High Current, High Power-Density Intermediate Bus Converters for Vertical Power Delivery to Next-Generation Processors

Pranav Raj Prakash, *Graduate Student Member, IEEE*, Ahmed Nabih, *Member, IEEE*, Yan Liang, *Student Member, IEEE*, Sudhir Kudva, *Member, IEEE*, Mostafa Mosa, C. Thomas Gray, *Senior Member, IEEE*, and Qiang Li, *Senior Member, IEEE*

***Abstract*—Advancements in artificial intelligence and machine learning are driving the development of a new generation of extremely powerful graphics processing units (GPUs) with exponentially increasing transistor counts, requiring larger die sizes and more power. Today, employing the conventional two-stage lateral power delivery with a 12 V intermediate bus voltage (IBV) requires most of the real estate in the GPU cards to be occupied by power delivery circuits. However, this is impractical for the next-generation GPUs with larger sizes and higher power consumptions, resulting in the interest to transition to vertical power delivery, where a lower IBV is used to reduce the size of the voltage regulators and move them vertically underneath the GPU. However, this increases the required step-down ratio for the first-stage intermediate bus converter (IBC), making it challenging to maintain high efficiency while constrained to a limited footprint. This article proposes a high-density transformer unit-cell for the first-stage LLC-based IBC, one or more of which can be connected to achieve different IBVs. To demonstrate an implementation of the transformer, an 840 W 48 V/1.8 V LLC converter is designed with an array of 14 transformers built into seven unit cells, distributed along the edge of the GPU package to minimize the power distribution networkPDN losses. The converter module can achieve up to 95.5% efficiency while showcasing a power density of 2200 W/in^3.**

***Index Terms*—Graphics processing unit (GPU), high-density dc–dc, LLC converter, planar magnetics, vertical power delivery.**

I. Introduction

IN RECENT decades, power management for data centers and telecom applications has received significant attention due to its exponential increase in market growth and electricity demand. By 2030, the total electricity demand for information and communication technology is expected to increase to approximately 20% of the projected global energy demand, with hyperscale data centers accounting for 40% of this energy consumption [1]. In the past decade, the common practice was to employ a 12 V bus architecture to power the compute trays, which was adequate for the lower power levels. However, the recent boom in energy demand has led to an overhaul of the data center power architecture, with the bus voltage to the server racks increased to 48 V to reduce power delivery network (PDN) losses and conversion stages [2]. As a result, many companies such as Facebook and Google have transitioned to the 48 V bus architecture to reduce electrical conversion losses by up to 30% [3].

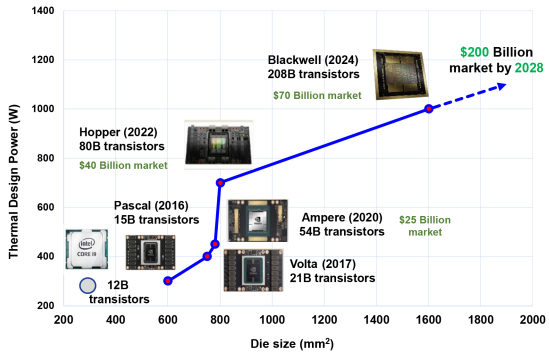


Fig. 1. Power and die size trends of NVIDIA's datacenter GPUs.

The two main options for general-purpose computing are central processing units (CPUs) and graphical processing units (GPUs). Fig. 1 shows the transistor count, power consumption, and die size of the Intel Core i9 CPU compared to the latest high-performance NVIDIA GPUs [4], [5], [6]. Even the newest CPU does not possess enough computing power to handle super-charged productivity, speed up workflows, and unlock the full potential of AI/ML that data scientists are striving for. Therefore, ultra-powerful GPUs are required to accelerate the pace of innovation. The increasing transistor counts have also brought increasing thermal design powers (TDP) and die sizes over the recent generations of GPUs. This is especially evident in the most recent Blackwell GPU, which increases the transistor count, die size, and TDP over the previous generation by 160%, 200%, and 33%, respectively. Consequently, the global market

Received 29 December 2024; revised 5 April 2025; accepted 7 May 2025. Date of publication 23 May 2025; date of current version 5 August 2025. Recommended for publication by Associate Editor M. Monfared. *(Corresponding author: Pranav Raj Prakash.)*

Pranav Raj Prakash, Yan Liang, and Qiang Li are with the Center for Power Electronics Systems (CPES), Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA (e-mail: pranavraj@vt.edu).

Ahmed Nabih and C. Thomas Gray are with the NVIDIA Corporation, Durham, North Carolina 27713 USA.

Sudhir Kudva and Mostafa Mosa are with the NVIDIA Corporation, Santa Clara, California, CA 95051 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPEL.2025.3572841.

Digital Object Identifier 10.1109/TPEL.2025.3572841

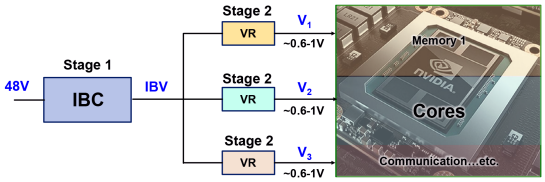


Fig. 2. Two-stage power delivery to the different domains in a GPU.

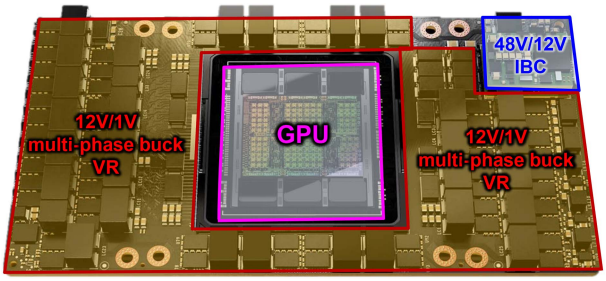


Fig. 3. Lateral power delivery to NVIDIA's Hopper GPU.

for GPUs is also expected to increase to $477 billion by 2030 [7], indicating the need for faster and more powerful chips. However, the higher power requirements for the next-generation GPUs, albeit the increasing compute efficiencies, pose a big challenge for power management systems.

The GPU chip has multiple independent frequency domains such as memory and core (graphics), where each requires a precise frequency and voltage to operate optimally, usually between 0.6 V to 1 V. Moreover, these domains have different load transient requirements. Therefore, as shown in Fig. 2, a two-stage power delivery system is preferred to convert the 48 V dc entering the server rack to the 0.6 V–1 V dc required by the GPU. The commonly implemented power architecture has the first stage as an unregulated intermediate bus converter (IBC) that steps down the 48 V to an intermediate bus voltage (IBV) with high efficiency. As the second stage, different voltage regulators (VRs) provide the various GPU domains with precisely regulated voltage and frequency while dealing with their rigorous transient requirements.

In addition to the losses of the two power stages, the total efficiency of the two-stage power delivery is also impacted by the PDN losses between stage 1 and stage 2 and between stage 2 and the GPU load. A higher IBV would result in a lower bus current I_{bus}, and hence a lower $P_{PDN} = I_{bus}^2 R_{PDN}$ for the same power. Fig. 3 shows the power delivery architecture of NVIDIA's Hopper GPU, which requires around 700 W of power [6]. Here, an IBV of 12 V is implemented, with a 48 V/12 V IBC as the first-stage (area in blue), followed by 12 V/1 V multiphase buck VRs as the second-stage (area in red) [8]. Since the two stages are cascaded on the same side of the GPU card, it is called lateral power delivery (LPD). As can be noted, the 1st-stage 4:1 IBC only occupies a small area in the corner (blue box), whereas the 60 phases of the second-stage VRs require most of the real estate on the GPU card due to the large sizes of the discrete inductors needed for the 12 V/1 V VRs.

TABLE I
VOLTAGE REGULATORS WITH DIFFERENT IBVS

Parameter	MPC22163	Fe1736
IBV(V)	10 – 16	1.6 – 2
Typical V_o(V)	0.8	0.8
Maximum I_o(A)	130	56
f_{sw}(MHz)	0.7	75
Bandwidth (MHz)	0.2 MHz–0.3 MHz*	20 MHz
Dimensions (mm)	9 × 10 × 7.65	5.64 × 3.54 × 0.6
Current density (A/mm^2)	1.3	2.8

*estimated

An example of a 12 V/1 V VR is MPC22163 from MPS, as summarized in Table I. It typically switches at less than 1 MHz for high efficiency, resulting in a low bandwidth, a large inductor size and high output capacitance to meet the steady-state and transient requirements. Although having a 12 V bus can reduce the PDN loss, the large inductor sizes are unsuitable for future-generation GPUs, which have a much larger die size and require higher power, as seen in Fig. 1. This is a bottleneck with the existing lateral power delivery to supply next-generation GPUs effectively.

This article provides a more in-depth and generalized analysis of the design and optimization process of the high-density IBC, building upon the ideas presented in a previously published conference paper [9].

The rest of this article is organized as follows. Section II explores the innovations in the power delivery architecture to the GPU that require increasingly high powers and the required form factor of the IBC to minimize the PDN loss. Section III describes the topology selection of the LLC-based bus converter. Section IV describes the design of a high-density transformer unit-cell that can be cascaded to modify the IBV easily. Section V describes the design guideline for a 48 V/1.8 V IBC using the proposed unit-cell. Section VI presents the hardware prototype and testing of the 48 V/1.8 V IBC and compares its performance with estimations of IBCs with other IBVs using the same transformer unit-cell. Finally, Section VII concludes this article.

II. VERTICAL POWER DELIVERY TO GPUS

The limitations of the existing LPD for future GPUs requiring higher power demand a paradigm shift in the power delivery architecture. The PDN losses arising from the high load current can be reduced by placing the VRs vertically underneath the GPU, hence limiting their distance to the thickness of the system-on-chip (SoC) card and the GPU substrate. This method, often labeled as vertical power delivery (VPD), can be implemented for both single-stage and two-stage power delivery solutions.

Fig. 4(a) illustrates the VPD implementation for a single-stage VR, as described in [10] and [11]. The mini-LEGO PoL converter [10] integrates a switched-capacitor (SC) stage and a multiphase buck stage with a floating bus, vertically stacking them in a power module housed beneath the GPU. The switching bus converter in [11] vertically stacks a SC front-end stage with two series-capacitor-buck (SCB) back-end stages, utilizing

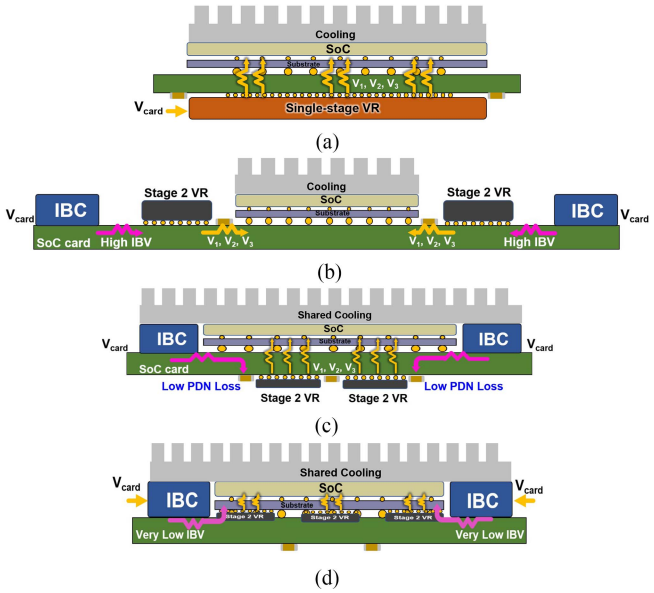


Fig. 4. (a) VPD with single-stage power delivery, (b) LPD with high IBV and large VRs, (c) reducing the IBV, moving the VRs vertically underneath the GPU, and moving the IBCs closer to GPU to reduce PDN loss, and (d) possible arrangement with very low IBVs to utilize ultralow profile IVRs that fit directly underneath the substrate.

switching buses to merge them. These single-stage VRs offer appealing options, as they eliminate the power distribution network (PDN) losses that must be accounted for in two-stage solutions. However, the operating frequency of the single-stage VRs is restricted to under 2 MHz to maintain optimal efficiency ([11] runs at 600 kHz while [10] runs at 1.5 MHz), thereby limiting the control bandwidth. In addition, the profile constraints on the bottom side of the SoC card (3–5 mm) restrict the peak load capability and, consequently, the current densities (0.6–0.7 A/mm^2) achievable, which is insufficient for future GPUs. Furthermore, GPUs typically implement liquid cooling on the top side of the SoC card. While a few bottom-side cooling solutions do exist, the limited mechanical clearance and component density often restrict the effectiveness of thermal management for single-stage VRs on the underside. Consequently, splitting the high voltage step-down and regulation into distinct stages demands further investigation to address the previously mentioned limitations associated with single-stage VRs.

The transition from LPD to VPD for a two-stage power conversion is described in Fig. 4. Fig. 4(b) shows the existing LPD with a high IBV and large VRs that hinder the GPU size and power expansion. Similar to the single-stage power delivery, the VRs must be moved to the bottom side of the SoC card, vertically underneath the GPU to accommodate the increasing size of the GPU die and enable VPD, as shown in Fig. 4(c). This also reduces the distance, and hence the PDN resistance from the VRs to the load. However, the profile constraints on the bottom side are very tight (3–5 mm) for the legacy 12 V VRs. Although some innovations, such as a printed circuit board (PCB) embedded coupled-inductor structure [12], have been made, the low current densities ($0.5A/mm^2$) are not adequate for the high-current GPUs. Although improvements to the magnetic design may reduce the profile of the VRs [13], a pursuit of higher efficiencies and power densities is leading the industry away from the 12 V IBV to lower voltages [14].

A lower IBV (e.g., 6 V, 3.3 V, 1.8 V) increases the duty cycle required by the buck VRs for the same output voltage. Therefore, for a fixed inductance, the peak inductor current, and hence the device switching current, decreases to maintain charge balance. The lower switching current, coupled with the lower input voltage, reduces the device switching loss considerably, thereby facilitating the safe increase of switching frequencies. This results in smaller inductor sizes due to lower peak magnetic flux. Moreover, with the recent development of new low-voltage devices with higher figures of merit (FOMs) and integrated gate drivers (DrMOS), the efficiency and densities can be improved [14]. An example of low IBV VRs is Fe1736 from Ferric, a 1.8 V/0.8 V VR as summarized in Table I. Thanks to the low IBV, switching the frequency can be increased to over 50 MHz, resulting in a smaller size, higher bandwidth, and reduced capacitance required for output regulation.

However, reducing the IBV increases the PDN loss between the first and the second stages. For example, with a $R_{PDN} = 300\ m\Omega$ for a 1000W GPU, reducing the IBV from 12 V to 1.8 V would increase the PDN loss by 44 times to around 100W, which would drastically reduce the total efficiency by 10%. Therefore, the first stage IBCs must be moved closer to the GPU to minimize the PDN resistance, as shown in Fig. 4(c). This also enables the IBCs to share the liquid cooling thermal management system for the GPU, hence reducing the complexity and cost. Furthermore, with current and future innovations in ultra-low profile (<1 mm), ultra-high frequency (>100 MHz) integrated voltage regulators (IVRs), they can be pushed closer to the GPU by embedding directly in the GPU substrate, thereby further minimizing the PDN losses and improving overall system efficiency [15], [16], [17]. Such high frequencies also virtually eliminate the requirement for output filter capacitors, further reducing the total real estate required for power delivery.

In summary, a two-stage VPD architecture can effectively be implemented for high-current GPUs, but the IBV must be reduced from the traditional 12 V to ensure high current density and bandwidth. However, the optimal IBV to maximize overall system efficiency is still an ongoing research problem since it heavily influences the IBC, PDN, and VR losses. While research on all three fronts is ongoing, this work focuses on proposing a versatile, high-density design for the first-stage IBC that can easily be scaled to accommodate various low IBVs. The transition to VPD also instills a unique design challenge on the IBCs, wherein they must have: 1) a narrow width to fit between the GPU and the edge of the GPU card; 2) a wide length to distribute and supply the output current evenly and provide a strong connection to the VRs to minimize the PDN loss [4]; and 3) a top-side cooled power module with all power devices on the top layer.

III. LLC-Based Intermediate Bus Converter

This section will discuss the selection of the preferred topology, devices, and resonant tank parameters.

A. Topology Selection

To maximize the efficiency while conforming to the high power density requirements, the choice of converter topology is important [18]. Capacitor-based switched tank converters (STCs) can achieve high efficiencies and power densities with their ability for zero current switching (ZCS) and hence are potential candidates for this solution [19], [20]. However, an additional buck converter is required for startup and protection. Moreover, due to multiple resonant branches, it is sensitive to component tolerances and requires a special controller to adaptively tune the operating frequency [21]. Besides component tolerances, the large number of components connected in series leads to significant conduction losses as the current increases. This constrains the STC from extending to lower output voltages and higher output currents, where additional resonant branches are necessary to achieve a higher conversion ratio. Recently, hybrid resonant switched tank-based converters were proposed, combining the switched-capacitor networks with multitapped autotransformers to reduce the number of stages and the component count, improving the efficiency [22]. However, due to the low output voltage and low turns ratio, the effective peak magnetizing current i_Lm to charge/discharge the device C_oss is low. Hence, the converter must be switched at $f_s > f_o$ to increase the turn-off current, resulting in zero voltage switching (ZVS) loss at low to medium loads.

Another well-reported topology is the unregulated LLC converter (DCX) [23], [24]. Due to the higher turns ratio n, the LLC DCX can achieve soft-switching at all load conditions and hence can be operated at high frequencies with high efficiency and power density. It is also less sensitive to component tolerances since there is only one resonant branch to be tuned, where the transformer's leakage inductance L_k is utilized as the resonant inductance L_r and Class I capacitors are used for the resonant capacitance C_r to stabilize the resonant frequency f_o across temperatures and voltages. Moreover, it is highly scalable and can easily be extended to lower IBVs by simply changing the turns ratio n.

However, for such high-current transformers, several significant challenges must be addressed, including paralleling multiple secondary rectifiers (SRs) and secondary windings to distribute the high secondary current and placing the multiple parallel SRs close to the termination to mitigate the secondary winding leakage inductances. To address these issues, the concept of matrix transformers can be implemented, wherein the output current can be divided among several smaller elementary transformers whose primary windings are in series, and the output sets are connected in parallel, thereby distributing the high current and facilitating easier SR termination, as shown in Fig. 5, [25].

B. Primary Inverter Configuration

The inverter on the primary side could either be of a full-bridge (FB) or a half-bridge (HB) configuration as shown in Fig. 5 and whose differences are summarized in Table II. Although FB inverters reduce the losses per device, the device number and transformer turns ratio are higher, resulting in higher costs and added complexities in the transformer design, especially for such high step-down applications. The HB inverters would be preferred at low power levels due to lower driving loss and easier transformer winding layout. However, the devices would reach their thermal limit at higher power levels, requiring the switch to FB rectifiers.

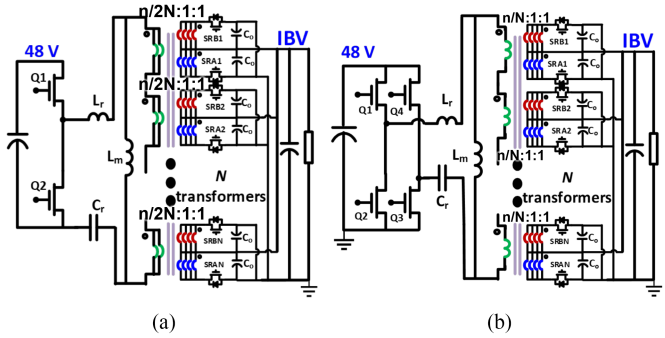


Fig. 5. n:1:1 LLC converter with N transformers using (a) half-bridge primary with $n/2N$ turns for each transformer, and (b) full-bridge primary with n/N turns for each transformer.

TABLE II
PRIMARY INVERTER CONFIGURATION

Configuration	Half-bridge	Full-bridge
Number of devices	2	4
Average resonant tank voltage	$V_{in}/2$	V_{in}
Turns ratio, n for unity gain at $f_s = f_o$	$V_{in}/(2V_o)$	V_{in}/V_o
Primary resonant current, i_{Lr} (RMS)	$2I$	I
Primary device conduction loss	$8I^2R_{ds_{on}}$	$4I^2R_{ds_{on}}$
Primary device driving loss	$2Q_gf_s$	$4Q_gf_s$

TABLE III
PRIMARY DEVICE COMPARISON

Technology	GaN	Si	
Part Number	EPC2302	NTMFS6H800NL	
$V_{br}(V)$	100	80	
$C_{oss}(pF)$	1000	800	
$V_{dr}(V)$	5	5	10
$Q_g(nC)$	23	60	112
$R_{ds_{on,max}} at 25^0C(m\Omega)$	1.8	2.4	1.9
$R_{ds_{on,max}} \cdot Q_g(p\Omega C)$	41.4	144	212.8
$R_{ds_{on,max}} \cdot C_{oss}(f\Omega F)$	1800	1920	1520

To determine the maximum power the devices can handle in an HB configuration, selecting the best device is paramount. Although Si MOSFETs have been traditionally preferred in this voltage range, recent advances in GaN technology have resulted in low on-resistance GaN devices with better figures of merit (FoMs) than the best Si devices, as highlighted in Table III. As expected, the very low gate charges Q_g and the low on-resistances $R_{ds_{on}}$ provide a lower FoM to the GaN devices than the Si counterpart. Although the $R_{ds_{on}}$ of the Si device can be reduced by increasing the driving voltage V_{dr} to result in a lower switching-loss FoM $R_{ds_{on,max}} \cdot C_{oss}$, it increases Q_g significantly, thereby paying the price of higher driving loss and light-load efficiency. Therefore, the 100V GaN FET can be selected as the primary device.

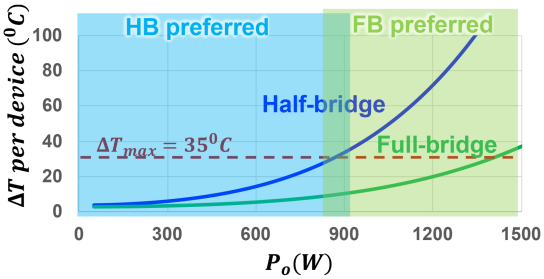


Fig. 6. Preferred rated powers with HB and FB inverters.

TABLE IV
TRANSFORMER AND SR SPECIFICATION AT DIFFERENT IBVS

IBV	6.25V	3.1V	1.8V
n (half-bridge)	4:1	8:1	14:1
n (full-bridge)	8:1	16:1	28:1
SR device	IQE006	IQE004	
$V_{br}(V)$	25 V	15 V	
$R_{ds_{on,max}}(m\Omega)@5V$	0.975	0.74	
$I_{RMS,max}$ for $\Delta T_c = 35^0C$	28	32	

The thermal limit for the devices can be estimated based on the junction-to-ambient thermal resistance, $R_{th,J-A}$. which can either be obtained from device data sheets for a rough approximation, calculated using tools like [26] for a more accurate estimation, without having to run time-consuming finite element analysis (FEA) simulations. For this evaluation, the low-profile IBCs placed close to the GPU share the same liquid cooling for thermal management. In addition, the primary devices have bottom-side cooling with thermal vias, and the power module is installed with an aluminum heat-spreader to reduce the thermal resistance from the case to the heatsink, as implemented in [27]. Specifically, a thermal interface material (TIM) with a thermal conductivity $k = 17.8$ W/m $\cdot K$, a $1-$ mm thick aluminum heat-spreader, and a convection coefficient $h = 100$ W/m^{2}K is used to result in an $R_{J-A} = 12.2\,^\circ$C/W. It is noteworthy that R_{J-A} would change with heatsink dimensions and flow speed and only provides a rough estimate for choosing the primary inverter configuration or number of devices required in parallel.

Based on this, the approximate temperature rise of the primary devices with increasing rated power can be evaluated for the HB and FB configurations. Assuming negligible dead time and operation when switching frequency f_s equals the series resonant frequency f_o, the device losses can be calculated as described in [28], and the temperature rise can be plotted for the devices in both HB and FB inverters. Assuming an ambient temperature of 60 °C–65 °C close to the GPU, a maximum temperature rise of 35 °C is permitted for safe operation. Based on this analysis, Fig. 6 shows the recommended power ranges for HB and FB inverters, where HB inverters would offer lower device count and easier transformer winding layout for less than 850W, whereas FB inverters would be required for thermal management for higher power levels. Alternatively, multiple LLC converters with HB inverters can be paralleled to increase the total delivered power.

C. Secondary Rectifier Configuration

The choice between center-tap (CT) and FB rectifiers comes down to the trade-off between requiring two devices with a breakdown voltage of at least $2V_o$ (for CT rectifiers) or four devices with a breakdown voltage of at least V_o (for FB rectifiers). For such low V_o applications, CT rectifiers are a better choice than FB rectifiers since a device of slightly higher, if not similar, breakdown voltage can be used even for a voltage stress of $2V_o$. Although GaN devices have recently started emerging in the low voltage domain, Si MOSFETs still have the better FoM at the time of this article, and hence are chosen as the synchronous rectifiers (SRs). Table IV shows the selected SRs for the different IBVs considered in this article due to their low $R_{ds,on}$ small 3.3 mm $\times$ 3.3 mm package and pin-to-pin compatibility across IBVs.

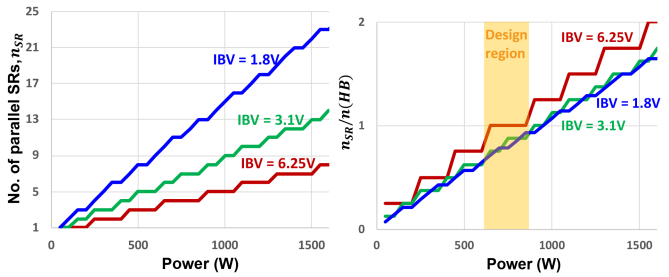


Fig. 7. (a) Number of parallel SRs required to limit $\Delta T_c = 35\,^\circ$C for different IBVs with increasing output power, and (b) number of parallel SRs required per transformer turn with a HB inverter; each 1:1 transformer would require one SR in parallel in the design region.

The number of parallel SRs required depends on their thermal capabilities. However, unlike the primary devices, Since the SRs are placed on the PCB windings, only case-side cooling is possible, and hence $R_{J-A} = 46.4\,^\circ$C/W is obtained using the same setup as the primary devices. Assuming negligible driving loss, the maximum RMS current $I_{RMS,max}$ to limit the temperature rise to 35 °C can be calculated as listed in Table IV. Based on this, the number of parallel SRs required for different IBVs at increasing power levels can be determined, as shown in Fig. 7(a). The IBC modules described in this article are top-side cooled only; hence, all devices must be placed only on the top-side. This complicates paralleling multiple SRs on a single transformer. Therefore, it is preferred to have the same number of transformers as parallel SRs for modularity and ease of design. To determine the design range for the rated power, the ratio of the number of parallel SRs n_{SR} to the required turns number (half-bridge) is plotted against power for different IBVs in Fig. 7(b). The ratio is close to and less than one around 800W, so the power can be maximized without over-stressing the devices. The number of SRs and transformers would remain the same for FB inverters, but each transformer would have a turns ratio of 2:1 instead.

IV. High-Density PCB Transformer Module

Based on the analysis in the previous sections, a 1:1 (for HB primary) and 2:1 (for FB primary) transformer is required to construct the IBC module. However, with such a large number of elemental transformers to be connected, it is clear that optimizing the transformer design is key to maximizing the efficiency and power density of the converter. This includes considerations with the optimal number of PCB layers, type, and material of the transformer's core, core and winding dimensions, switching frequency, and the SR rectifier placement to maximize space utilization.

A. Transformer Unit-Cell With High Space Utilization

Each elemental transformer can be implemented using a UI/UU core, where the windings are wound around one leg, and the other leg is the magnetic flux return path. However, two such transformers can be wound inversely such that the magnetic flux in the return legs is equal and opposite. This way, the flux return legs can be removed and the two winding-carrying legs can be integrated into a single UI/UU core [29].

With increasing switching frequencies and reducing magnetic sizes, PCB-based planar magnetics are preferred over the traditional litz wire-based transformers due to easier manufacturability, strong repeatability to minimize leakage inductance mismatch, and a larger surface area resulting in low-profile cores and better thermal capabilities [30]. However, multiple winding layers must be paralleled for high-current applications to prevent overheating. Based on this, a 12-layer PCB is implemented in [29], where the primary and secondary layers are interleaved to minimize the magneto-motive force (MMF) build-up. The four secondary layers per CT turn are connected in parallel through vias and then connected to the SRs and capacitors which are placed outside the transformer footprint to complete the ac SR termination loop. This termination pattern may be referred to as lateral SR termination since the SRs and capacitors are on the same layer as one of the secondary layers and the SR termination loop. However, this results in asymmetrical current-sharing between the parallel secondary windings, where the secondary winding on the top layer got the most current due to its lowest impedance path to the SRs on the same layer. This causes increased winding losses and a large footprint.

To address this issue, [23] proposed a vertical SR termination technique, where two additional layers dedicated to housing the SRs and capacitors are added to the top and bottom of the PCB stack (14 total layers) and are connected to the parallel secondary layers with vias, as shown in Fig. 8(a). Fig. 8(b) shows the SRA and SRB termination loops of the two half-cycles, which connect the switching nodes SWA and SWB to the corresponding SRs and capacitors to return to the V_o node. Fig. 8(c), (d), and (e) show the secondary current flow in the two half-cycles through the SA and SB layers to the termination layers. Since all the parallel secondary layers are now connected to the SRs through vias, the impedance is more symmetrical, resulting in improved current-sharing.

The winding areas occupied by the SWA and SWB nodes are highlighted in green and pink, respectively. It may be noted

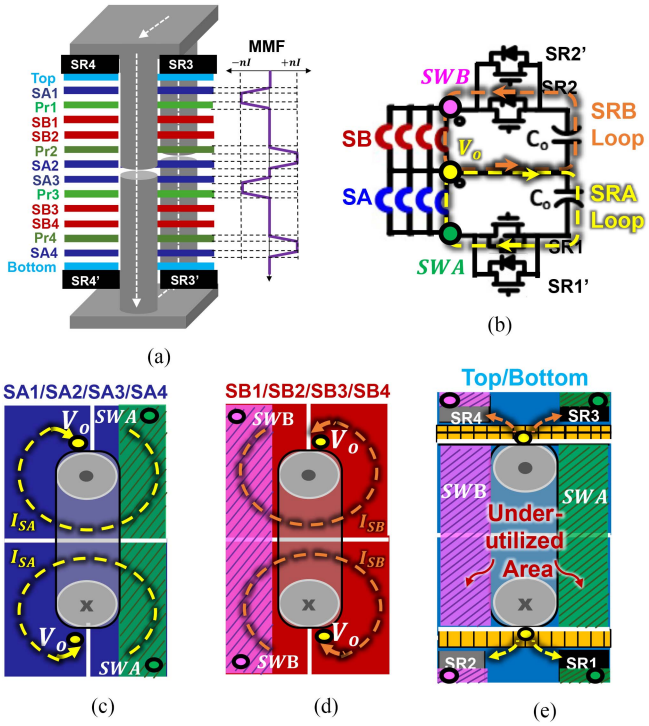


Fig. 8. Transformer design in [23]: (a) 14-layer PCB stack-up with top and bottom layers dedicated for SR termination, (b) SR termination loops in the two half-cycles, (c) SRA loop current flow and SWA node area, (d) SRB loop current flow and SWB node area, and (e) under-utilized area in the top/bottom termination layers.

that the switching nodes occupy half of each winding, although they only connect the winding terminal to the drain terminal of the device. This results in significant underutilized space on either side of the UI core [Fig. 8(e)], thereby limiting the achievable power density. Moreover, the capacitors and SR devices are arranged along the converter's width, which could require additional length to accommodate them. Therefore, although this termination method can improve the current sharing between the parallel secondary layers, poor space utilization is the bottleneck to maximizing its density.

Other works have attempted to solve some of these challenges. A similar design was implemented in [31], where the capacitors were moved to a separate submodule (double-layered vertical SR termination), and the devices are moved closer to reduce the length. However, the underutilization of the space between the cores persists, and the multiple-module system could result in complex implementation and packaging. The design in [32] utilized the transformer footprint effectively, with the capacitors and SRs all placed in a high-density form factor. However, the SRs are covered by the magnetic core, which would prevent proper thermal management of them, thereby limiting the maximum power that could be pushed. Furthermore, the core legs must be elongated beyond the SRs' profile, increasing the total module profile.

To address the challenges with the previous literature, this article proposes a high-density transformer unit-cell, as shown

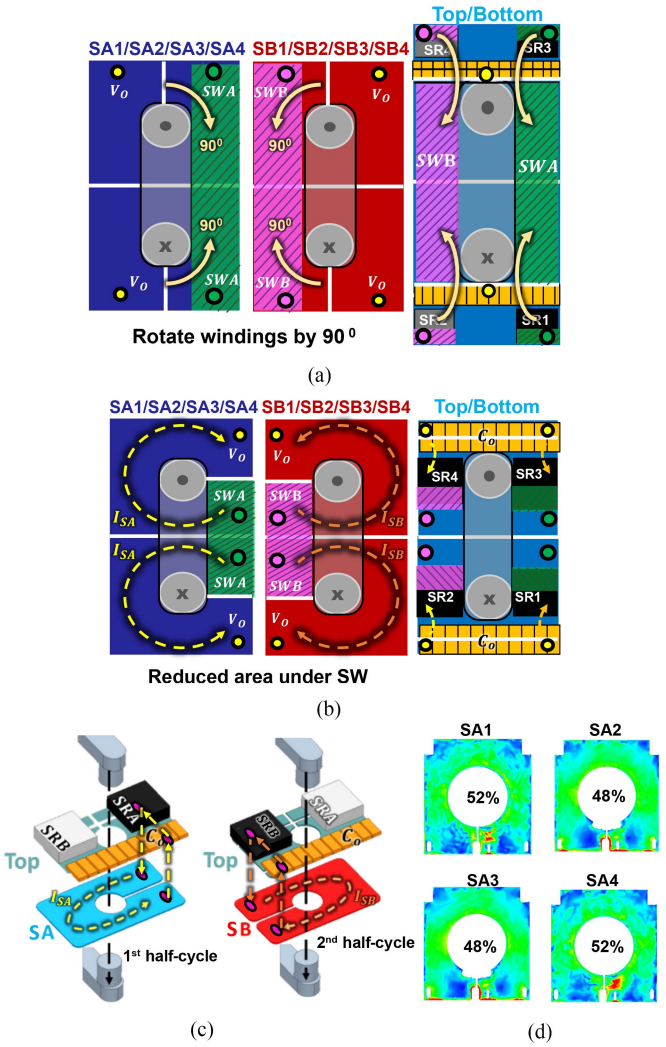


Fig. 9. Construction of proposed transformer unit-cell: (a) Rotating the secondary windings of [23] by 90°, rotating the SRs by 180° and moving them next to the core, (b) resulting transformer unit-cell with high space-utilization, (c) Secondary current flow in the two half-cycles, and (d) Current-sharing between parallel secondary layers.

in Fig. 9. Considering the 14-layer PCB stack-up and winding arrangement as [23] as the starting point in Fig. 9(a), the secondary windings are rotated by 90° toward the transformer core, such that the area under the switching nodes SWA and SWB are reduced by half, as shown in Fig. 9(b). Now, the SRs must be rotated by 180° and be placed adjacent to the transformer core to connect to the switching nodes. This now frees up space on the top and bottom of the transformer core to place the filter capacitors to complete the SR termination loop. This reinstates the single-layered vertical SR termination from [31], where a separate submodule is not required for the capacitors. This results in a high-density transformer unit-cell with an optimal space-utilization of the transformer footprint. Fig. 9(c) shows the secondary ac current flow during the two half-cycles. The symmetry ensures equal leakage inductance along the two paths, thereby ensuring symmetrical operations too. Furthermore, the proposed arrangement does not impact the

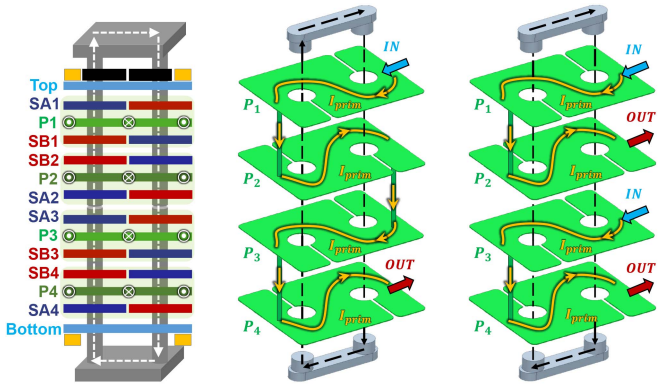


Fig. 10. "Serpentine" half-turn primary winding arrangement within a unit-cell: (a) PCB stack-up, (b) 2:1 ratio with all primary layers in series for FB inverter, and (c) 1:1 ratio with parallel layers P1/P3 in series with parallel layers P2/P4 for HB inverter.

current-sharing between the parallel secondary layers [Fig. 9(d)] since all the windings are still connected symmetrically through vias to the SRs.

B. Connecting Multiple Transformer Unit-Cells

The primary windings within the proposed unit-cell are arranged as shown in Fig. 10. Since each unit-cell consists of two elemental transformers, four and two primary turns per transformer for FB and HB inverters, respectively, are required. Using four primary layers [Fig. 10(a)], each layer can be implemented as a half-turn twisted to change direction from one leg to the other. This way, four primary turns can be implemented by connecting the four layers in series using limited vias only at the two ends of the unit-cell, as shown in Fig. 10(b) [33]. Similarly, to implement two turns for the HB inverter, the primary current flows in through layers P1 and P3 in parallel and flows out through layers P2 and P4 in parallel, as shown in Fig. 10(c). Interleaving the input and output paths also helps minimize the leakage inductance. Furthermore, the current densities of the primary layers would ideally be similar in Fig. 10(b) and (c) since the $2I$ current would be distributed between the two parallel layer sets.

To implement the large turns ratio required by the IBC, the proposed unit-cells can easily be connected and arranged in a linear array to minimize the total width of the converter module and distribute the output current along its length. Moreover, certain improvements are made to reduce the winding loss. The "serpentine" primary winding structure can be extended to connect multiple unit-cells, as shown for two unit-cells in Fig. 11(a) [33]. The secondary windings can also be arranged by interleaving the SA and SB layers for symmetrical leakage inductance and connecting the outputs in parallel. The two switching and output nodes SWA, SWB, and V_o are also labeled, where it can be observed that the V_o nodes between the two unit-cells are adjacent to each other. This means that the secondary layers can be merged between two unit-cells, thereby reducing the current density in that area [34]. Furthermore, the twisting in the primary

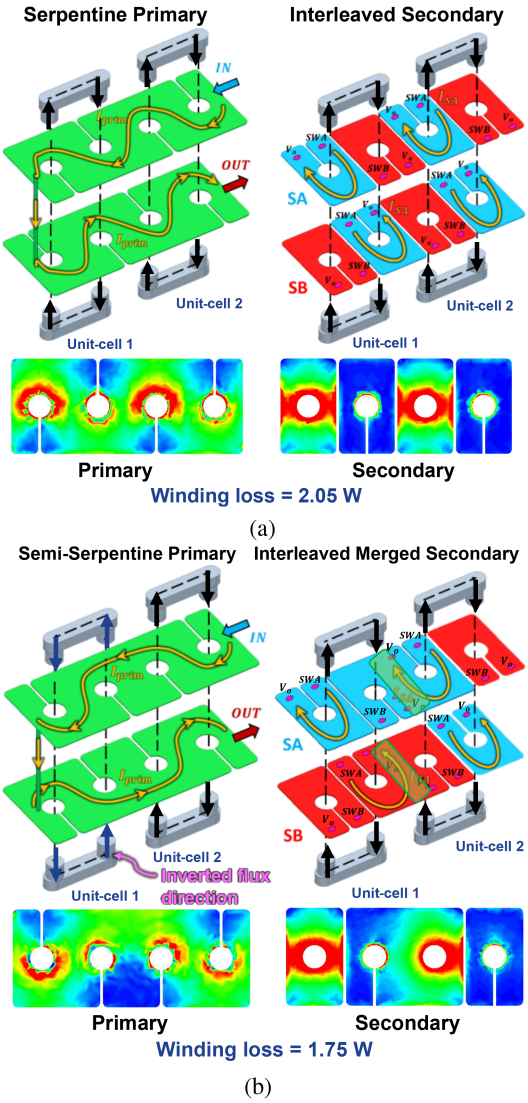


Fig. 11. Winding arrangement and current densities connecting two unit-cells: (a) "Serpentine" primary and interleaved secondary and (b) "semi-serpentine" primary and merged secondary between unit-cells.

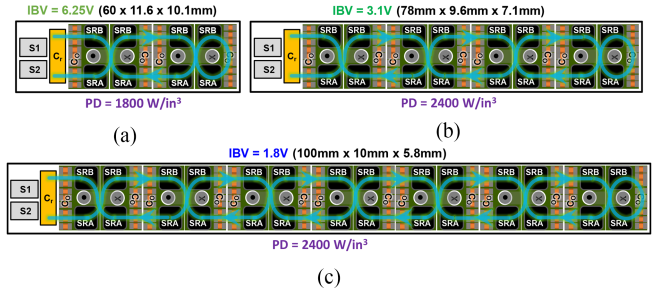


Fig. 12. Unit-cell-based IBCs for different IBVs, their estimated dimensions and power densities: (a) 2 unit-cells for IBV = 6.25 V, (b) 4 unit-cells for IBV = 3.1 V, and (c) 4 unit-cells for IBV = 1.8 V.

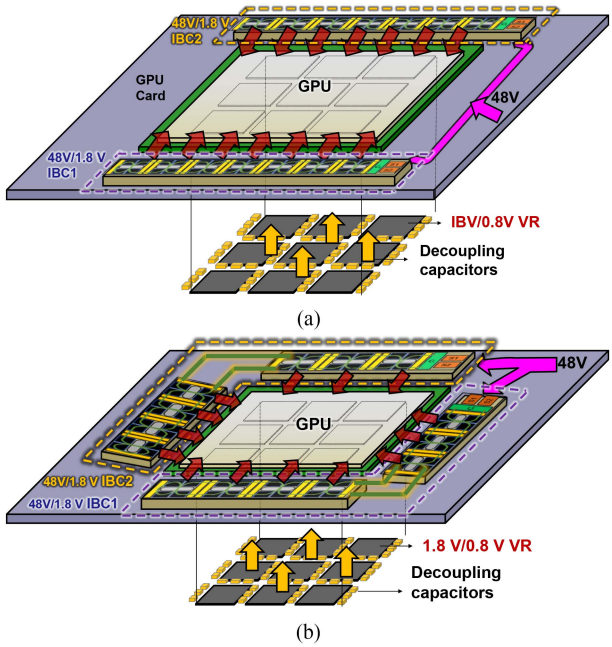


Fig. 13. Possible arrangements of two parallel IBCs around the GPU: (a) Each IBC is a single module on either side of the GPU, and (b) Each IBC is two modules connected in series and placed around the GPU.

windings can also be reduced by switching from a "serpentine" winding structure to a "semi-serpentine" structure, as shown in Fig. 11(b). To achieve this, The flux direction of the adjacent unit-cell is inverted, thereby resulting in the straightening of the primary winding between unit-cells. The improvements in winding arrangement help reduce the winding loss by 15%.

Based on this concept, IBCs of various IBVs can easily be designed by connecting multiple 2:1 transformer unit-cells in series. For a given output power, a larger unit-cell number increases the transformer turns ratio, hence reducing the IBV. Simultaneously, more SRs are added in parallel to distribute the increasing output current evenly along the module length, thereby highlighting the scalability of the proposed design. Fig. 12 shows conceptual designs of HB LLC-based IBCs for 6.25 V, 3.1 V and 1.8 V IBVs, which require 2 (4:1 ratio), 4 (8:1 ratio), and 7 (14:1 ratio) unit-cells, respectively. The estimated dimensions and power densities are also provided, highlighting the small size and high power density possible with the proposed transformer design.

V. Intermediate Bus Converter for 1.8 V IBV

To demonstrate the scalability of the proposed transformer design and utilizing the ultra-low profile, ultra-high bandwidth 1.8 V-0.8 V VRs, a low IBV of 1.8 V is chosen, and an 840 W 48 V:1.8 V HB LLC converter is designed. The 14:1 turns ratio is implemented with 14 1:1 transformers, each delivering 60 W, and implemented by connecting seven transformer unit-cells as demonstrated in the previous section. For future GPUs requiring >1500 W of power, [5], two such modules may be paralleled and, based on the GPU size and available space around the die, may be arranged in various configurations to minimize the PDN loss thanks to the modularity of the transformer unit-cell. Fig. 13(a) shows one arrangement where each module contains all seven unit-cells and is placed on either side of and close to the GPU to provide the IBV symmetrically to the IVRs with

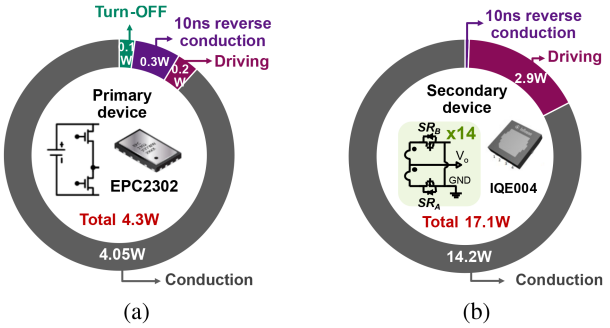


Fig. 14. Loss evaluation at 600 kHz, 840 W for the (a) primary and (b) secondary devices.

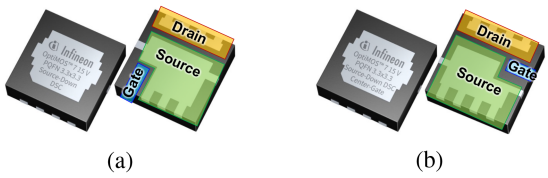


Fig. 15. Infineon 15 V device with different gate-pin locations: (a) IQE004NE1LM7SC and (b) IQE004NE1LM7CGSC.

minimum PDN loss. Alternatively, depending on the GPU size, the IBCs can be separated into multiple modules, wherein three unit-cells and the HB inverter on one module are connected in series with the remaining four unit-cells on another module, as shown in Fig. 13(b). The intermodular connection would be made on the primary high-voltage side, thereby incurring negligible i^2R losses. This highlights the system-level versatility of the proposed transformer unit-cell. For this article, however, a single power module with seven unit-cells is designed. Since the transformer largely dominates the size and loss of the converter, its design is paramount in optimizing the converter's performance.

A. Device Selection

Based on the analysis in Section III, the primary HB inverter uses the 100 V GaN FETs EPC2032. In addition to the low FoM, a 3 × 5 mm package combined with an exposed top for top-side thermal management makes it an excellent fit for the converter. Fig. 14(a) shows the primary device loss breakdown at full load, which adds up to 4.3 W.

Similarly, the secondary CT rectifiers employ the 15 V Si MOSFET, which offers a low FoM and a top-cooled package with a small 3.3 mm × 3.3 mm footprint. Fig. 14(b) shows the SR loss breakdown of all 28 devices at full load, which adds up to 17.1 W. Moreover, it is available with both center and side gate-pin configurations, as shown in Fig. 15, helping to minimize the gate loop inductance and easing the layout.

B. Transformer Core Material Selection At High $B_{\max}$

It is imperative to select the right Mn-Zn ferrite material for the transformer's core to minimize the core loss and, hence, the transformer loss. For such high-density transformers, the limited available area results in the core suffering higher peak

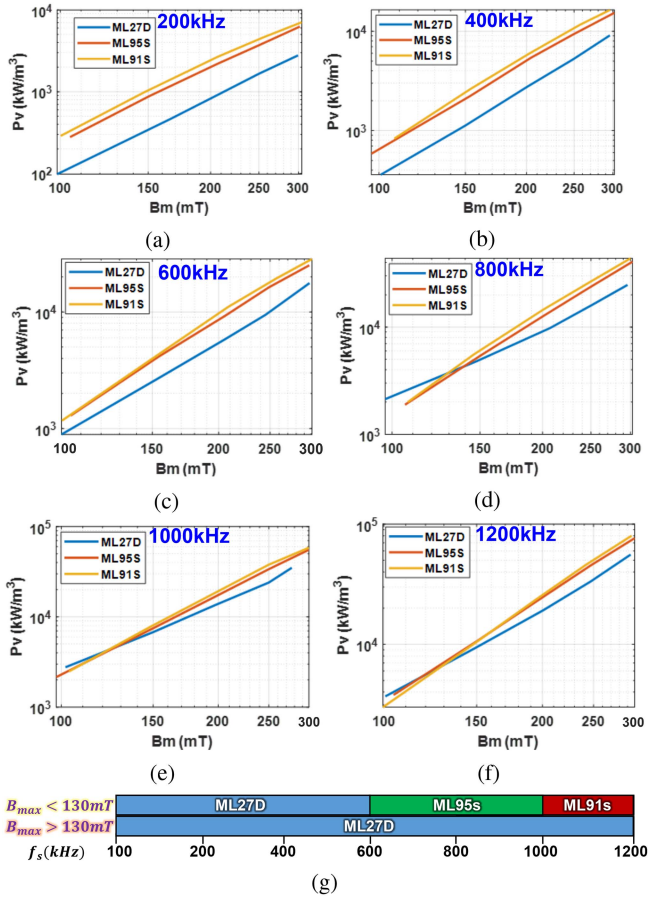


Fig. 16. Measured core loss density P_v of toroidal samples of three Proterial materials at high $B_{\max}$ conditions: (a) 200 kHz, (b) 400 kHz, (c) 600 kHz, (d) 800 kHz, (e) 1000 kHz, (f) 1200 kHz, and (g) summary of optimal materials.

magnetic flux density ($B_{\max}$) than what is provided in their data sheets. For example, the Mn-Zn ferrite materials from Proterial, ML27D, and ML95s are recommended to be operated at $B_{\max} < 150$ mT, and ML91s at $B_{\max} < 80$ mT to provide the lowest core loss density P_v at their recommended frequency ranges [35]. However, no information is provided for high $B_{\max}$ conditions (less than the saturation flux density B_{sat}), which would be commonplace for high-density magnetics. Therefore, the P_v of toroidal core samples of the three materials (OD = 9.9 mm, ID = 5.9 mm, H = 3 mm) is measured at different frequencies up to $B_{\max} = 300$ mT using the partial-cancellation method described in [36], as shown in Fig. 16. Table V summarizes the measurement results, showing that the material recommendations essentially hold in the recommended $B_{\max}$ operating range. However, at high $B_{\max}$ conditions, the P_v of ML95s and ML91s increase significantly, resulting in ML27D providing the lowest P_v from 200 kHz–1.2 MHz. Therefore, this material is used for the transformer core in this article.

The core loss is calculated using Mu's model for rectangular ac voltages [37] as

$$P_{core} = \frac{8}{\pi^2} k (B_{\max})^{\beta} f_s^{\alpha} V_{core} \tag{1}$$

TABLE V
LLC CONVERTER SPECIFICATIONS

Parameters	Value
Input Voltage	46 V-52 V
Output Voltage	1.6–1.9 V
Rated Power	840 W
Resonant Frequency	660 kHz
Turns Ratio	14:1
Primary Device	EPC EPC2302
Secondary Device	Infineon IQE004
Core Material	Proterial ML27D
Magnetizing Inductance	1.2 μH
Leakage Inductance	18 nH
Resonant Capacitance	3.3 μF
Core radius	0.9 mm
Winding width	1.65 mm
Converter dimensions	110 mm x 10 mm x 5.8 mm
Power Density	2200 W/in^3

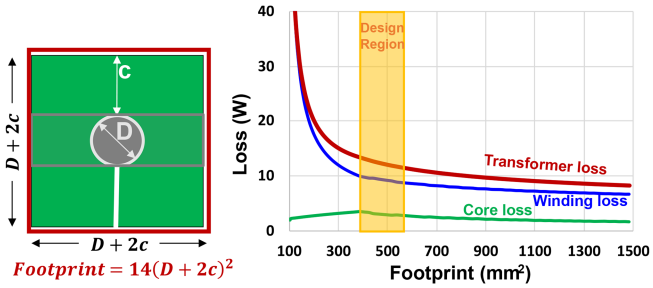


Fig. 17. Minimum half-load transformer loss vs footprint at 600 kHz.

where P_{core} is the core loss, k, α and β are the core loss coefficients for the core material, $B_{\max}$ is the peak magnetic flux density, f_s is the switching frequency, and V_{core} is the effective core volume. The core loss coefficients are curve-fitted based on the measurements in Fig. 16 and used to calculate the core loss.

C. Transformer Loss Optimization

The winding loss is calculated based on the Dowell's model [38] as

$$P_w = R_{DC}\left[Re(M_w) + \frac{m^2-1}{3}Re(D_w)\right] I_{RMS}^2 k_{fr}. \quad (2)$$

Here, R_{dc} is the winding dc resistance, M_w and D_w are the Dowell's ac coefficients, m is the magnitude of the relative MMF in the layer, I_{RMS} is the winding RMS current and k_{fr} is the fringing effects coefficient obtained from 2-D FEA simulations by dividing the simulated winding loss by the calculated winding loss using the Dowell's model. The winding dc resistance R_{dc} calculation for a circular PCB winding and the loss evaluation process was described in [25].

Based on the fact that GPUs operate at low-to-half load most of the time, the transformer's dimensions are optimized to maximize the half-load efficiency. The transformer loss, $P_{tr} = P_{core} + P_w$, is calculated for different core radii r and winding widths c, and the dimensions yielding minimum loss at each footprint are selected. This process is repeated at different footprints, resulting in the curves shown in Fig. 17. For such low-voltage, high-current transformers, the winding loss is higher than the core loss, and the transformer loss reduces with increasing footprint at the penalty of lower power density. The core radius is limited such that $B_{\max}$ does not exceed 300mT. This results in higher losses at low footprints since the resulting dimensions are not optimal. Therefore, the design region is selected based on the available footprint and around the "knee region" of the curve just beyond the nonoptimal points, where the tradeoff between efficiency and power density can be balanced.

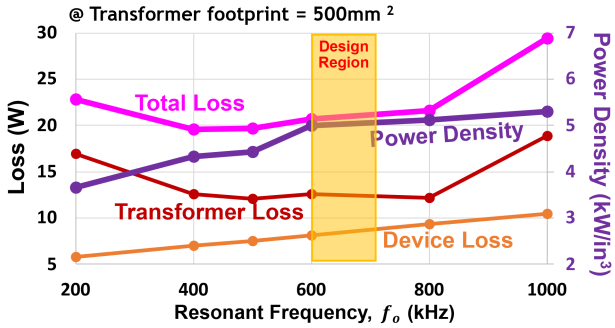
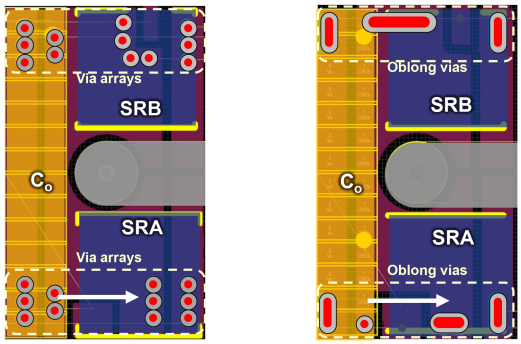


Fig. 18. Converter loss at half-load and power density vs f_o.

Because the LLC-DCX converter switches solely at the resonant frequency f_o, choosing the optimal f_o is essential for optimal performance. As the frequency increases, core loss in ferrite materials decreases for a given voltage within the recommended operating frequency range [29]. However, higher frequencies also increase RMS current and winding ac resistance, resulting in higher device and transformer winding losses [38]. Therefore, the converter losses (at half-load) and power densities are evaluated at different resonant frequencies as described in [39] and plotted in Fig. 18. As discussed, ML27D is used across the entire frequency range as it has the lowest P_v under high $B_{\max}$ conditions. Moreover, as the RMS currents increase with frequency, the device losses also increase. However, due to the nonlinearity of the increase in core loss (due to its exponential dependence), the design region around 600–700 kHz can be selected as a good tradeoff between efficiency and power density.

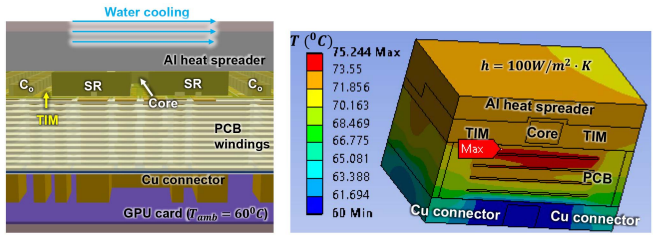
D. Further Design Optimizations

The vias linking the parallel secondary layers to the SRs handle significant ac currents, reaching approximately 60 A per transformer. These same vias must also conduct the filtered dc current from the filter capacitors to the load. Therefore, meticulous design is essential to minimize via losses. Furthermore, the large cutouts in the PCB layers, necessitated by these through-hole vias, cause current crowding in the intermediate primary layers, which leads to increased winding losses. To overcome these challenges, the arrays of multiple vias are substituted with large, custom-shaped, copper-filled pads, as illustrated in Fig. 19. This change increases the effective hole circumference by approximately 30% and the area by about 120%, thereby decreasing both the ac and dc resistance of the termination pathways. As a result, FEA simulations indicate a significant 25% reduction in total via losses, which is crucial for high output-current applications.



(a) (b)

Fig. 19. Improving via design by replacing (a) arrays of multiple vias to and (b) custom-shaped large pads.



(a) (b)

Fig. 20. Thermal management of transformer: (a) Front-view of one unit-cell with an Al heat spreader on the top and Cu connectors on the bottom and (b) FEA steady-state thermal simulations of half a unit-cell (1 transformer) with $h = 100$ W/m^{2}K and $T_{amb} = 60\,^{\circ}$C.

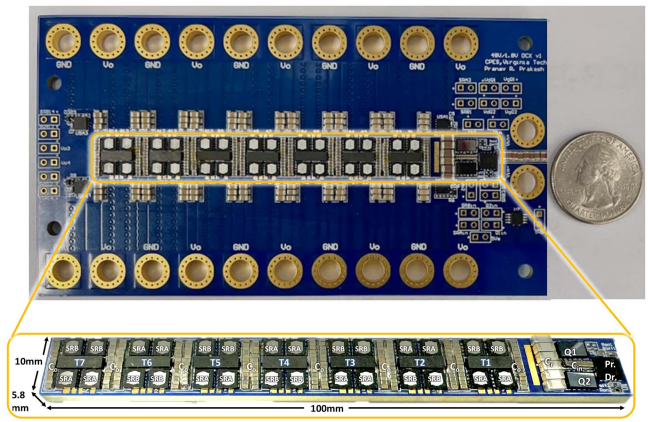


Fig. 21. Hardware prototype of the 840 W 48 V/1.8 V IBC module.

To minimize the output voltage ripple, the maximum possible output filter capacitance is desirable. However, it depends on the space available on the PCB winding. Based on the narrow winding width, 10 V 0402-sized ceramic capacitors are selected. However, the proposed unit-cell design allows many of them to be paralleled on both the top and bottom layers to increase the total capacitance (14 per layer per elemental transformer in this case).

A noteworthy phenomenon is that paralleling capacitors on the top and bottom layers, connected by the custom vias, introduces some parasitic inductance L_s between them, thereby causing an increase in the transformer's ac resistance around the resonant frequency f_s between C_o and L_s [40]. To mitigate this, the parasitic impedance of the termination vias is simulated using Ansys Q3D, and finally, 22 μF capacitors for the top layer and 10 μF for the bottom layer are selected to move f_s away from f_o.

E. Thermal Analysis With FEA Simulations

To minimize the thermal resistance from the heat sources with varying profiles on the top layer (SR devices, transformer core and output capacitors) to the heatsink on the top, a customized aluminum heat spreader is added between the components and liquid cooling plate to essentially replace the thermal interface from a TIM of $k = 17.8$ W/m $\cdot$ K to aluminum of $k = 240$ W/m $\cdot$ K [27]. In addition, since FR4 is not a good conductor of heat ($k = 0.2$ W/m $\cdot$ K), the heat from the lower PCB winding layers requires an alternate thermal path to the ambient environment. Therefore, large copper ($k = 400$ W/m $\cdot$ K) connectors for the ground output terminal are used, which double as a low thermal and electrical impedance path from the module bottom layer to the GPU card, as shown in Fig. 20(a).

It is impractical to set up an accurate thermal management system as in a real compute tray, where the GPUs are running next to the IBC power modules and sharing a common liquid cooling system. Therefore, steady-state thermal simulations were performed on ANSYS Workbench on one transformer (half a unit-cell) with the proposed thermal management and the results with a convection coefficient of $h = 100$ W/m^{2}K (nominal for liquid cooling) and ambient temperature of GPU card $T_{amb} = 60\,^{\circ}$C. is shown in Fig. 20(b). The maximum temperature increase from T_{amb} is 15.2 °C in the middle of the transformer, leaving enough thermal margin from the 35 °C rise limit for unknown factors such as thermal coupling between transformers. However, the results do indicate effective thermal dissipation from both sides of the converter module.

VI. Hardware Demonstration

The proposed 840 W 48 V/1.8 V IBC module is developed as shown in Fig. 21. It measures 100 mm $\times$ 10 mm $\times$ 5.8 mm and conforms to the required form factor to provide the high current to the VRs distributed equally along its length. Since the primary GaN devices require a short gate loop for optimal operation, the primary half-bridge gate driver LM5113-Q1 is placed on board. On the other hand, the drivers for the Si MOSFET SRs, FAN3122, are placed on the motherboard to minimize the module size. However, they can be moved to the power module to increase the length to 110 mm, resulting in a total converter power density of 2200 W/in^3.

Fig. 22 shows the operating waveforms at light-load and full-load, where it can be noted that ZVS for all switches is achieved throughout the entire load range. The $V_{tr,sec}$ waveform is the voltage across the two switching nodes SW1 and SW2 [refer Fig. 8(b)] which combines the V_{DS} of both SR devices. Furthermore, this voltage can be used to calculate and plot the

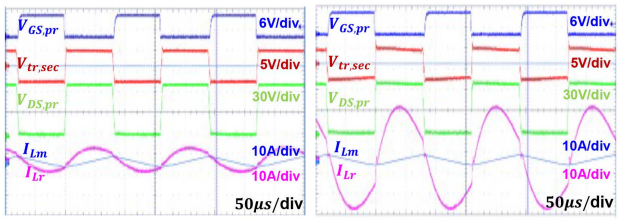


Fig. 22. Operational waveforms: (a) 20% load and (b) 100% load.

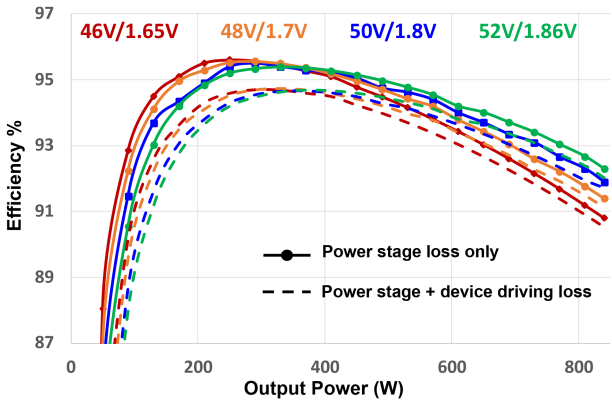


Fig. 23. Measured converter efficiencies at various operating conditions.

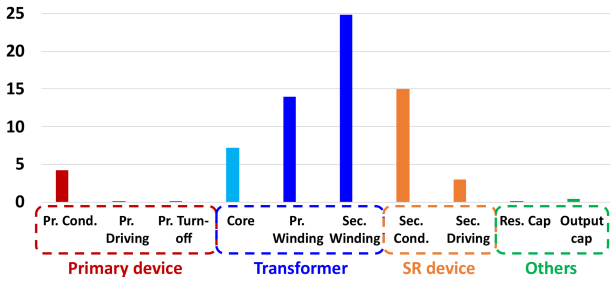


Fig. 24. Loss breakdown for 50–1.8 V at full-load.

magnetizing current i_{Lm} on the oscilloscope as

$$i_{Lm} = \int V_{tr,sec} \cdot n \cdot L_m/2. \tag{3}$$

Fig. 23 shows the measured efficiencies (excluding the device driving losses) in the V_{in} range of 46–52V, a peak efficiency of 95.6% at 46V and a full-load efficiency of 92.3% at 52V is achieved. The lower core loss at lower voltage results in higher peak efficiency, whereas the peak current at the same load is higher, reducing heavier load efficiencies. Fig. 24 shows the breakdown of the total converter loss at full load when operated at 50 V/1.8 V. Due to the low-voltage, high-current nature of the converter, the transformer winding loss and SR conduction losses are understandably dominant.

The thermal performance of the converter operating for 5 minutes at 80% load under 800 linear feel per minute (LFM) of air cooling from the right to left and no heat sinks is shown in Fig. 25. It may be noted that the hot spot of 102.3 °C is on the transformer core which is excited with a high $B_{\max}$ of 300 mT. The furthest

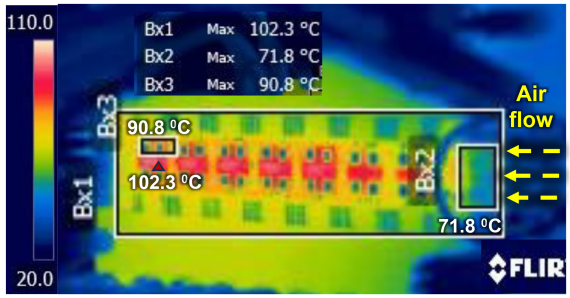


Fig. 25. Steady-state thermal performance at 80% load under 800 LFM air cooling and no heat sinks.

TABLE VI
PERFORMANCE COMPARISON OF BUS CONVERTERS WITH LOW IBV

Parameters	[32]	[41]	[42]	[43]	This work
V_{in}(V)/V_o(V)	36/0.75	48/1.0	40/1.0	48/1.0	48/1.8
P_o(W)	112	300	200	350	840
I_o(A)	150	300	200	350	460
f_s(MHz)	1.0	1.0	1.0	1.0	0.67
Peak eff.	94.1%	93.8%	97.0%	94.0%	95.5%
Full-load eff.	86.0%	89.0%	91.8%	85.0%	92.2%
PD (W/in^3)	1050* †	1010	1394	412†	2200

*Estimated for entire power stage
†Power density excludes driving circuitry.

SR devices have a temperature of 90.8 °C and primary devices have a temperature of 71.8 °C, thanks to the multiple thermal vias underneath them. Although heatsinks would be required to run the converter at full power under air cooling, the proposed thermal management described in Section V should help achieve the temperature rise requirements in the actual implementation.

The performance of the proposed converter is compared with recent research on unregulated low voltage, high-current dc–dc converters for this application, as highlighted in Table VI. All efficiencies listed exclude the device driving losses. The optimized frequency of operation of the proposed converter can enable higher currents to be pushed without incurring significant ac losses, thereby increasing the power density significantly while still maintaining high efficiencies across the load range.

VII. CONCLUSION

The power delivery must keep up with the rapidly increasing size and power consumption of SOCs. However, the existing two-stage lateral power delivery architecture with a 12 V IBV results in the second-stage VRs taking up a significant chunk of the available real estate on the GPU card, thereby limiting the size and power increase of the GPUs. Moreover, the VRs operate at frequencies around 1 Mhz, which limits the bandwidth to hundreds of kHz and requires large capacitance to meet the steady-state and transient requirements. Therefore, vertical power delivery architecture is attracting significant interest to address this challenge. By reducing the IBV, the size of the second-stage VR can be reduced significantly and placed directly underneath the GPU, thereby delivering the high current vertically with very low PDN loss. Furthermore, the operating frequencies of these low-IBV VRs can be increased 50–100

times, thereby increasing the bandwidth and reducing the required capacitance significantly. However, the first stage IBC must be designed with a long and thin top-cooled form factor and placed very close to the GPU to reduce the high PDN loss due to the low IBV.

To achieve this, the article proposed a highly scalable transformer unit-cell structure consisting of two transformers of 1:1 or 2:1 turns each, coupled with a simple UU/UI core—furthermore, the high space utilization of the footprint results in high power densities.

To implement the design, an 840 W, 48 V/1.8 V HB LLC-DCX converter is designed with seven unit-cells connected in series to complete the required 14:1 turns ratio. For GPUs requiring even higher power, such modules may be connected in parallel from the input and placed on either side of the GPU to provide the high current symmetrically to the VRs. The design of the LLC-DCX is described in detail, which can achieve a power density of 2200 W/in3 and a peak efficiency of 95.5%. Furthermore, the scalability of the proposed transformer unit-cell is highlighted, where high-density IBCs with different IBVs can efficiently be designed. This work validates the feasibility of the vertical power delivery solution to GPUs, thereby paving the way for future innovations in optimizing the IBV and further improving the performance.

ACKNOWLEDGMENT

The authors would like to thank the CPES Power Management Consortium (PMC) and High Density Integration (HDI) mini-consortia for sponsoring this research.

REFERENCES

[1] A. S. G. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[2] S. Oliver, "From 48V direct to intel VR12.0: Saving 'Big Data' 500,000 per data center, per year," Jul. 2012. [Online]. Available: http://www.vicorpower.com/documents/whitepapers/wp_VR12.pdf

[3] S. Taranovich, "Data center next generation power supply solutions for improved efficiency," Apr. 2016. [Online]. Available: https://www.edn.com/data-center-next-generation-power-supply-solutions-for-improved-efficiency/

[4] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "Nvidia A100 tensor core GPU: Performance and innovation," *IEEE Micro*, vol. 41, no. 2, pp. 29–35, Mar./Apr. 2021.

[5] A. Tirumala and R. Wong, "NVIDIA blackwell platform: Advancing generative AI and accelerated computing," in *Proc. IEEE Hot Chips 36 Symp.*, 2024, pp. 1–33.

[6] J. Choquette, "NVIDIA hopper H100 GPU: Scaling performance," *IEEE Micro*, vol. 43, no. 3, pp. 9–17, May/Jun. 2023.

[7] R. K., "Graphics processing unit GPU market report 2024 (Global edition)," Cognitive Market Research, 2023.

[8] S. Kudva, "Challenges to enabling vertical power delivery in high-power GPU applications," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, 2024.

[9] P. R. Prakash et al., "A 2400 W/in3 1.8 V bus converter enabling vertical power delivery for next-generation processors," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, 2024, pp. 910–917.

[10] Y. Elasser et al., "Mini-LEGO CPU voltage regulator," *IEEE Trans. Power Electron.*, vol. 39, no. 3, pp. 3391–3410, Mar. 2024.

[11] Y. Zhu et al., "A compact 48-V-to-sub-1-V switching bus converter with 4.7-mm height for processor vertical power delivery," in *Proc. IEEE Energy Convers. Congr. Expo.*, 2024, pp. 2596–2603.

[12] F. Zhu and Q. Li, "A novel PCB-Embedded coupled inductor structure for a 20-MHz integrated voltage regulator," *IEEE Trans. Emerg. Sel. Topics Power Electron.*, vol. 10, no. 6, pp. 7452–7463, Dec. 2022.

[13] A. M. Naradhipa, F. Zhu, and Q. Li, "Ultra-low-Profile twisted core inductor for vertical power delivery voltage regulator," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, 2024, pp. 918–924.

[14] S. Jiang and Z. Ye, "Next TLVR innovations: Topologies, magnetics and control," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, 2024.

[15] K. Radhakrishnan, M. Swaminathan, and B. K. Bhattacharyya, "Power delivery for high-performance microprocessors—Challenges, solutions, and future trends," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 11, no. 4, pp. 655–671, Apr. 2021.

[16] Empower, "3.3 V and 1.8 V IVRs," [Online]. Available: https://www.empowersemi.com/integrated-voltage-regulators-ivr/

[17] Ferric, "Fe1736," [Online]. Available: https://www.ferric.com/products

[18] M. Chen, S. Jiang, J. A. Cobos, and B. Lehman, "Design considerations for 48-V VRM: Architecture, magnetics, and performance tradeoffs," in *Proc. 4th Int. Symp. 3D Power Electron. Integration Manuf.*, 2023, pp. 1–9.

[19] Y. Li, X. Lyu, D. Cao, S. Jiang, and C. Nan, "A 98.55% efficiency switched-tank converter for data center application," *IEEE Trans. Ind. Appl.*, vol. 54, no. 6, pp. 6205–6222, Nov./Dec. 2018.

[20] S. Jiang, S. Saggini, C. Nan, X. Li, C. Chung, and M. Yazdani, "Switched tank converters," *IEEE Trans. Power Electron.*, vol. 34, no. 6, pp. 5048–5062, Jun. 2019.

[21] Y. Li, X. Lyu, Z. Ni, D. Cao, C. Nan, and S. Jiang, "Adaptive on-time control for high efficiency switched-tank converter," in *Proc. 1st Workshop Wide Bandgap Power Devices Appl. Asia*, 2018, pp. 169–175.

[22] A. Dago, M. Balutto, S. Saggini, M. Leoncini, S. Levantino, and M. Ghioni, "Hybrid resonant switched tank converters for high step-down voltage conversion," *IEEE Trans. Power Electron.*, vol. 39, no. 11, pp. 14838–14851, Nov. 2024.

[23] M. H. Ahmed, F. C. Lee, and Q. Li, "Two-stage 48-V VRM with intermediate bus voltage optimization for data centers," *IEEE Trans. Emerg. Sel. Topics Power Electron.*, vol. 9, no. 1, pp. 702–715, Feb. 2021.

[24] M. de Rooij and A. Negahdari, "Beyond 4 kW/in3 power-density for 48 V to 12 V conversion using eGaN FETs in an LLC DC-DC bus converter," in *Proc. PCIM Europe; Int. Exhib. Conf. Power Electron., Intell. Motion, Renewable Energy, Energy Manage.*, 2022, pp. 1–9.

[25] C. Fei, F. C. Lee, and Q. Li, "High-efficiency high-power-density LLC converter with an integrated planar matrix transformer for high-output current applications," *IEEE Trans. Ind. Electron.*, vol. 64, no. 11, pp. 9072–9082, Nov. 2017.

[26] *EPC Thermal Model. Calculator Quick Start Guide*, 3rd Ed., El Segundo, CA, USA: EPC Corporation, 2022. [Online]. Available: https://epc-co.com/epc/Portals/0/epc/documents/guides/GaN-FET-Thermal-Calculator-qsg.pdf

[27] P. R. Prakash, A. Nabih, S. Wang, P. P. Hieu, Y. Ruan, and Q. Li, "GaN-Based 400 V/48 V DC-DC converter with 97% efficiency and PCB magnetics for automotive applications," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, 2023, pp. 3201–3208.

[28] A. Nabih and Q. Li, "Design of 98.8% efficient 400-to-48-V *LLC* converter with optimized matrix transformer and matrix inductor," *IEEE Trans. Power Electron.*, vol. 38, no. 6, pp. 7207–7225, Jun. 2023.

[29] D. Reusch and F. C. Lee, "High frequency bus converter with low loss integrated matrix transformer," in *Proc. 27th Annu. IEEE Appl. Power Electron. Conf. Expo.*, 2012, pp. 1392–1397.

[30] Z. Ouyang and M. A. E. Andersen, "Overview of planar magnetic technology—Fundamental properties," *IEEE Trans. Power Electron.*, vol. 29, no. 9, pp. 4888–4900, Sep. 2014.

[31] P. Vinciarelli, "Delivering power to semiconductor loads," US Patent US11398770B1, Jul. 2022.

[32] H. Wu, Y. Song, X. Zhang, and Y. Zhang, "Current-cancellation-based heterogeneous integration of LLC-DCX with ultralow-voltage high-current output for data centers," *IEEE Trans. Power Electron.*, vol. 39, no. 8, pp. 9144–9149, Aug. 2024.

[33] K. Ngo, E. Alpizar, and J. Watson, "Modeling of losses in a sandwiched-winding matrix transformer," *IEEE Trans. Power Electron.*, vol. 10, no. 4, pp. 427–434, Jul. 1995.

[34] Y. Cai, M. H. Ahmed, Q. Li, and F. C. Lee, "Optimal design of megahertz *LLC* converter for 48-V bus converter application," *IEEE Trans. Emerg. Sel. Topics Power Electron.*, vol. 8, no. 1, pp. 495–505, Mar. 2020.

[35] *Mn-Zn Soft Ferrite Cores for High Frequency Power Supplies MaDC-F, Proterial Ltd*. Tokyo, Japan: Hitachi Metals, 2019.

[36] D. Hou, M. Mu, F. C. Lee, and Q. Li, "New high-frequency core loss measurement method with partial cancellation concept," *IEEE Trans. Power Electron.*, vol. 32, no. 4, pp. 2987–2994, Apr. 2017.

[37] M. Mu and F. C. Lee, "A new core loss model for rectangular AC voltages," in *Proc. IEEE Energy Convers. Congr. Expo.*, 2014, pp. 5214–5220.
[38] P. L. Dowell, "Effects of eddy currents in transformer windings," in *Proc. Inst. Elect. Engineers*, 1966, vol. 113, no. 8, pp. 1387–1394.
[39] P. R. Prakash, A. Nabih, and Q. Li, "Design optimization of PCB-Winding matrix transformer for 400 V/12 V unregulated LLC converter," in *Proc. 2021 IEEE Energy Convers. Congr. Expo.*, 2021, pp. 1777–1784.
[40] P. Prakash, A. Nabih, and Q. Li, "Termination design optimization of high-current PCB-Winding matrix transformers," *IEEE Trans. Power Electron.*, vol. 38, no. 4, pp. 4957–4971, Apr. 2023.
[41] X. Ren, J. Zhang, Y. Jiang, X. Li, and T. Long, "A 48-to-1 V LLC DC transformer," in *Proc. IEEE 24th Workshop Control Model. Power Electron.*, 2023, pp. 1–5.
[42] K. Wang, Y. Ning, H. Li, and X. Yang, "Integrated planar transformer for high-voltage-ratio LLC DCX with high current and ultra-low voltage output," *IEEE Open J. Power Electron.*, vol. 5, pp. 1782–1791, 2024.
[43] A. Figueroa, P. Mazariegos, J. Goicoechea, A. Castro, and J. A. Cobos, "Low-profile direct power converter: 350 A/48V-1 V with planar matrix transformer using standard PCB and commercial cores," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, 2024, pp. 2172–2177.

Pranav Raj Prakash (Graduate Student Member, IEEE) received the B.Tech. degree in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, India, in 2018, and the M.S. degree in electrical engineering, in 2021, from the Center for Power Electronics Systems (CPES), Virginia Tech, Blacksburg, VA, USA, where he is currently working toward the Ph.D. degree in electrical engineering.

His research interests include the design and control of high-frequency, high-density dc–dc converters employing wide band-gap devices, and planar magnetics.

Ahmed Nabih (Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Cairo University, Giza, Egypt, in 2014 and 2017, respectively, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2023.

He is currently a Research Scientist with Nvidia Corporation, Durham, NC, USA. His research interests include wide bandgap technology, high-power drivers, digital control, high-frequency resonant converter design, and planar magnetics.

Dr. Nabih was a recipient of the Best Paper Award of the 2020 OCP Future Technology Symposium, the 2023 IEEE TPEL prize paper award, and multiple APEC outstanding presentation awards.

Yan Liang (Student Member, IEEE) received the B.S. degree in electrical engineering from Beijing Jiaotong University, Beijing, China, in 2021. She is currently working toward the Ph.D. degree in electrical engineering with the Center for Power Electronics Systems (CPES), Virginia Tech, Blacksburg, VA, USA.

Her research interests include high-frequency, high-density dc–dc converters, and planar magnetics.

Sudhir Kudva (Member, IEEE) received the B.E. degree in electronics and communcation engineering from the National Institute of Technology, Suratkal, India, in 2004, the M.E. degree in microelectronics from the Indian Institute of Science, Bangalore, India, in 2006, and the Ph.D. degree in electrical engineering from the University of Minnesota, MN, USA, in 2013.

He is a Senior Research Scientist with the Circuit Research Lab of Nvidia Research, which he joined in 2013. From 2006 to 2008, he also worked as a Design Engineer with AMD, India.

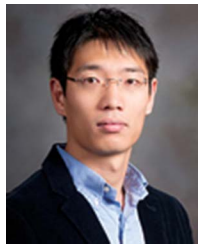
His research interests include power delivery network design, regulator module implementation, fully-integrated voltage regulator, and regulator design for security-sensitive circuits.

Mostafa Mosa received the B.S. degree in electrical engineering from South Valley University, Egypt, in 2008, the M.S. degree in electrical engineering from Aswan University, Egypt, in 2012, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2018.

He has authored or coauthored more than 22 journal and conference papers, as well as one book and one book chapter in his area of expertise. He is also the Holder of two U.S. patents in the field of power electronics and energy systems. He is currently working as a Power System Design Engineer with NVIDIA Corporation, Durham, NC, USA, where he focuses on developing advanced power solutions for data center applications. His research interests include power electronics for data center and renewable energy.

C. Thomas Gray (Senior Member, IEEE) received the B.S. degree in computer science and mathematics from Mississippi College, Clinton, MS, USA, and the M.S. and Ph.D. degrees in computer engineering from North Carolina State University, Raleigh, NC, USA.

From 1993 to 1998, he was an Advisory Engineer with IBM, Research Triangle Park, NC, USA, working on transceiver design for communication systems. From 1998 to 2004, he was a Senior Staff Design Engineer with the Analog/Mixed Signal Design Group, Cadence Design Systems, working on SerDes system architecture. From 2004 to 2010, he was a Consultant Design Engineer with Artisan/ARM and Technical Lead of SerDes architecture and design. In 2010, he joined Nethra Imaging as a System Architect. His work experience includes digital signal processing design and CMOS implementation of DSP blocks as well as high-speed serial link communication systems, architectures, and implementation. In 2011, he joined NVIDIA, Inc., Durham, NC, USA, where he is currently Senior Director of Circuit Research, leading activities related to high-speed signaling, photonics, security circuits, low-energy and resilient memories, circuits for machine learning, and variation-tolerant clocking and power delivery.



Qiang Li (Senior Member, IEEE) received the B.S. and M.S. degrees in power electronics from Zhejiang University, Hangzhou, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2011.

He is currently an Associate Professor with the Center for Power Electronics Systems, Virginia Tech. His current research interests include power management for distributed power systems, applications of wide bandgap power devices, high-frequency power conversion and controls, magnetics and electromagnetic interference, high-density electronics packaging and integration, and renewable energy.

Dr. Li was the recipient of the First Place Prize Article Award for IEEE TRANSACTIONS ON POWER ELECTRONICS in 2016, and the 2017 U.S. National Science Foundation (NSF) Career Award.