How to Increase Tokens per Watt by 30 percent

Taner Dosluoglu
weeteq Ltd
Glasgow, UK
https://orcid.org/0000-0003-3738-0476

Steve Conner
weeteq Ltd
Glasgow, UK
steve@weeteq.com

Martin MacDonald
weeteq Ltd
Glasgow, UK
martin@weeteq.com

***Abstract*—Power delivery solutions within AI and high performance compute (HPC) data centers have received significant recent attention, with a higher voltage input bus being planned for server trays. In parallel with this effort, high performance processor manufacturers integrate on-chip solutions to address the power delivery network challenges of fast load transients within power domains. Higher power requirements have necessitated higher voltage components to be placed closer to the package, while the efficiency requirements have limited the available solutions with high bandwidth to provide active droop compensation. This paper presents a dual-path power delivery solution, optimizing both efficiency and fast load transient response orchestrated by a central controller, reducing total power consumption.**

Keywords—power delivery network, droop compensation, vertical-parallel delivery, high voltage ratio, rack power solutions

I. Introduction

The increasing energy consumption of data centers is a global concern requiring a major collaborative effort. Significant effort has already been made towards efficiency improvements of power converters, the remaining few percent is unlikely to improve due to physics constraints, hence any suggestion of 30% energy improvement is understandably challenged by an industry which recognises the daunting task of improving the efficiency of high voltage converters. At the same time, such converters typically provide a highly efficient solution down to only 12V, leaving the rest to be optimized by board level voltage regulators and on-chip power management functions. There remains room for improvement, since any on-chip integrated power solution that does not address the voltage droop by less than 25% is considered ineffective. The fundamental issue for on-chip power is potential timing failures due to large voltage droops in the power delivery network (PDN) due to aggressive load changes. These dynamic load changes are required in order to maximize the performance of power domains within the thermal and energy density constraints. The PDN constraints are most apparent in application processors used in mobile phones where the maximum clock frequency has stayed unchanged around 1.5GHz for more than a decade, despite the multiple process nodes introduced, each improving power, performance, area (PPA) significantly. The high performance compute processors have pushed this issue to its logical conclusion from chip level thermal constraints to a worldwide total energy constraint, without solving the fundamental problem. The optimization challenge is pushed down to integrated power management solution providers where the boundary between the high voltage and chip level power domain is more likely to be a large bank of capacitors in real world examples.

The proposed solutions for addressing voltage droop is either limited to performance constraints that mitigate the issue, or provide a relative improvement addressing a very specific problem. Even when the solution is pushed to the ultimate limit of instant energy transfer from higher voltages, input supply and PDN constraints limit the performance.

II. Requriements and System Level Constraints

A. Power Delivery Network

A system level architectural optimization is required that removes the artificial boundary between the higher voltages (>12V) and the power domains of processor cores on-chip. This requires consideration of the impedance profile of the entire power delivery network in the frequency domain and the corresponding time domain signature that defines the first droop that has to be addressed within 10 nanoseconds, the second droop within 100 nanoseconds.

B. Power Domains - Fast Load Transient

On-die power delivery solutions need to address both planned and unplanned load transients. Unplanned load transients are an inherent part of the processor operations due to the complex dependency of the processor workloads to the memory cache operations and most likely the underlying cause of the first droop encountered on-chip. An example of planned load transient is dynamic voltage frequency scaling (DVFS) functions, setting target voltage in order to ensure the processor does not operate at voltages higher than what is minimally required for the specified clock frequencies. Use of hierarchical voltage regulator architectures for efficient, fine-grained DVFS implementation in multi-domain SoCs have been reported to save 25% to 45% in peak power.

In this paper, we introduce a Fast Load Transient path that can handle the input requirements of on-chip power solutions using hybrid topology of switched capacitor converters and inductor based voltage regulators.

C. High Current – High Efficiency Path

The use of transformer-based converters facilitates high efficiency conversion at high input-to-output voltage ratios. Using a two stage approach where the first stage is a fully regulated buck and the second stage is an unregulated fixed ratio converter, allows higher voltages to be brought closer to the package and reduces the high current distribution losses. Efficient implementation of such an architecture requires switching frequencies and components that would not support the fast load transient requirements at the input of the on-chip power solutions. An ideal solution would be able to combine the high efficiency path with the fast load transient path without sacrificing the power density that is required for package integration.

III. Proposed Architecture

The proposed architecture anticipates future data centers with a high voltage DC power distribution and delivery to the rack. This will initially have an active front end power shelf converting 3-phase 400V AC to the 400V or 800V DC without isolation. It replaces the relatively bulky multi-stage converter that performs power factor correction of the AC input, voltage step down to 48V. In future data centers with 800V DC; this power shelf and its cooling needs will disappear from the rack.

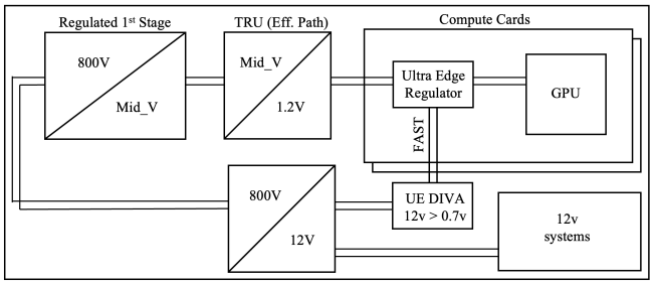

Fig. 1. Compute Tray Power Distribution Architecture.

Within the compute tray the power distribution board has 800V DC input and provides a regulated, non-isolated Mid-level high voltage to a transformer-rectifier unit (TRU) and an isolated 12V auxiliary output for a parallel path. The high level architecture is provided in Figure 1 showing the high efficiency path via TRU and the parallel path that supports the FAST load transient connected to the GPU, and auxiliary 12V providing power to all the other 12V components on the tray.

The transformer-rectifier unit (TRU) further converts the Mid-level (150V-250V) voltage down to the processor chip IO voltage levels that are in the range of 1.2V to 1.8V. For this purpose, we will be using an isolated forward converter with a matrix transformer, consisting of multiple small transformers with primaries in series and secondaries in parallel. This is the main high efficiency path for the current delivery to the processor units (GPU/CPU).

In parallel with TRU which provides bulk of the power needed by the processor directly from the high voltage bus, an auxiliary 12V VRM provides power to the processor. The reduction in power throughput provides the needed relaxation of constraints for the VRM for improving the fast load transient response with smaller components and faster switching frequencies at the expense of efficiency. This is similar to the design constraints of on-chip and in-package integrated voltage regulators which can initially be used together with the proposed parallel path delivery system.

IV. Discussions on Implementation

Special consideration has to be given to the physical placement of the output stages of the TRU which will have a very high current and need to be placed on the opposite side of the processor package for efficient vertical power delivery. It should be noted that the proposed architecture is a natural match for immersion liquid cooling. The high voltage circuity can be made more compact because the cooling fluid has better electrical properties than air. The fast load transient response connections ('FAST' in Fig. 1) of VRM on the other hand, can be placed on the same side of the board for lateral power delivery. We have taken this into consideration and use the inductance of the package and PCB trace within the design parameters for up to 3nH inductance and 66% component variation (+/- 2nH).

Typical voltage and current waveforms are illustrated in Fig. 2 where the efficiency path response (TRU Current) and FAST path response requirements are captured. The TRU response is designed to accommodate low frequency (below 1MHz) resonant peak of power distribution network while the mid and high frequency resonant peaks are addressed by the FAST VRM (up to 100MHz) and on-chip power solutions. The most critical contribution of the fast parallel path is to eliminate the second droop which is traditionally managed by

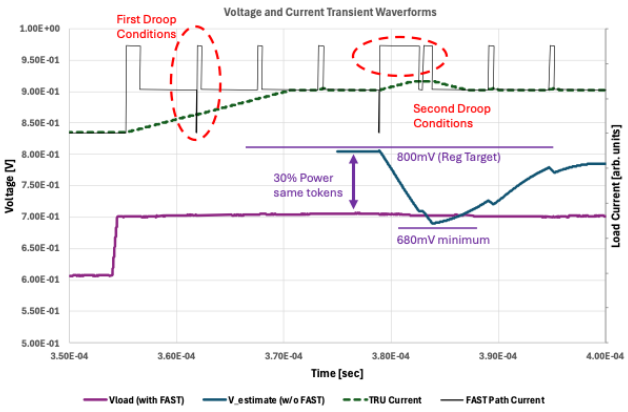

Fig. 2. Power Domain voltage and current waveforms.

adding more capacitors in the package and on the PCB. This will no longer be an option for the future high performance compute trays. The voltage droop with and without the parallel 'FAST' path is shown in Fig. 2 where the target reference voltage needs to be 100mV above the proposed solution to accommodate minimum guaranteed voltage with more than two hundred 100uF capacitors placed next to the package to accommodate the same load transient as the proposed dual path architecture. The 100mV regulation margin directly translates into 30% more power for the same amount of compute performance.

The integrated voltage regulators and in-package VRM solutions aim to address this issue but they suffer from efficiency due to compromise between the efficiency and system response bandwidth which is fundamental regardless of the topology. Comparison of total power consumption and voltage droop compensation between different architectures provided in fig. 3. We compare the total power consumption of an efficient, lower bandwidth two stage solution with 98% and 95% efficiencies respectively (93% end-to-end efficiency) with the multi-stage solution having a high bandwidth final stage with 85% efficiency (75% end-to-end) and additional 100mV droop compensation. Even with the lower end-to-end efficiency the total power consumption is improved by 5.5%.

	Best Efficiency	Best IVR/400VAC	Dual Path
End-to-End Efficiency (%)	93%	75%	91% / 81%
Vtarget / Vdroop	0.8V / 130mV	0.7V / 30mV	0.7V / 30mV
Power at Load [normalized]	1000	766	766
Power saved on-chip	-	30%	30%
Power at Source [normalized]	1074	1018	862
Power saved at HV supply	-	5.5%	25%

Fig. 3. Comparison of power and voltage droop compensation.

The proposed dual path solution is included in Fig. 3 with the efficiency path providing 80% of the current at 91% efficiency and the higher bandwidth path with the remaining current at 81% efficiency. The overall power saving is 25% with additional savings in cooling requirements.

V. Conclusions

A new dual-path architecture is proposed as a data center tray power delivery solution that provides very efficient high voltage to low voltage power conversion and active voltage droop compensation at high currents. A use case is provided to illustrate a reduction of the processor chip power by 30% and reduction of the overall power by 25%, compared to existing solutions. Design progress of prototype will be presented.