressed. More subarrays mean more spatial degrees of freedom, at the cost of complexity.

Minimum variance distortionless response (MVDR) beamforming and multiple signal classification (MUSIC) remain the standard tools. MVDR minimizes interference power while preserving the wanted signal, and MUSIC is used for high-resolution direction finding. Machine-learning methods are also being explored for direction-of-arrival estimation and adaptive beamforming.

Reflectors will remain efficient where high gain and aperture efficiency are the primary requirements. But the growth of LEO constellations and emerging multi-orbit networks is shifting the balance toward architectures that deliver agile beamsteering, wide-angle scanning, handover, and stronger interference mitigation.

Phased arrays, hybrid beamforming and DDBF all address these demands, and together they show that antenna design is no longer just an RF front-end problem — it's closely coupled to

digital processing, system partitioning and implementation scale. At [EnSilica](#), we focus our innovation on silicon, algorithms, and reference designs that make these architectures practical at scale for satellite user terminals and payloads.

[view this article online](#)

 **BACK TO TABLE OF CONTENTS**