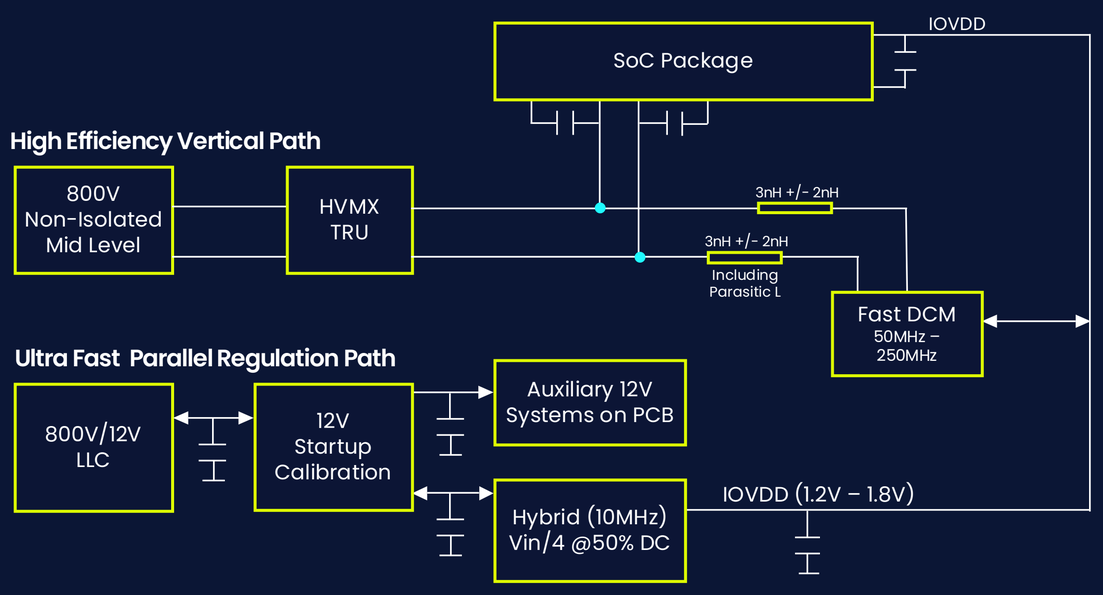
30% More Tokens Per Watt

The Dual-Path Power Revolution

Dr Taner Dosluoglu
taner@weeteq.com

As HPC data centers face unprecedented energy demands, power delivery networks are bottlenecked by fast load transients causing on-chip voltage droops. Systems overcompensate with large capacitor banks and a 100mV regulation margin, wasting significant power. weeteq's dual-path architecture eliminates the artificial boundary between higher voltages and on-chip power domains, pairing a high-efficiency bulk path with a high-bandwidth fast transient path to actively compensate for voltage droops.
The result: 30% less processor chip power, 25% less overall consumption, and 30% more Tokens per Watt.

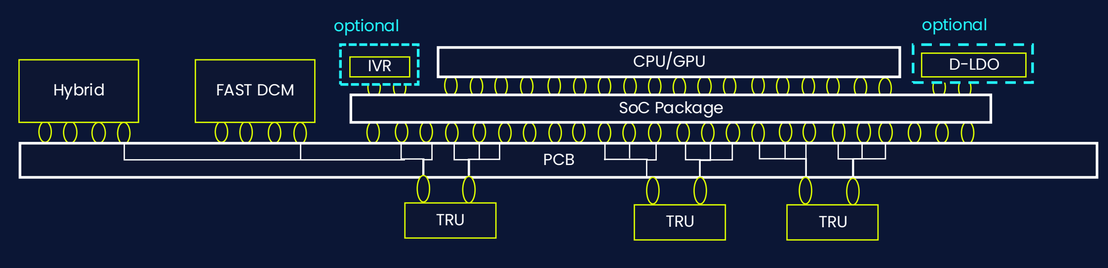
D/VA Parallel Architecture



The High Efficiency Vertical Path provides direct 800V DC distribution, single-step high-ratio conversion to processor domain, optimised for bulk current delivery with reduced bus current and copper loss. The TRU delivers estimated average current at >90% efficiency.

The Ultra Fast Parallel Regulation Path provides localised transient absorption, >10 A/ns dynamic current response, >1 mV/ns voltage scaling, and sub-10ns droop correction. The High Voltage Matrix Transformer (HVMX) transient response is controlled by the Ultra Edge Controller which provides optimum average current and monitors real-time deviation from ideal transient response (100ns). The deviation from ideal transient response of HVMX is compensated by FAST DCM path providing faster response at point of load (10ns). FAST DCM can also provide required transient response to improve voltage droop due to package power delivery network frequency response (50MHz - 250MHz).

Physical Layout



TRU output stages carry very high current and are placed on the opposite side of the processor package for efficient vertical power delivery. The FAST load transient response connections can be placed on the same side of the board for lateral power delivery. The bi-directional Hybrid (LC) VRM recycles excess power and compensates HVMX, while Fast DCM provides up to 100MHz droop compensation. On-chip and in-package integrated voltage regulators can be combined with the proposed parallel path delivery system as DIVA can work with existing IVR & digital LDO solutions.

Comparison of Architectures

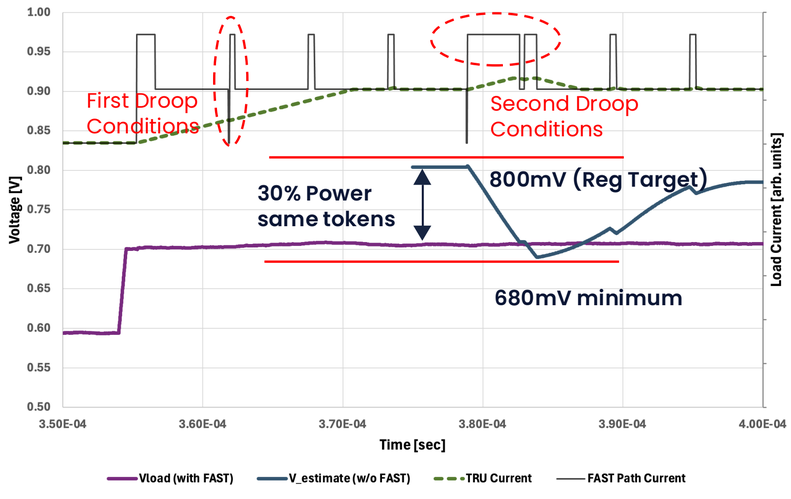
	Best Efficiency	Best IVR/400VAC	D/VA Dual Path
End-to-End Efficiency (%)	93%	75%	91% / 81%
Vtarget / Vdroop	0.8V / 130mV	0.7V / 30mV	0.7V / 30mV
Power at Load [normalized]	1000	766	766
Power saved on-chip	-	30%	30%
Power at Source [normalized]	1074	1018	862
Power saved at HV supply	-	5.5%	25%

The integrated voltage regulators and in-package VRM solutions aim to address voltage droop but they suffer from lower efficiency due to compromise between efficiency and system response bandwidth, which is fundamental regardless of topology. Different approaches are compared in the table above. An efficient lower bandwidth two stage solution can achieve 93% end-to-end efficiency, requiring a 0.8V target with 130mV droop is provided as a reference. A multi-stage solution with a high bandwidth final stage can provide 100mV droop compensation but with lower efficiency (typically 75% end-to-end efficiency). Even with this low end-to-end efficiency, the total power consumption is improved by 5.5% compared to reference solution without active droop compensation. The DIVA dual path solution provides 80% of the current via the efficiency path at 91% efficiency and the remaining current is provided by the higher bandwidth path; meeting 680mV minimum guaranteed voltage requirement with 0.7V target voltage and only 30mV droop; reducing power at source for the same performance without any change to the workload or the DVFS functionality. Power saved on-chip is 30%. The overall power saving is 25% with additional savings in cooling requirements.

At rack level, the existing power architecture requires 120kW per rack at 400V 3-phase AC with 13kW per compute node. The Ultra Edge 800V DIVA Power Architecture requires 90kW per rack at 800V DC with 10kW per compute node. Same TPS achieved for 0.7V target power domain voltage instead of 0.8V. A reduction of the processor chip power by 30%, reduction of the overall power by 25%, and 30% more Tokens per Megawatt.

weeteq whitepaper

Voltage and Current Transient Waveforms



The voltage droop with and without the parallel FAST path is shown. Without the FAST path, the target reference voltage needs to be 100mV above the proposed solution to accommodate minimum guaranteed voltage. This would require more than two hundred “ideal” 100uF capacitors placed next to the package to accommodate the same load transient. The dual path architecture eliminates this requirement.
The efficiency path response (TRU Current) provides bulk current delivery while the FAST path responds to the transient event, keeping the load voltage above the 680mV minimum requirement. The 100mV regulation margin reduction directly translates into 30% less power for the same amount of compute performance.