WHITEPAPER

Powering AI-RAN: Evolving telecom power architectures for the GPU era

How telecom power systems must adapt to GPU-class baseband loads across rectifiers, point-of-load conversion, and battery backup

Abstract

AI-RAN promises a more intelligent and more adaptable radio access network, but it also introduces a power paradox: the same GPU-class acceleration that enables AI-driven spectral efficiency, adaptive beamforming, and new edge services also brings faster current transients, wider load swings, and higher thermal density into telecom infrastructure. Power systems that were historically engineered for comparatively deterministic baseband behavior must now support a more dynamic electrical profile without sacrificing availability, efficiency, or footprint.

Recent over-the-air work across the United States, Japan, and Indonesia shows that AI-RAN is no longer hypothetical. At the same time, mobile traffic continues to grow globally, with India, Nepal, and Bhutan already among the highest smartphone data-consumption markets and still climbing [2]. This combination of rising traffic and AI-native processing shifts RAN modernization from a compute-only question to a power-architecture question.

This paper examines that transition across three practical domains: AC-DC telecom rectifiers, DC-DC conversion for GPU-class rails, and battery backup. It uses stage-specific examples from Infineon's CoolGaN™, CoolSiC™, CoolMOS™, OptiMOS™, XDP™, and EiceDRIVER™ portfolios to illustrate the semiconductor classes increasingly relevant to AI-RAN power design, while remaining focused on architecture rather than promotion.

Table of contents

1 The AI-RAN shift — From deterministic to dynamic power 3
1.1 Why AI in RAN is accelerating 3
1.2 GPU-class compute enters the baseband 4
1.3 How AI-RAN power behavior shifts: Deterministic → dynamic, spiky, nonlinear 4
2 System-level power architecture for AI-RAN 5
2.1 The classical telecom power chain 5
2.2 The AI-RAN power chain: What changes most 5
2.3 Quantifying the power demand shift 6
3 AC/DC telecom rectifiers — Scaling for GPU-Era power density 7
3.1 Classic rectifier architectures: Where they fall short 7
3.2 Wide bandgap semiconductors: The GaN and SiC advantage 7
3.3 Infineon solutions: CoolGaN™ and CoolSiC™ in high-efficiency rectifiers 8
4 DC/DC Conversion — Taming GPU rail transients 10
4.1 The 72 W L4 GPU: A case study in efficiency 10
4.2 Why GPU rails are different: 300–600 A, sub-microsecond transients [10], [11], [12] 10
4.3 48 V DC distribution: The telecom-to-AI architecture bridge 11
4.4 Multiphase buck converters and digital control: Infineon's OptiMOS™ and EiceDRIVER™ 11
5 Battery backup under AI-RAN load profiles 13
5.1 Why AI-RAN changes battery behavior: Shorter hold-up times 13
5.2 Bidirectional power flow and fast-charging requirements 13
5.3 Power MOSFETs and integrated gate drivers: Infineon's approach 14
6 The iron triangle — Efficiency, density, and reliability 15
6.1 Efficiency: Every percentage point counts 15
6.2 Power density: Shrinking footprints in edge deployments 16
6.3 Reliability: Telco five-nines under dynamic loads 16
6.4 System-level integration: Coordinated power, thermal, and telemetry 16
7 Conclusion 17
8 References 18

1 The AI-RAN shift — From deterministic to dynamic power

AI-RAN is best understood as the moment when telecom infrastructure begins to behave less like fixed-function communications hardware and more like distributed edge compute. The AI-RAN Alliance organizes this transition around three complementary models [15]:

- Using AI to improve RAN performance
- Sharing infrastructure between AI and RAN workloads
- Running AI applications on RAN infrastructure

Industry platforms from Nokia and NVIDIA illustrate how these models are moving toward implementation [16], [17]. That architectural move changes not only what the RAN can do, but also how it draws power.

1.1 Why AI in RAN is accelerating

Three forces are driving adoption.

- First, traffic demand keeps rising and is becoming more heterogeneous, increasing the value of real-time, AI-assisted optimization [18]
- Second, AI is now being tied directly to concrete RAN use cases rather than generic future promises that include:
  - using traffic and channel measurements to optimize beamforming parameters
  - applying learned interference patterns to support intercell coordination
  - monitoring radio and equipment telemetry to identify abnormal behavior before it develops into a service- fault
- Third, the ecosystem has moved far enough that operators, infrastructure vendors, and silicon providers are now aligning around deployment roadmaps rather than isolated experiments:
  - adaptive MIMO beamforming that tunes transmission to user and channel conditions in real time
  - interference mitigation that predicts and suppresses intercell or intracell interference more intelligently than static rule sets
  - anomaly detection for hardware faults, spectrum intrusions, and site-level performance drift

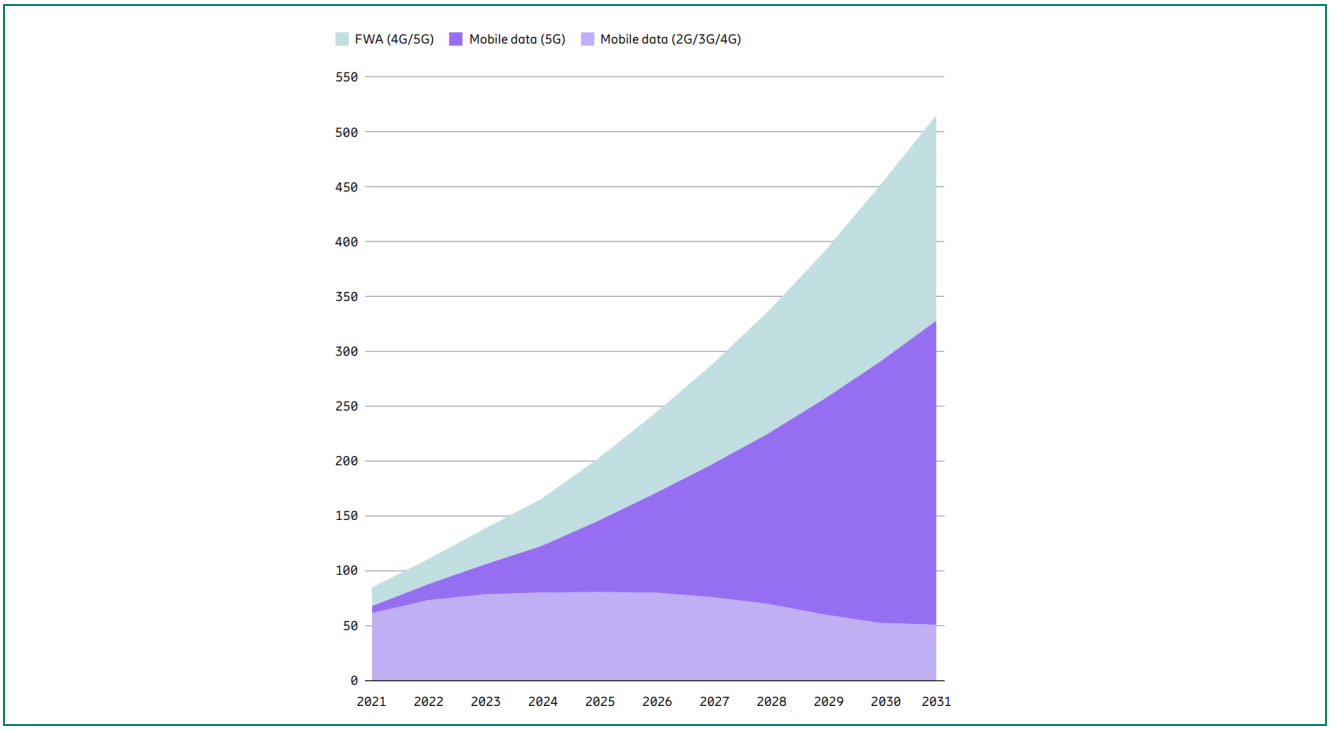


Figure 1 Global mobile network data traffic growth and forecast [18]

The timing also matters. 5G Advanced brings AI and machine learning more visibly into standardization and optimization workflows, while the path to 6G is increasingly framed as AI-native by design. In practical terms, that means AI in the RAN is no longer a distant research topic; it is part of near-term network planning.

1.2 GPU-class compute enters the baseband

In classic basebands, Layer-1 and adjacent processing were dominated by ASICs, DSPs, and FPGAs with relatively stable and predictable electrical behavior. AI-RAN changes that by introducing heterogeneous compute: CPUs, GPUs, and sometimes additional accelerators share the same platform. The move is attractive because GPUs handle matrix-heavy AI inference efficiently and support software-defined evolution, but it also changes the electrical personality of the baseband.

NVIDIA L4 is a useful efficiency-focused example [3]. Its 72 W thermal design power and low-profile form factor show that not every AI-RAN platform is a hyperscale-class power monster. Yet even relatively efficient accelerators still require low-voltage, high-current rails and can demand a faster transient response than classic telecom silicon.

1.3 How AI-RAN power behavior shifts: Deterministic → dynamic, spiky, nonlinear

The electrical difference can be summarized simply: legacy basebands were comparatively deterministic, while AI-RAN basebands are more dynamic, spiky, and nonlinear. In a classical ASIC-dominated design, power might vary only modestly across traffic conditions. In a GPU-assisted design, utilization can swing far more aggressively as inference bursts, model transitions, and mixed workloads change current demand in short time intervals.

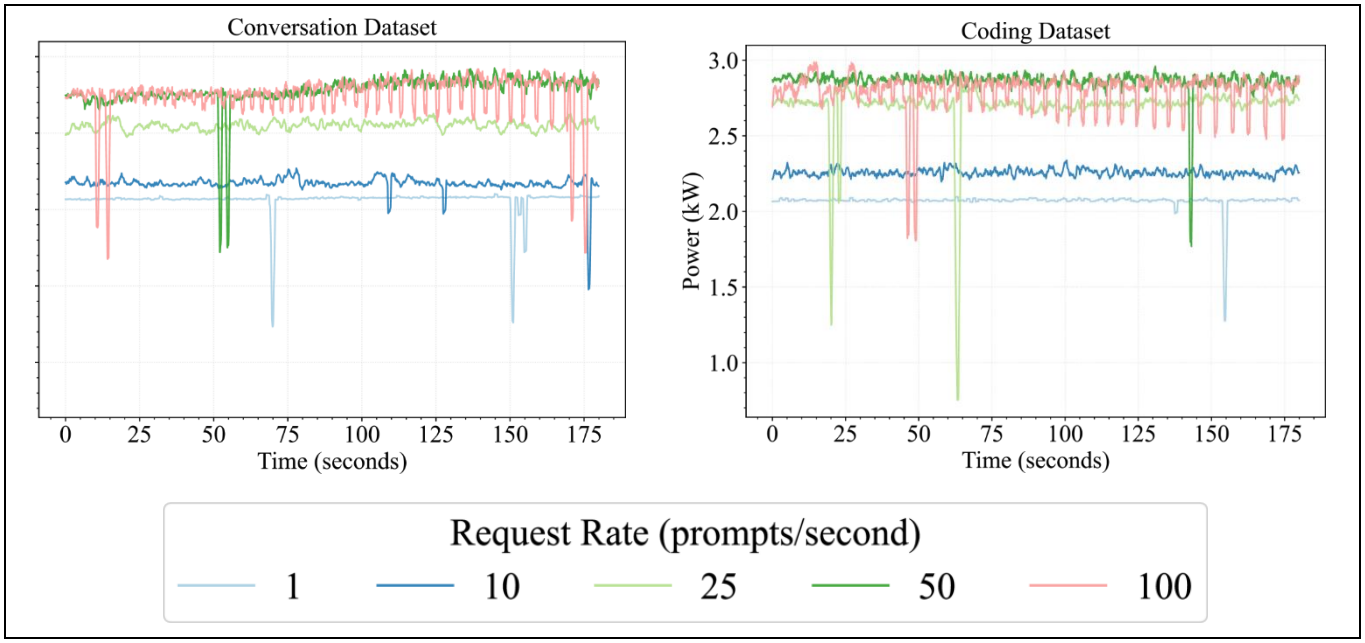


Figure 2 Measured **power profiles for a Llama-3 70B online inference workload under sustained request rates. The** profiles illustrate the rapid and workload-dependent power variations associated with accelerator-based inference [24]

That matters because the change propagates upstream. Rectifiers see a wider and less forgiving load window. The −48 V distribution bus must tolerate sharper current demand shifts. Point-of-load converters must supply low-voltage rails with tighter transient control. And battery systems must support a site whose average and instantaneous behavior are both changing.

2 System-level power architecture for AI-RAN

Before looking at individual stages, it helps to compare the classical telecom power chain with the version AI-RAN increasingly demands. The overall topology remains familiar, but the behavior of the loads hanging from that topology changes enough to reshape design priorities throughout the chain.

2.1 The classical telecom power chain

The conventional telecom site follows a well-understood path: AC mains feed a rectifier shelf, the rectifier produces a regulated −48 V bus, distributed DC/DC conversion generates the lower rails needed by radios and baseband electronics, and a battery backup system maintains continuity during outages or transfers. The architecture is robust and serviceable, and it became dominant precisely because it balanced safety, availability, and maintainability well.

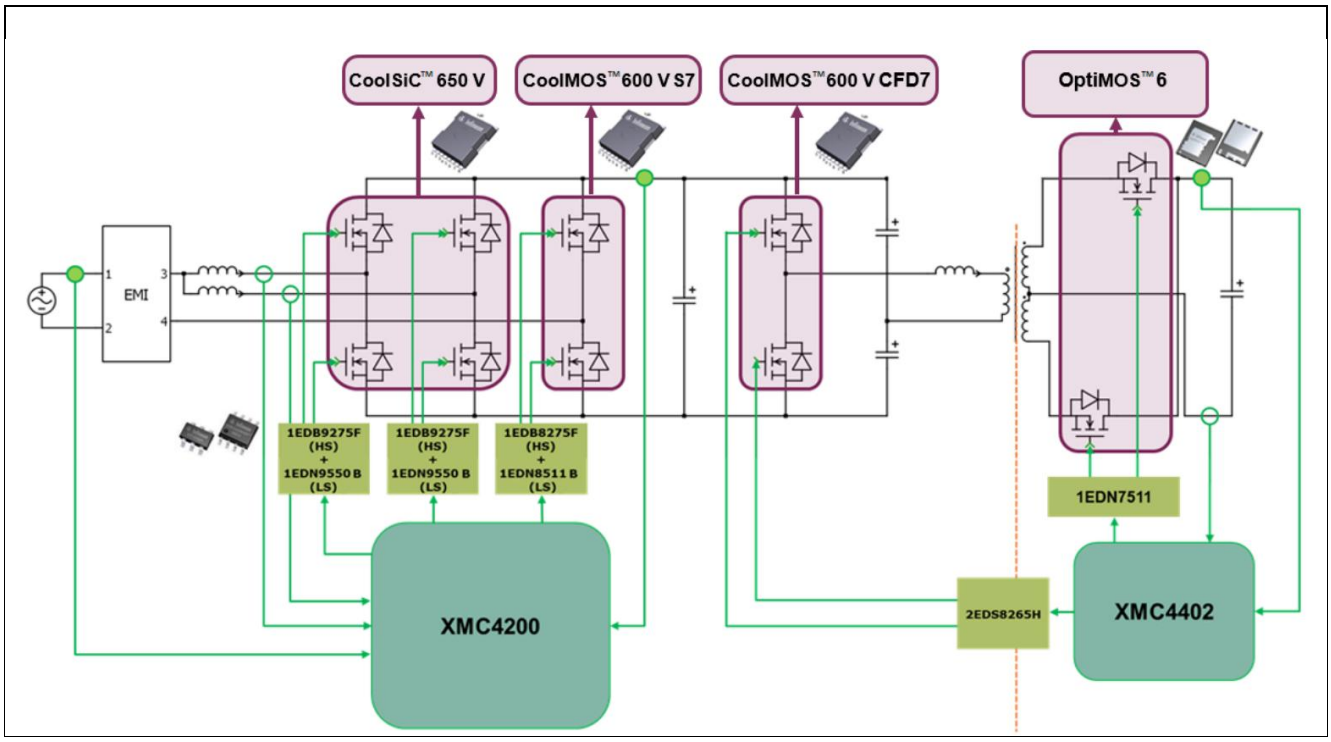


Figure 3 Simplified topology of the 1 kW telecom power supply, showing the implemented PFC and isolated DC/DC stages and the main Infineon power semiconductors

In that classical setting, load profiles were meaningful but manageable. Power demand correlated more directly with radio traffic and changed on timescales that the rest of the system could absorb without unusual stress.

2.2 The AI-RAN power chain: What changes most

AI-RAN deployments remain at an early stage, and published, system-level power measurements are consequently limited. Although AI-RAN systems differ from hyperscale AI servers in scale and operating environment, both use accelerator-based compute that introduces dynamic loads, low-voltage high-current rails, and demanding thermal constraints. This paper therefore draws selectively on established AI-server power architectures to identify techniques relevant to emerging AI-RAN systems.

AI-RAN keeps the same broad structure but alters the sensitivity of each stage. The front end must sustain higher efficiency over a wider operating window. The −48 V bus must remain quieter and more stable under sharper current excursions. The DC/DC stage must support low-voltage, high-current accelerator rails with tighter transient control. And the backup domain must respond to a site that no longer draws power in a smooth, largely predictable way.

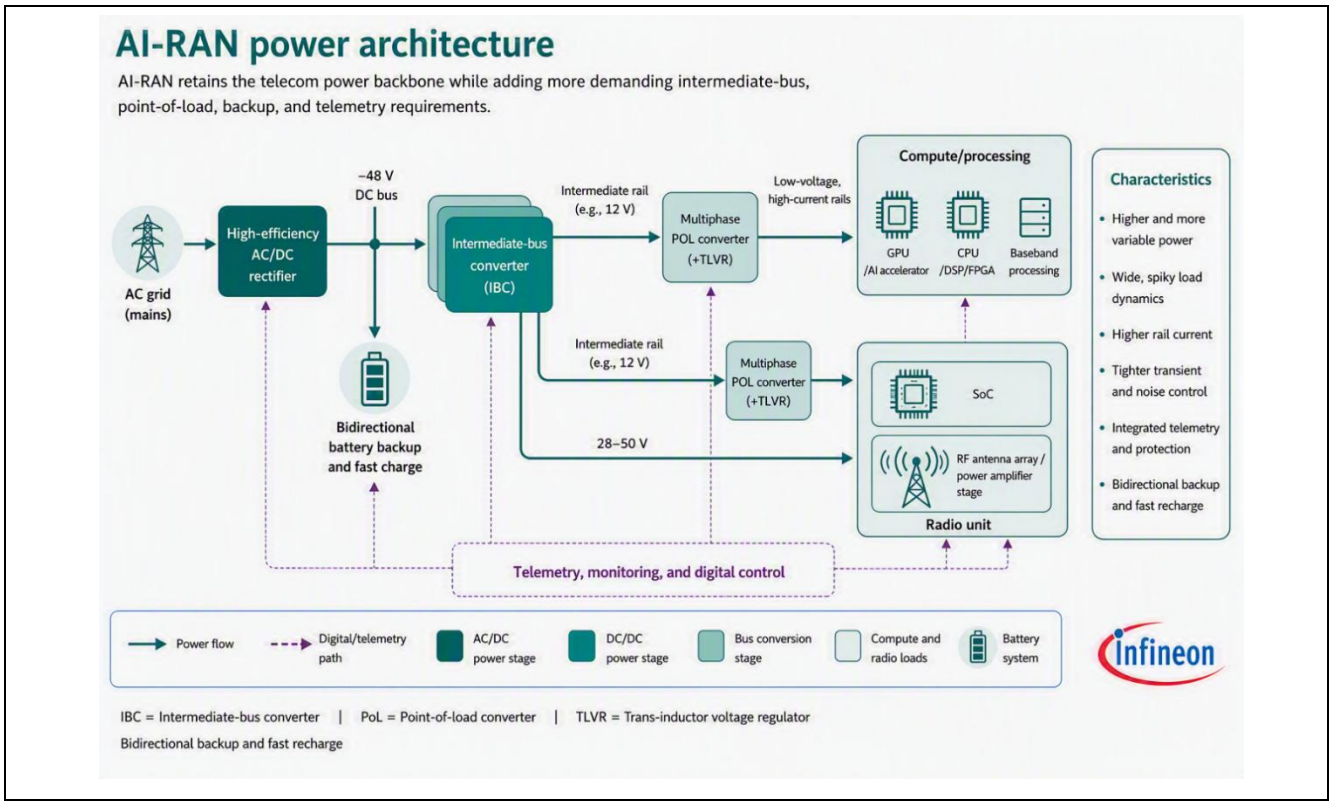


Figure 4 **Evolution from the classic telecom power chain to an AI-RAN power architecture. AI-RAN retains the** telecom power backbone while adding more demanding intermediate-bus, point-of-load, backup, and telemetry requirements (AI generated image)

These demands become especially visible in distributed outdoor deployments. 5G pushes radios, edge compute, and power electronics into poles, rooftops, and building facades where there is no room for oversized cabinets and no controlled indoor environment. In that context, electrical inefficiency shows up twice: once on the utility bill and again as trapped heat inside a constrained enclosure.

2.3 Quantifying the power demand shift

Table 1 **Classic telecom power chain versus AI-RAN power chain**

Metric	Classic RAN baseband	AI-RAN site	Implication
Baseband behavior	Comparatively predictable and traffic-correlated	Dynamic, burst-driven	Wider control and thermal envelope required
Site power demand	Determined primarily by radio and conventional baseband loads	Adds accelerator and associated cooling loads	Rectifier and distribution capacity must include accelerator headroom
48 V bus current	Follows the existing site power budget	Increases with accelerator and cooling demand	Higher conduction losses and connector stress
Load-step behavior	Typically less abrupt	Faster and wider current excursions	Tighter bus stability and downstream transient control
Backup demand	Longer hold-up at lower load	Shorter hold-up at higher variable load	Battery strategy must change

A useful mental model is – 5G raised the baseline, and AI-RAN makes the baseline more dynamic. That is why the power system must be treated as an architectural concern from the beginning, not as a support layer adjusted after the compute platform is selected.

3 AC-DC telecom rectifiers — **Scaling for GPU-Era** power density

If AI-RAN changes the behavior of the load, the AC-DC rectifier is where the site first feels that change in aggregate. It remains responsible for high-reliability conversion from the grid to the site's −48 V bus, but it now does so under wider load variation, higher density expectations, and more restrictive thermal conditions.

3.1 Classic rectifier architectures: Where they fall short

Telecom rectifiers span a broad range of designs. High-end silicon rectifiers have achieved efficiencies approaching 97% for more than a decade. Commercial silicon rectifiers also span multiple performance generations. The examples in Figure 5 show an earlier 3 kW rectifier with 96.2% peak efficiency and a later high-end 3 kW rectifier designed to maintain at least 95% efficiency from 30% to full load at 230 V AC.

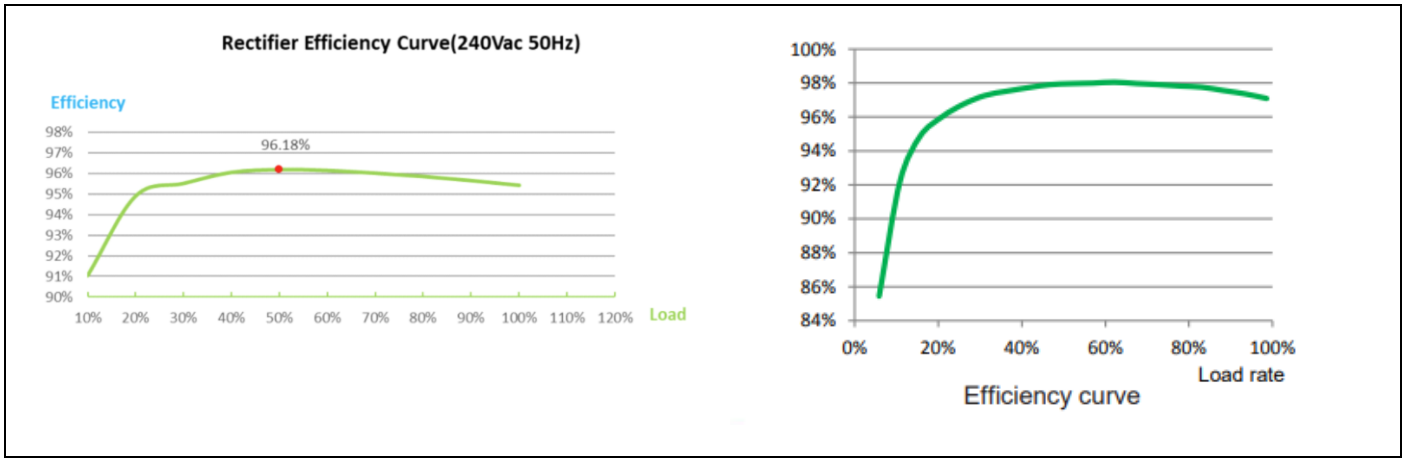


Figure 5 Efficiency characteristics of two generations of commercial silicon telecom rectifiers; Rectronic's R830H 3.0 kW 48 V high efficiency DC modular rectifier [22] (left) and Huawei's R4850G2 rectifier [23] (right). The comparison illustrates that rectifier performance depends on design generation, topology, and optimization priorities

The AI-RAN design challenge is therefore not simply to replace the perceived inefficiency of silicon-based rectifiers, but to sustain high and relatively flat efficiency across a wider load range while increasing power density and managing thermal constraints.

Many solutions used silicon superjunction MOSFETs in two-stage architectures built around PFC and isolated high-voltage DC/DC conversion. Those designs still work, but they are less well suited to the combination of thin form factors, natural or restricted convection, and wide load swings that AI-RAN introduces.

The problem is not simply that AI-RAN draws more power. It also takes more time away from a narrow efficiency sweet spot. Classic designs tuned for high efficiency close to rated load can lose proportionally more under wide 30-100% swings, and their thermal burden scales poorly in small cabinets.

3.2 Wide bandgap semiconductors: The GaN and SiC advantage

Wide-bandgap devices improve the front end in two practical ways: they cut losses, and they allow higher switching frequencies that shrink magnetic components and heatsinks. Gallium nitride is especially attractive in high-frequency, mid-voltage switching roles, while silicon carbide offers strong benefits in high-voltage and high-current positions where ruggedness and thermal conductivity matter most.

Table 2 Indicative front-end comparison: legacy silicon versus WBG-enabled designs [5], [8]

Parameter	classic silicon design	GaN / SiC enabled design	Typical effect
Peak efficiency	Up to 97%	97–98%+	Lower heat generation
Switching frequency	50–250 kHz	65–600 kHz	Smaller magnetics
Power density	30–40 W/in³	50–70 W/in³	Smaller shelf footprint

In practical telecom terms, that means the front end can be made flatter in efficiency across load range, denser in packaging, and less punishing to the enclosure thermal budget.

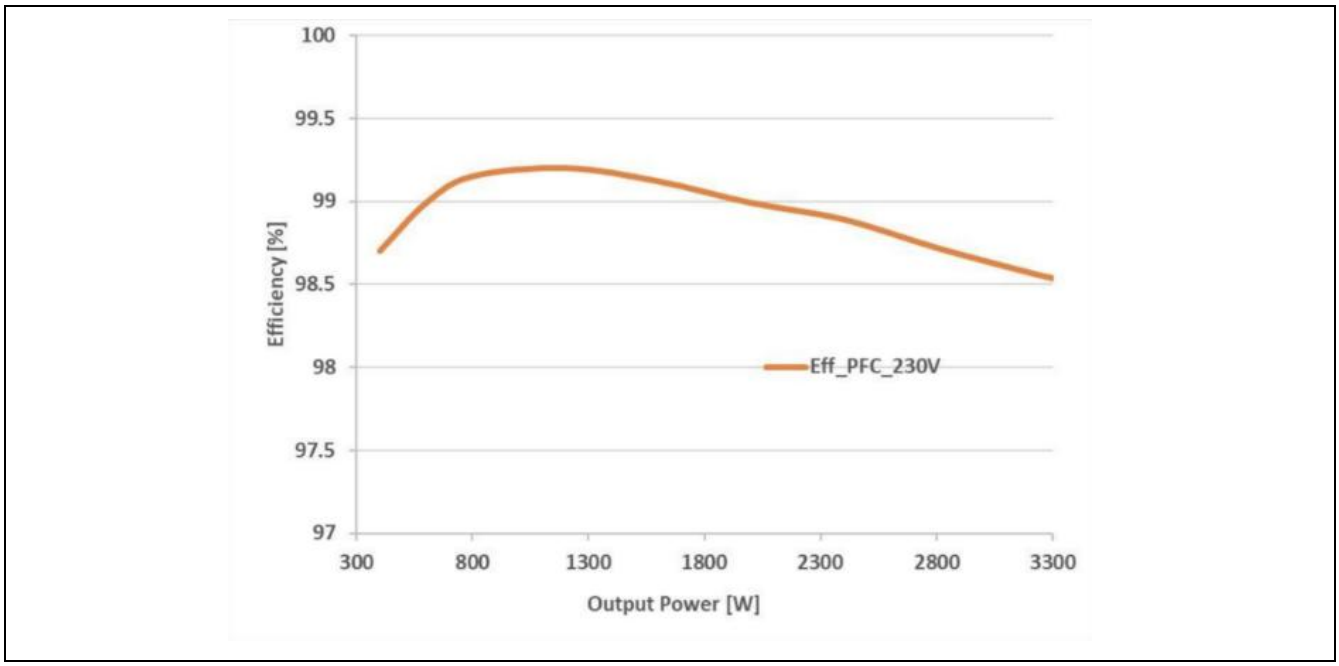


Figure 6 Measured efficiency of a 3.3 kW bridgeless totem-pole PFC using CoolSiC™ G2 and CoolMOS™ devices

3.3 Infineon solutions: CoolGaN™ and CoolSiC™ in high-efficiency rectifiers

Infineon's telecom AC/DC portfolio illustrates the stage-level design pattern now emerging in the market. CoolGaN™ supports hard- or soft-switching roles where high frequency and compactness matter; CoolSiC™ addresses high-voltage, high-current switching with strong thermal and surge performance; CoolMOS™ remains relevant in hybrid designs; and EiceDRIVER™ devices help implement these stages cleanly and compactly.

Infineon's 1 kW and 500 W telecom power-supply designs target outdoor small-cell and edge-computing applications [5], [6], [7]. 3.3 kW REF-3K3W-HFHD-PSU extends these design principles to a higher-power, high-frequency, high-density implementation [21]. These designs illustrate how Si, SiC, and GaN can be combined according to topology, switching-frequency, efficiency, and power-density requirements. For AI-RAN, the rectifier should therefore be evaluated not only by peak efficiency, but also by its efficiency across the operating range, power density, and contribution to the enclosure thermal budget.

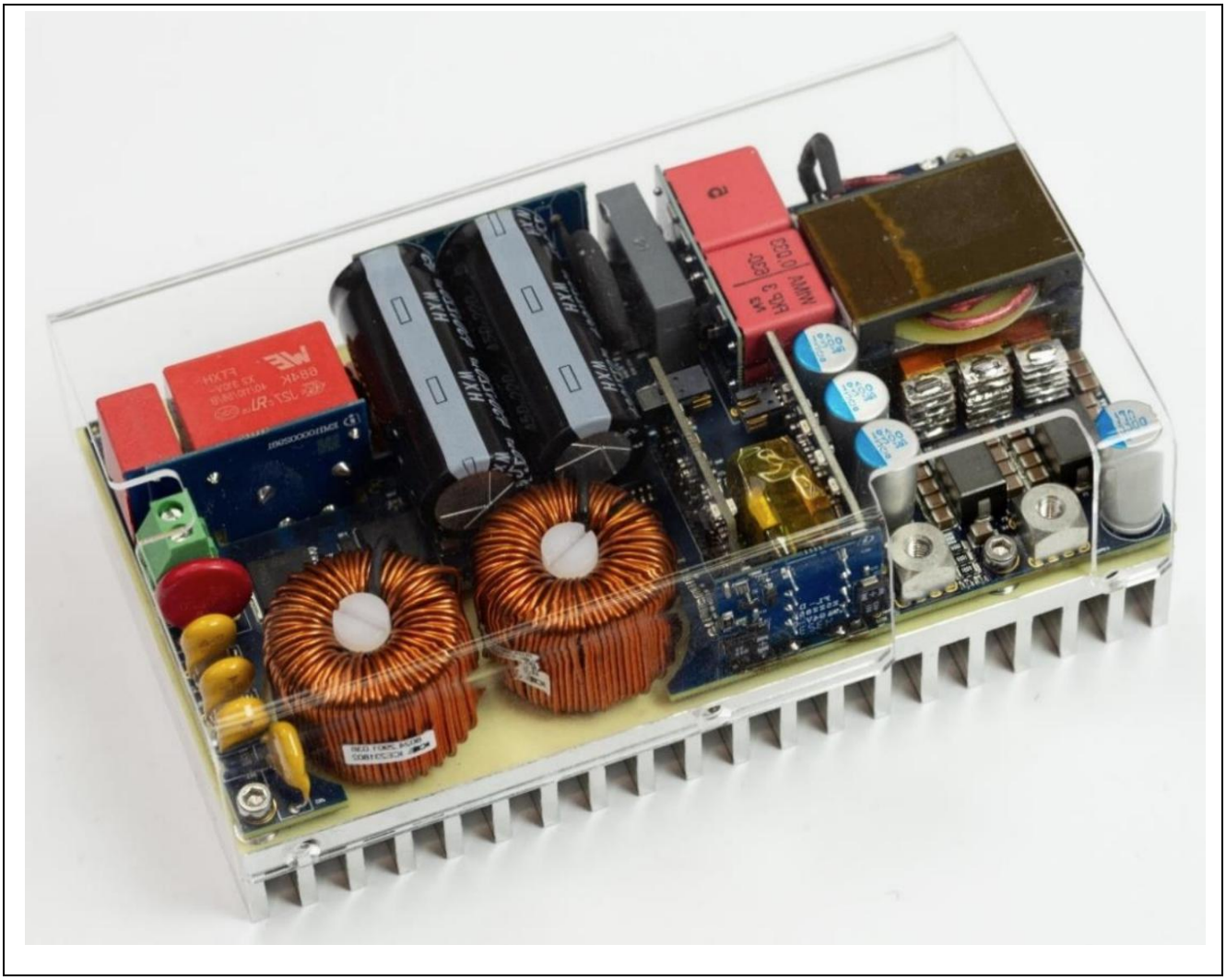


Figure 7 Compact 1 kW fanless power supply for edge-computing and outdoor small-cell applications using CoolSiC™ and CoolMOS™ devices

4 DC/DC Conversion — Taming GPU rail transients

If the rectifier determines how efficiently power enters the site, the DC/DC stage determines whether it arrives at accelerator rails in a form the compute platform can tolerate. In AI-RAN, that makes point-of-load conversion the most timing-critical electrical stage in the site.

4.1 The 72 W L4 GPU: A case study in efficiency

A commercial 72 W, low-profile inference accelerator provides a useful case study because it shows that AI-RAN does not always imply the highest-power accelerator classes. NVIDIA L4 is one implementation in this class [3]. Even at this comparatively modest board-power level, integration into a telecom baseband or edge-compute node requires a carefully engineered rail environment.

Table 3 **Illustrative rail view for an efficiency-oriented AI-RAN node based on** [3]

Rail / domain	Typical voltage	Peak current	Role
GPU memory	~1.35 V [19]	~10 A	GDDR or equivalent memory rail
System rail	12 V [20]	~15 A	NICs, fans, storage, auxiliaries

The important lesson is not the exact wattage, but the combination of low voltage, meaningful current, and fast rail movement that the converter must absorb.

4.2 Why GPU rails are different: 300–600 A, **sub-microsecond transients** [10], [11], [12]

As AI-RAN platforms scale upward, rail difficulty rises quickly. High-end compute power delivery is already engineered for currents well above those of classic telecom logic, and the same design logic appears here even when the absolute node power is lower. The challenge is not just step-down ratio. It is current slew rate, rail ripple, and the need to keep droop and overshoot inside very tight tolerance windows during abrupt workload changes.

That is why low-ESR capacitance, short current paths, fast control loops, and phase-interleaved conversion become first-order concerns rather than layout details. Inference behavior can trigger rapid current movement, and the converter must respond before the rail leaves its safe operating window.

High-current delivery also depends on maintaining balanced phase currents during load transitions. [12] shows that paired power stages in a dual-phase multiphase implementation remain closely balanced during a 300 A load step.

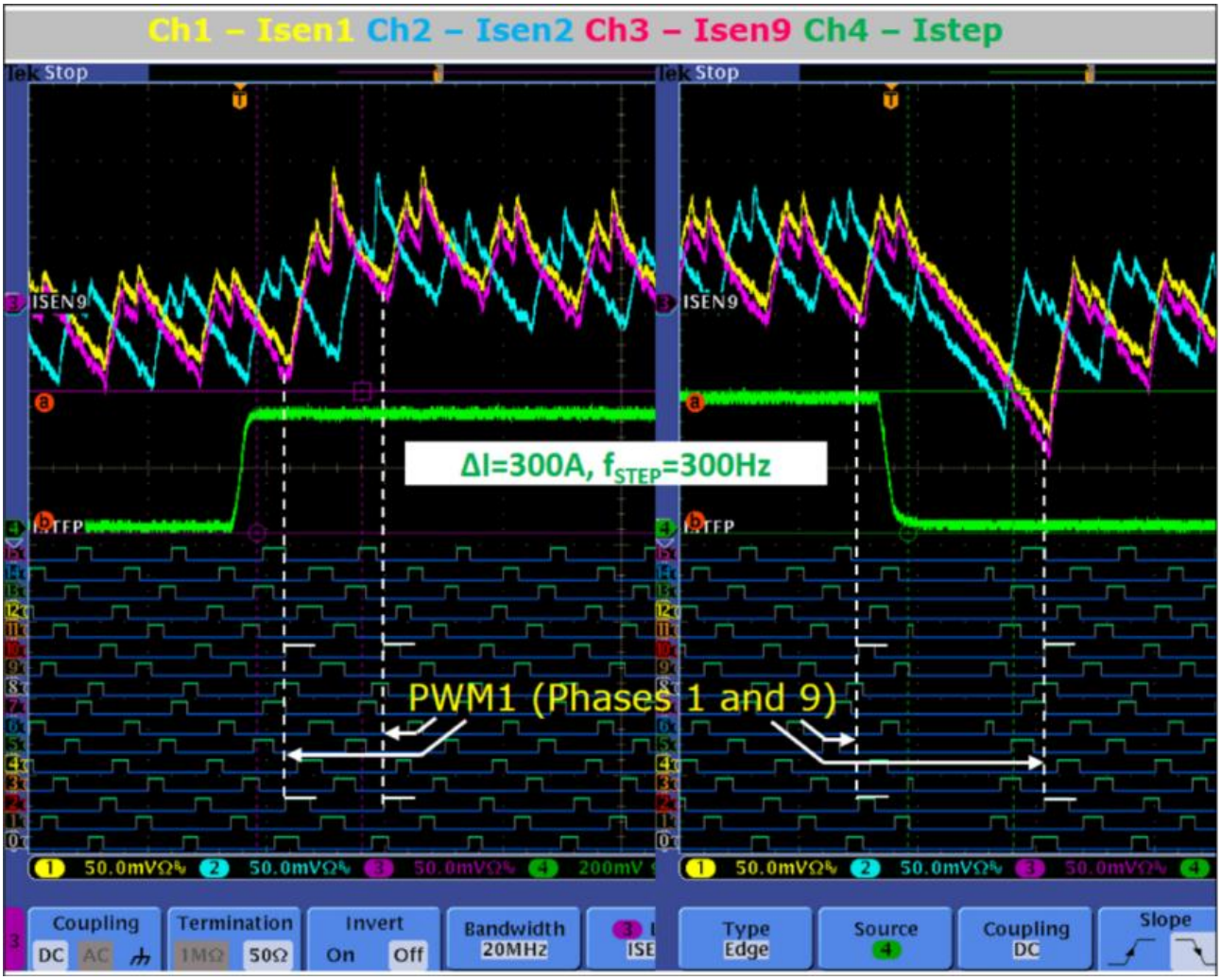


Figure 8 Balanced phase-current response during a 300 A load step. Multiphase conversion distributes the rapidly changing current demand across paired power stages, supporting stable delivery to GPU-class rails

4.3 48 V DC distribution: The telecom-to-AI architecture bridge

One strategic advantage telecom infrastructure already has is its familiarity with 48 V distribution. Compared with lower-voltage bulk distribution, 48 V dramatically reduces current for a given power level, which in turn reduces copper loss and eases distribution. That matters more as rack or cabinet power rises and as rail sensitivity downstream becomes stricter.

In AI-RAN, 48 V is not simply legacy convention. It is the practical bridge between a telecom site and a compute-heavy node. It supports more efficient intermediate conversion, aligns naturally with existing backup practices, and mirrors the direction already taken in AI server and data center power architectures.

4.4 Multiphase buck converters and digital control: Infineon's OptiMOS™ and EiceDRIVER™

The dominant point-of-load answer is the digitally controlled multiphase buck converter. Following the intermediate-bus converter, which steps the 48 V bus down to an intermediate rail, the POL converter uses multiple interleaved phases to spread current, raise effective ripple frequency, and improve transient response. In AI-RAN, that architecture becomes especially valuable because it supports both rail stability and compact thermal distribution.

Infineon's XDP™ controller family, together with low-voltage OptiMOS™ stages and EiceDRIVER™ gate drivers, illustrates the design pattern across the DC/DC chain: digital control and telemetry at the IBC stage, adaptive transient management and high phase count conversion at the POL stage. This provides enough flexibility to support changing accelerator requirements without redesigning the whole power stage.

TLVR — trans-inductor voltage regulator — is the most important topology refinement used at the multiphase POL stage in this context. By coupling inductors and reshaping transient behavior, TLVR reduces output-capacitance requirements and makes fast response more achievable in physically constrained systems.

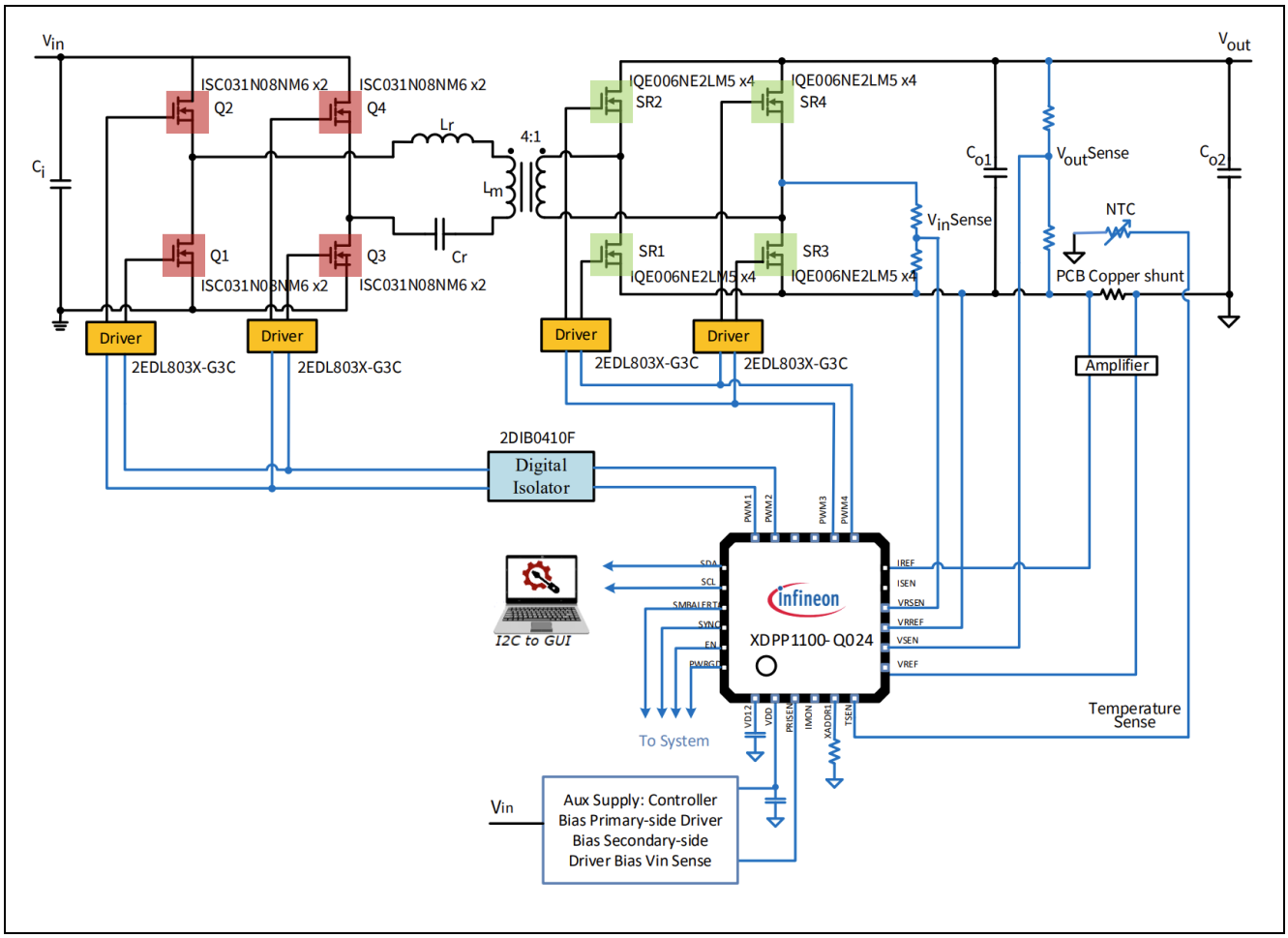


Figure 9 Simplified architecture of a 1 kW intermediate-bus converter operating from a 42–60 V input and producing a 10–15 V intermediate rail

5 Battery backup under AI-RAN load profiles

Battery backup is often treated as a quiet background subsystem in telecom design. AI-RAN changes that. Once the site's average draw rises and its load profile becomes more variable, the backup system becomes a more active part of resilience planning rather than a static hold-up resource.

5.1 Why AI-RAN changes battery behavior: Shorter hold-up times

The first consequence is simple arithmetic. If the same battery must support a site that now draws materially more power, hold-up time shrinks. But AI-RAN changes more than average power. It introduces more variable discharge behavior, faster current excursions, and stronger thermal coupling between the battery, rectifier losses, and enclosure conditions.

Similar bidirectional BBU architectures are already being developed for AI servers, where rapid load changes and stringent space constraints create comparable power-conversion challenges. Although telecom backup requirements differ in hold-up duration, deployment environment, and serviceability, these server designs provide useful insights into high-efficiency conversion, modular scaling, fast transient support, and coordinated battery management for future AI-RAN sites [13].

A useful planning lens is to compare three responses to that problem: increase stored energy, shed non-critical load, or recover charge faster after an event. In practice, all three can matter depending on site profile and outage behavior.

5.2 Bidirectional power flow and fast-charging requirements

As backup systems interact more tightly with the rest of the site, bidirectional conversion becomes more important. The same DC/DC stage can be used to manage controlled charging and regulated discharge, enabling faster recovery, smarter battery protection, and better interaction with the −48 V bus during events.

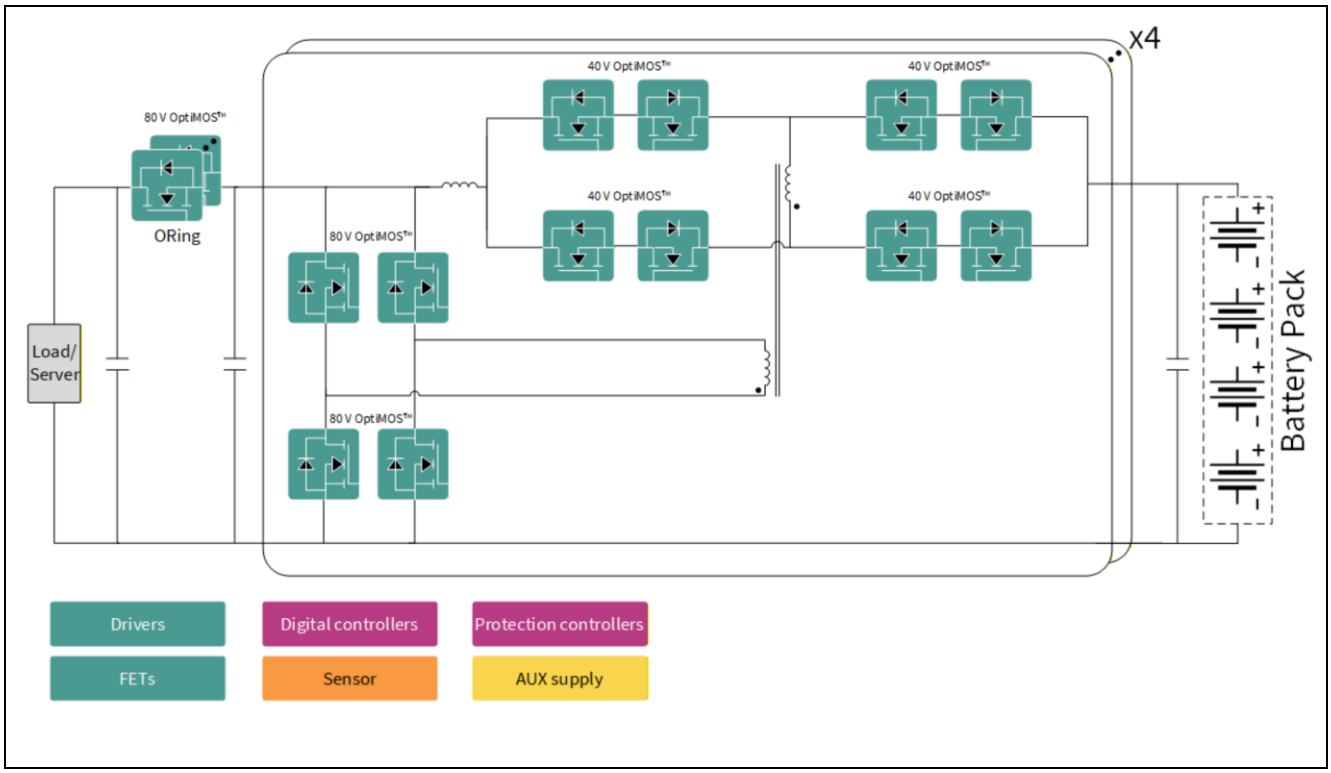


Figure 10 Functional architecture of a bidirectional 12 kW battery backup unit using partial power conversion. The system charges the battery during normal operation and supports the 48 V power bus during disturbances

Table 4 **Telecom backup chemistry comparison for AI-RAN contexts**

Metric	VRLA	Li-ion	Why it matters for AI-RAN
Energy density	Lower	Higher	Smaller battery footprint in constrained cabinets
Fast charge capability	Limited	Strong	Shorter recovery after outage
Cycle life	Lower	Higher	Better fit for variable discharge behavior
Capex per kWh	Lower	Higher	Trade off against lifetime TCO

For AI-RAN, the argument for Li-ion is therefore not merely about fashion or energy density. It is about matching battery behavior to a more variable electrical environment.

5.3 Power MOSFETs and integrated gate drivers: Infineon's approach

In battery-adjacent conversion, the same traits valued elsewhere in the chain remain relevant: lower losses, thermal restraint, and robust control. Infineon’s 48 V, 12 kW BBU design uses OptiMOS™ 7 40 V and OptiMOS™ 6 80 V power MOSFETs, together with EiceDRIVER™ gate-driver ICs, in a bidirectional partial-power-conversion architectures [13]. These components support high-efficiency charge and discharge operation while enabling a modular, high-density design. That is why the BBU problem increasingly looks like a power-conversion problem, not only a battery-sizing problem.

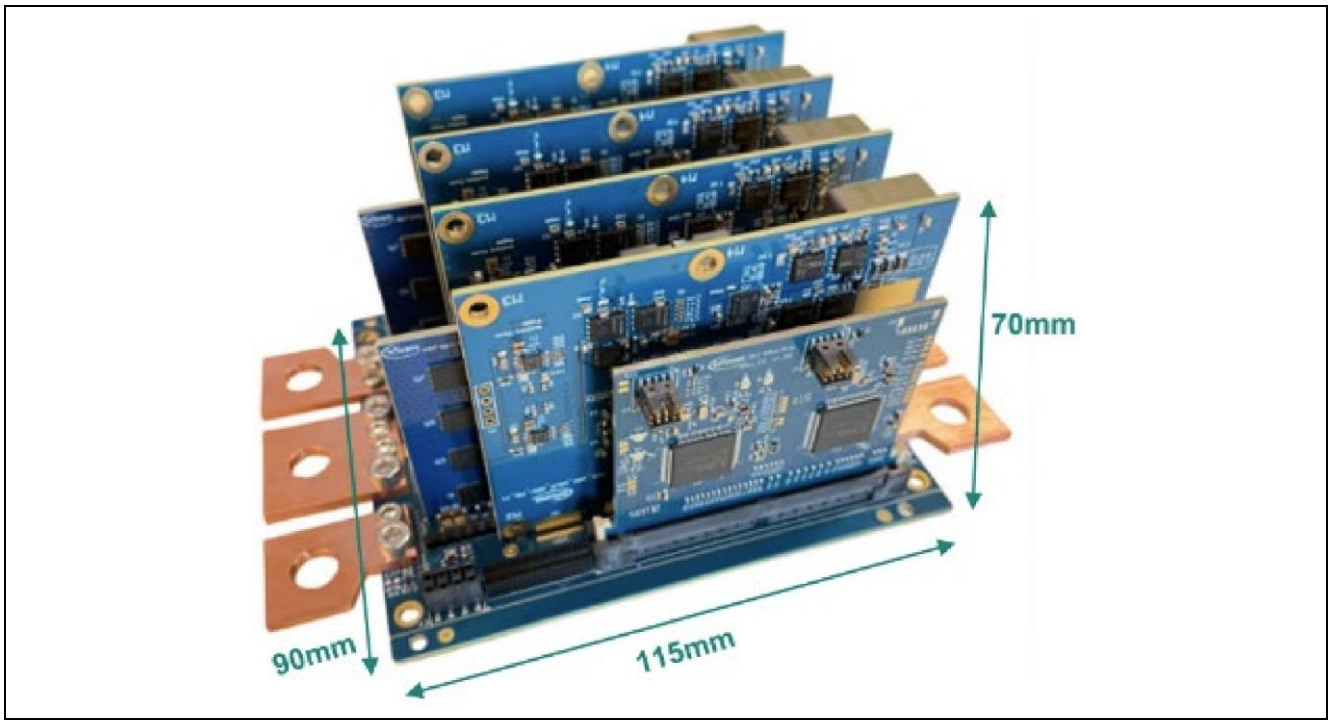


Figure 11 Hardware implementation of the modular 12 kW BBU, comprising four parallel 3 kW power boards together with control and protection boards

6 The iron triangle — Efficiency, density, and reliability

By this point, the pattern becomes clear. AI-RAN power design is governed by an iron triangle of efficiency, density, and reliability. No stage can optimize one of these in isolation without affecting the other two, which is why the best designs are the ones that treat the whole chain as one coordinated system.

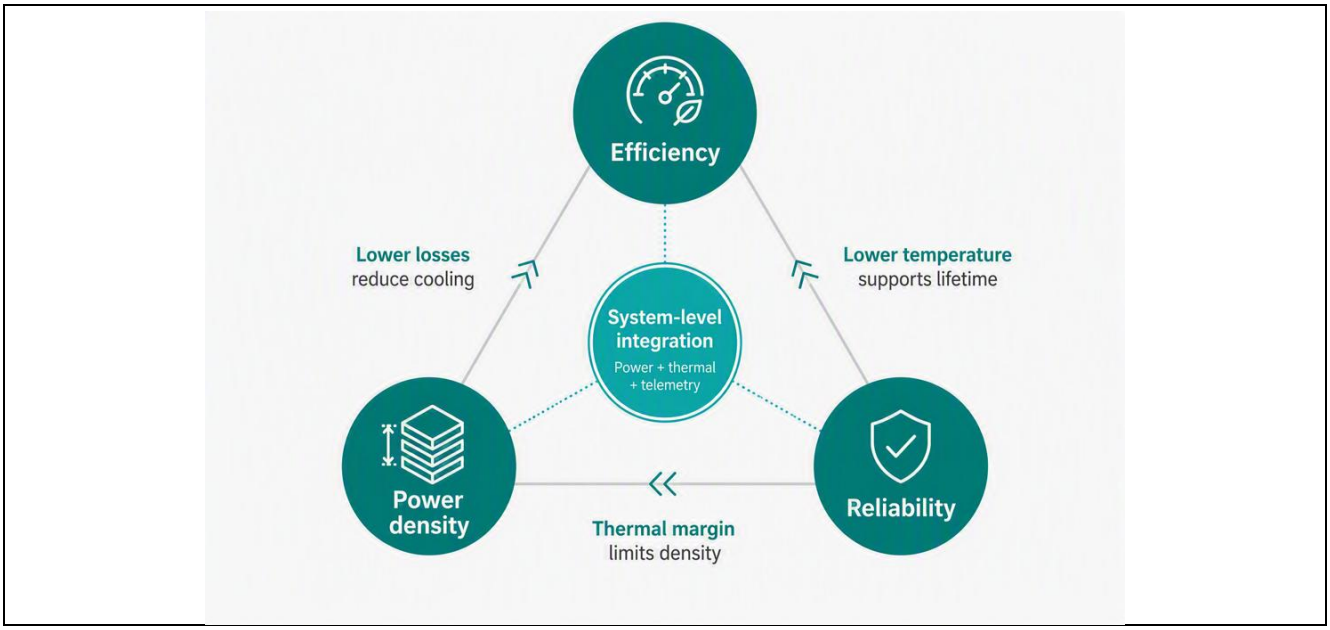


Figure 12 **Efficiency, power density, and reliability form the iron triangle of AI-RAN power design. Coordinated** system-level design is required to improve one objective without compromising the others

6.1 Efficiency: Every percentage point counts

Efficiency matters because every watt lost becomes both operating cost and thermal burden. In AI-RAN, gains compound across the chain: a better front end reduces heat; a better intermediate stage cuts distribution loss; a better point-of-load converter eases rail thermals; and a more efficient backup path reduces stress during both discharge and recharge.

Table 5 Illustrative efficiency chain comparison based on [9]

Stage	Legacy silicon baseline	Improved AI-RAN design	Impact
AC/DC rectifier	95-96%	97-97.5%	Lower enclosure heat
48 V intermediate conversion	92%	94%	Less bus-side loss
Low-voltage GPU rail conversion	87%	90%	Cooler point-of-load operation
Backup discharge path	92%	96%	Less battery-side thermal loss

The stage-level numbers are illustrative, but the principle is robust: apparently small percentage gains compound into meaningful site-level improvement.

6.2 Power density: Shrinking footprints in edge deployments

Power density matters because AI-RAN pushes more useful compute into the same or smaller physical envelope. In practical terms, gains in rectifier density, converter compactness, and battery energy density are what make room for the accelerators and radios the site is actually meant to host.

The important point is that density is not only a mechanical problem. It is an efficiency and thermal problem hiding in mechanical clothing.

6.3 Reliability: Telco five-nines under dynamic loads

Telecom still expects five-nines availability: roughly 99.999% uptime, equivalent to only a few minutes of annual downtime [14]. AI-RAN does not lower that bar. It makes the route to it more complex. The site must now preserve predictable behavior under more variable electrical and thermal conditions, which means reliability depends more heavily on protection, telemetry, and controlled degradation paths than before.

In this context, reliability is not just component ruggedness. It is the ability of the whole site to remain stable and interpretable when the load becomes dynamic.

6.4 System-level integration: Coordinated power, thermal, and telemetry

System-level integration is what keeps the iron triangle from collapsing. A transient on a GPU rail can change demand on the −48 V bus, shift the operating point of the rectifier, and alter thermal conditions that affect battery stress and airflow. Once that happens, the site can no longer be engineered as a collection of disconnected power boxes.

A mature AI-RAN power architecture therefore needs coherent telemetry and coordinated control across the chain: front-end loading and temperature, bus health, per-phase VRM behavior, battery state-of-charge and state-of-health, and enclosure thermal conditions.

When those signals are connected, the site can react more intelligently — moderating AI load before thermal throttling, preserving battery margin during outages, and diagnosing rail problems in context rather than in isolation.

This is the systems effect at the heart of the paper. Efficiency, density, and reliability reinforce each other only when power, thermal design, and telemetry are treated as one architecture rather than parallel workstreams.

7 Conclusion

AI-RAN is not only a compute transition. It is a power-architecture transition. GPU-class acceleration changes how the baseband behaves electrically, how the site stores and recovers energy, and how aggressively each stage of the power chain must protect thermal and reliability margins. The familiar telecom architecture built around the −48 V bus still remains the structural backbone; what changes is the standard each attached stage must now meet.

Five takeaways stand out for the next 24–48 months:

- Treat AI-RAN as a power-infrastructure challenge as much as a compute upgrade. Site-level success depends on power-system behavior, not accelerator choice alone
- Upgrade the front end. High-efficiency rectifiers built on GaN, SiC, and advanced superjunction devices are now a practical requirement for thermal and density reasons, not a nice-to-have improvement
- Engineer GPU rails as first-class infrastructure. Multiphase digital conversion, fast transient response, and TLVR-aware design are central to stable AI-RAN operation
- Redesign backup strategy for higher and more variable load. Battery sizing, bidirectional conversion, and recharge behavior must be treated as active design decisions
- Futureproof around 48 V and system integration. The best architectures are the ones that connect rectifiers, conversion, backup, thermal design, and telemetry into one coordinated system that can scale toward higher-power AI and 6G-native deployments

Put simply, the next generation of telecom infrastructure will be judged not only by how intelligently it computes, but by how intelligently it is powered.

8 References

[1] AI-RAN Alliance: AI-RAN Alliance | Shaping Future AI-Native Networks; https://ai-ran.org/

[2] Ericsson: Mobile Data Traffic Outlook; Ericsson; Available online

[3] NVIDIA: NVIDIA L4 GPU Accelerator Product Brief (PB-11316-001_v01); NVIDIA; March 9, 2023; Available online

[4] M. Mughees, Y. Li, Y. Chen, and Y. R. Li: Short-Term Load Forecasting for AI-Data Center; arXiv; March 10, 2025; Available online

[5] Infineon Technologies AG: Evaluation Board EVAL_500W_5G_PSU: 500 W Telecom Power Supply for 5G Small Cells Using 600 V CoolMOS™ G7 and CFD7 in DDPAK (Version 1.0); February 7, 2022; Available online

[6] Infineon Technologies AG: 1000 W Telecom Power Supply for 5G Edge Computing and Small Cells Using CoolSiC™ 650 V and CoolMOS™ 600 V CDF7 in TOLL Package; May 10, 2024; Available online

[7] Infineon Technologies AG: 1000 W Telecom Power Supply for 5G Edge Computing and Small Cells Using CoolGaN™ 600 V and CoolMOS™ 600 V CDF7 in TOLL Package; May 31, 2024; Available online

[8] Infineon Technologies AG: EVAL_3K3W_TP_PFC_SIC2: 3300 W CCM Bidirectional Totem-Pole with CoolSiC™ 650 V G2 and XMC™ (Version 1.0); July 24, 2024; Available online

[9] Infineon Technologies AG: 1 kW 48 V to 12 V Telecom Quarter-Brick Fixed-Frequency LLC with XDPP1100 Digital Controller; February 1, 2024; Available online

[10] Infineon Technologies AG: XDPE132G5H-G000 Digital Multi-Phase Controller Datasheet (Revision 1.0); Infineon Technologies AG; July 15, 2021; Available online

[11] Infineon Technologies AG: Multiphase Buck Converter with TLVR Output Filter; Infineon Technologies AG; May 21, 2025; Available online

[12] Infineon Technologies AG: Dual-Phase Operation in Multiphase Buck Applications (Version 1.0); Infineon Technologies AG; July 10, 2020; Available online

[13] Infineon Technologies AG: 12 kW BBU for AI Data Centers: REF_48V_12KW_BBU_PPC_Si; Infineon Technologies AG; Available online

[14] Light Reading: Telecom Aims for 99.999% Reliability. Telecom Policy Does Not; Light Reading; April 9, 2025; Available online

[15] AI-RAN Alliance: Working Groups and Committees; https://ai-ran.org/working-groups

[16] Nokia: Nokia Defines the Next Era of Radio with the Industry's First AI-Native RAN Platform; Available online

[17] Nokia: AI-RAN; https://www.nokia.com/radio-access/ai-ran/

[18] Ericsson: Ericsson Mobility Report; June 2026; Available online

[19] Micron Technology, Inc.: GDDR6 Graphics Memory; Available online

[20] PCI-SIG: PCI Express Card Electromechanical Specification, Revision 6.0.1; Available online

[21] Infineon Technologies AG: REF-3K3W-HFHD-PSU Evaluation Board; Available online

[22] Rectronic: R830H 3.0 kW 48 V High-Efficiency DC Modular Rectifier; Available online

[23] Huawei Technologies Co., Ltd.: R4850G2 Rectifier User Manual; Version 1.4; Available online

[24] R Vercellino, J Willard, G Campos, WS Pereira, O Hull, M Selensky, J Mueller: Measurement of Generative AI Workload Power Profiles for Whole-Facility Data Center Infrastructure Planning; Available online

Published by
Infineon Technologies AG
Am Campeon 1-15, 85579 Neubiberg
Germany

© 2026 Infineon Technologies AG.
All rights reserved.

Public

Version: V1.0_EN
Date: 08/2026

Stay connected!

Scan QR code and explore offering
www.infineon.com

Please note!
This Document is for information purposes only and any information given herein shall in no event be regarded as a warranty, guarantee or description of any functionality, conditions and/or quality of our products or any suitability for a particular purpose. With regard to the technical specifications of our products, we kindly ask you to refer to the relevant product data sheets provided by us. Our customers and their technical departments are required to evaluate the suitability of our products for the intended application.

We reserve the right to change this document and/or the information given herein at any time.

Additional information
For further information on technologies, our products, the application of our products, delivery terms and conditions and/or prices, please contact your nearest Infineon Technologies office (www.infineon.com).

Warnings
Due to technical requirements, our products may contain dangerous substances. For information on the types in question, please contact your nearest Infineon Technologies office.

Except as otherwise explicitly approved by us in a written document signed by authorized representatives of Infineon Technologies, our products may not be used in any life-endangering applications, including but not limited to medical, nuclear, military, life-critical or any other applications where a failure of the product or any consequences of the use thereof can result in personal injury.