Artificial intelligence: powering the deep-learning machines of tomorrow

Deep learning neural networks demand sophisticated power solutions

Abstract

Once very much a science fiction dream, artificial intelligence (AI) is rapidly becoming part of our daily lives. While AI takes many forms, the systems that mimic the human brain's learning and problem solving capability comprise increasingly capable computer based “neural” networks consisting of many paralleled processors that run complex learning and execution software algorithms.

While the algorithms are key to this technology, the computer power system demands are stretching the boundaries of existing power delivery technology. In this white paper, Infineon will look at the power demands for AI systems, as well as presenting some of the latest, state-of-the-art, power technologies that are enabling the advances in this exciting arena.

By
Danny Clavette, Director, Systems Applications
Infineon Technologies AG

Table of contents

1 Overview of artificial intelligence

Humans are smart, achieving intelligence through years of learning and data accumulation as well as arguably getting “wiser” with age. Computers could be considered “smart” due to data retention capabilities but until recently lacked the capability to autonomously learn from these large databases in order to execute tasks or make decisions. While a human brain consumes 20-30 W of power, the latest learning systems are consuming power at levels that would support a small town as they learn to become 'artificially intelligent'. While we can debate whether computing is getting 'smarter' than humans, it is impossible to debate that the requirements for powering this new generation of supercomputer have changed dramatically.

In some ways, the approach taken to AI deep learning is quite similar to human development where computers continue to learn through exposure. In the example below, a neural network is fed with thousands of training images that are processed via multiple layers in order to build experience and knowledge.

As a result of this computer intensive and power hungry learning process, the network is eventually able to distinguish a squirrel from a chipmunk or a fox. The goal is to achieve AI learning in the shortest amount of time, thus parallel computing power is maximized to linearly improve computation times.

The high power consumption of today's AI is driving changes in the computing architecture to replicate neural networks that mimic the human brain in an effort to reduce power needs. Traditional Central Processing Units (CPUs) are architected to be very flexible to support a wide variety of general-purpose programs and are not optimized for very specific and repetitive tasks such as AI learning.

Many of the necessary functions for AI can be performed by Graphics Processing Units (GPUs). These GPUs are designed to repeatedly perform complex mathematical functions more efficiently, can be conveniently connected in parallel to further increase computing power and be opportunistically applied to learning applications. With slight modifications, these latest GPU devices process 3x to 10x faster while consuming the same power as a CPU. The early AI market has been dominated by NVIDIA; their DX1 GPU super computer contains eight Tesla P100 GPUs, each capable of 21.2 TeraFLOPs, and requires 3200 W of total system power. Multiple DX1s connected in parallel are required to form an effective neural network.

Honing the technology even further, Tensor Processing Units (TPUs) are ASICs that have been developed specifically for machine learning. Based on GPU platforms, reduced floating-point accuracies allow more compute capability per clock cycle. Rasterization and texture mapping features are also removed to further improve computation efficiency. Google launched the first TPU in 2015 and Intel is expected to launch LakeCrest this year, targeting Deep Neural Network (DNN).

To learn, networks need to be able to sense. Local 'edge devices' include sensors, cameras, data collectors and local actuators. Connected to the central AI servers via high-speed wireless connections, these low power devices are the eyes, ears and hands of the neural network. Estimates predict that there will be over 50 billion edge devices connected to the network by 2020.

It should come as no surprise that, despite the power challenges, the market for AI is growing rapidly as demonstrated by the (approximately) 40-fold growth at Google in the past two years.

2 Challenges for power designers

The power levels required for this new technology are simply staggering. In order to match the processing power of a human brain, a system would need to perform more than 38 thousand trillion operations per second (or 38 PetaFLOPS according to Dharmendra Modha, IBM Fellow and Chief Scientist at the Almaden Research Center). For an interesting comparison, a server farm using NVIDIA's DX1's 21.2 TeraFLOPs per 3200 W advertised performance would require approximately 1800 DX1s consuming nearly 6 Megawatts (3200 W * $38e^{15}/21.2\ e^{12}$). The human brain on the other hand, requires only 20 W of power.

The challenge facing power designers is multi-faceted. Simply delivering these levels of power is challenging enough. Efficiency is absolutely critical, not only as energy costs are rising, but also as every watt of waste energy dissipated as heat increases the air conditioning requirements in the datacenter, further increasing operational costs and carbon footprint.

Real estate is also rising in cost and, as datacenters contain hundreds or thousands of processing units, size is important. A small reduction in the size of a single unit is replicated many times over, allowing more processing power to be located in the same space as larger solutions. Yet, this smaller size requirement rapidly increases power density and reduces the surface area available for dissipating heat, making thermal management one of the significant challenges in designing power for this new generation of AI super computers.

Computing systems are complex loads; while learning they are running at full power. As the activity drops, so does the power requirement, but the efficiency is required to remain as high as possible throughout the power band. With today's multi-phase power solutions this entails the designer building in provision for controlling the number of phases used to ensure that efficiency is optimized at all times.

3 Digital vs analog control

Clearly, a more sophisticated approach to power design is going to be required to meet the needs of this rapidly growing sector. In order to address this need, Infineon has introduced advanced digital control techniques, replacing the legacy analog-based solutions.

Digital control brings many benefits when designing high-end power solutions, not least overall system flexibility and adaptability. With digital technology, controllers can be customized without the need for expensive and time-consuming silicon spins. The customization extends to defining the configuration, telemetry for gathering system performance data, setting fault management and calibrating the device.

As power systems become more integrated into the overall solution, communication between the power solution and the main CPU / GPU / TPU is a new requirement. Infineon's mature digital controller technology facilitates market-leading solutions and includes a GUI that enables real-time system design, configuration, validation and monitoring.

Digital solutions simplify building the scalable power solutions required for AI. Yet with all of the included functionality and precision delivery of power, they are now price competitive with the analog solutions they are ultimately replacing.

4 Recent power technology advances supporting AI

Infineon is one of the leading designers and manufacturers of advanced power control and switching technologies. Their product offering is highly integrated and provides all of the key silicon elements required to build highly advanced power solutions for AI applications.

Infineon's complete portfolio includes huge breadth of products including digital controllers, integrated power stages, integrated power management ICs, Point-of-Load (POL) converters as well as discrete solutions including driver ICs, power blocks and discrete MOSFETs. The range is built upon Infineon's long history of innovation and comprises multiple market-leading technologies such as OptiMOS™, DrMOS™ and µDrMOS™.

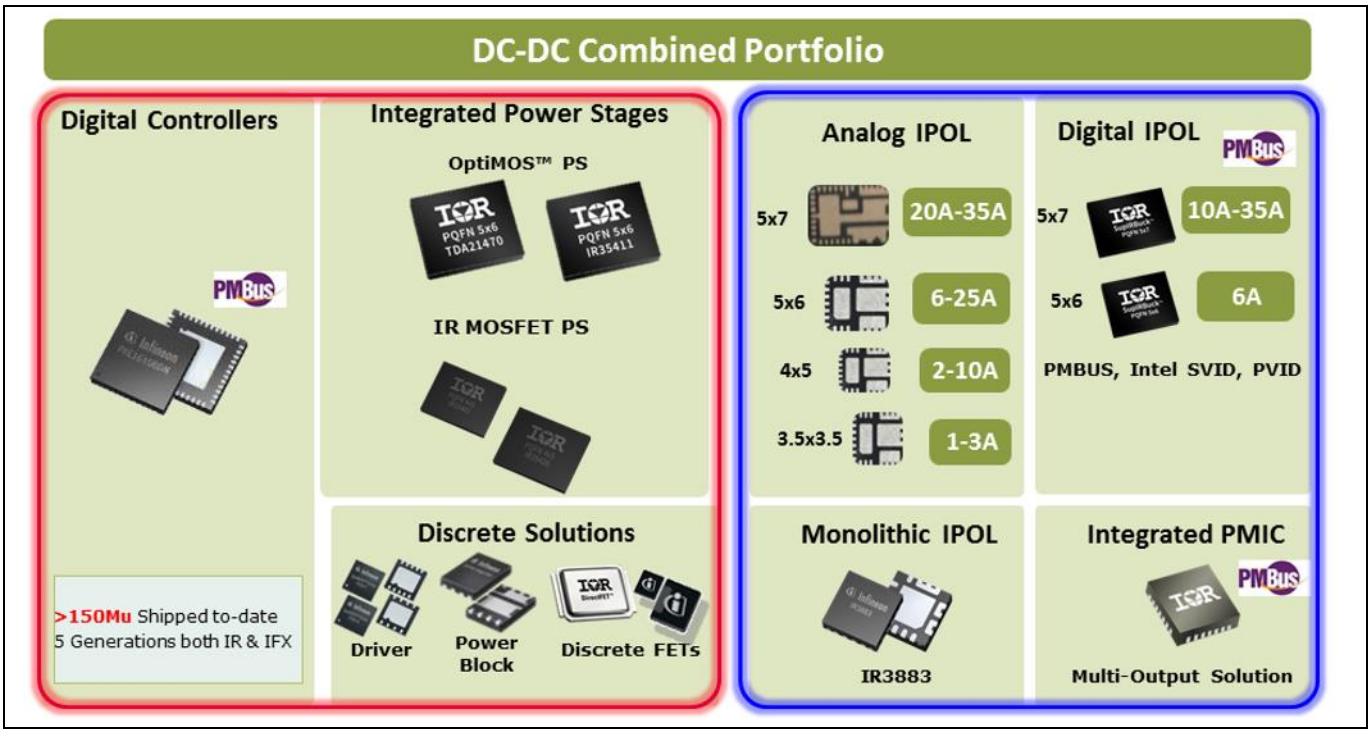

Figure 1 Infineon offers full flexibility in terms of space, performance and cost

4.1 Multi-rail & multiphase digital controllers

Central to Infineon's offering for servers (as well as workstations and high-end desktops) is a complete controller product family of multi-rail / multiphase digital controllers. These advanced controllers are compliant to Intel®, AMD® and support PMBUS with AVS (Adaptive Voltage Scaling) for voltage set-point control and system telemetry with up to 50 MHz max operation.

Infineon solutions are programmable to provide one, two or three fully digitally controlled voltage rails with up to 10 phases. Infineon also offers a family of doubling ICs and Drivers to further increase phase count.

Part Number	Max Phase Configuration	Communication Protocol
IR35219	2 Loop, 10+0 to 5+5	PMBUS with AVS, I2C
IR35221	2 Loop, 8+0 to 4+4	PMBUS with AVS, I2C
IR35217	2 Loop, 8+0 to 7+1	SVI2, SVID, PMBUS, I2C
IR35215	2 Loop, 8+0 to 4+4	SVID, PMBUS, I2C
IR35207	2 loop, 3+1 (memory control)	SVID, PMBUS, I2C
PXE1610CDM	2 Loop, 6+1	SVID, PMBUS, I2C
PXM1310CDM	2 loop, 3+1 (memory control)	SVID, PMBUS, I2C
PXM1330CDN	2 loop, 3+3 (memory control)	SVID, PMBUS, I2C
IR3598	Doubler/Quad Driver for discrete MOSFETs	
IR3599	Doubler/Quad IC for Powerstages	

Figure 2 Infineon's range of advanced controllers are compliant with Intel® and AMD® standards and also support PMBUS with AVS (Adaptive Voltage Scaling)

Efficiency across a wide load range is supported through the ability of designers to program autonomous phase addition or shedding. Other programmable features include PID loop compensation, loadline slope and offset as well as digital temperature compensation.

External load line setting components are eliminated by the digitally programmable load line. The controller is designed to work with RDS(ON) & DCR current sense power stages and provides accurate input and output current reporting.

Digital control enables proprietary non-linear control algorithms and provides excellent transient response with reduced output capacitance. Most of our controllers also support programmable cycle-by-cycle per phase current limit for superior dynamic current limiting.

These devices are easily configurable using our optimized Graphical User Interface (GUI) tools with final configuration settings that can be stored in our digital controller's on-chip non-volatile memory.

As would be expected of a sophisticated controller, significant fault detection and protection is in-built including IUVP, IOVP, CFP, OUVP and OOVP (Input Undervoltage Protection, Input Overvoltage Protection, Catastrophic Fault Protection, Output Undervoltage Protection and Output Overvoltage Protection). Over current protection (OCP) is provided as an instantaneous value, averaged for total current, by channel as well as pulse-to-pulse. There are multiple Over Temperature Protection (OTP) thresholds (internal and external) as well as open/short voltage sense line detection and negative current limit protection.

In some of Infineon's latest controllers, the combined state-machine and integrated microcontroller core architecture allow for maximum flexibility and the internal Non-Volatile Memory (NVM) stores the parameters of any custom configurations.

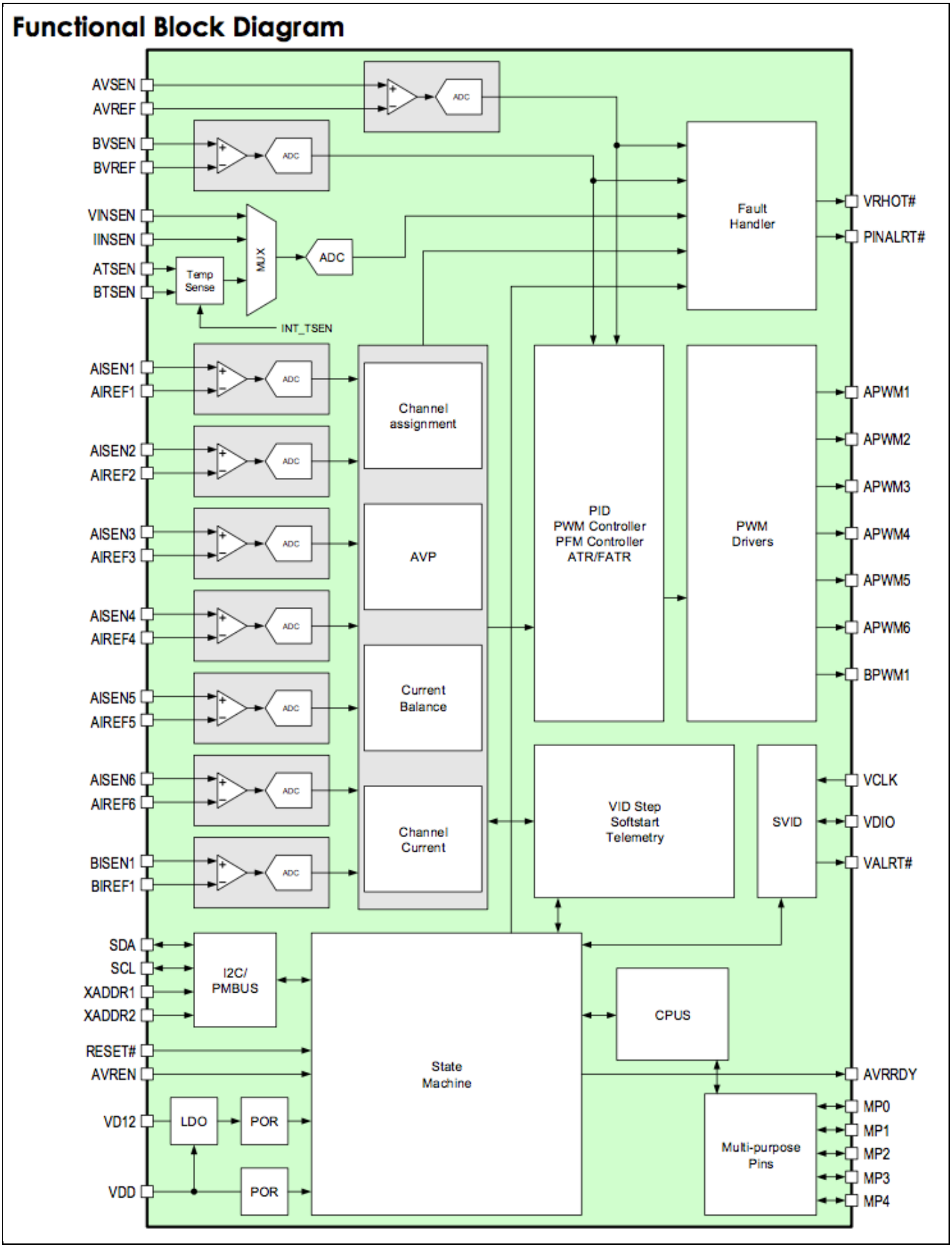

Figure 3 IR35215 block diagram

4.2 IR35411 and TDA21470 OptiMOS™ power stages

The IR35411 power stage contains a low quiescent current synchronous buck gate-driver IC, high-side and low-side MOSFETs and a Schottky diode in the same package to further improve efficiency. The package is optimized for PCB layout, heat transfer, driver/MOSFET control timing, and minimal switch node ringing when layout guidelines are followed. The paired gate driver and MOSFET combination enables higher efficiency at lower output voltages required by cutting edge CPU, GPU and DDR memory designs.

The IR35411 internal MOSFET current sense algorithm with temperature compensation achieves superior current sense accuracy versus best-in-class controller based inductor DCR sense methods. Protection includes cycle-by-cycle OCP with programmable threshold, VCC/VDRV UVLO protection, phase fault detection, IC temperature reporting and thermal shutdown.

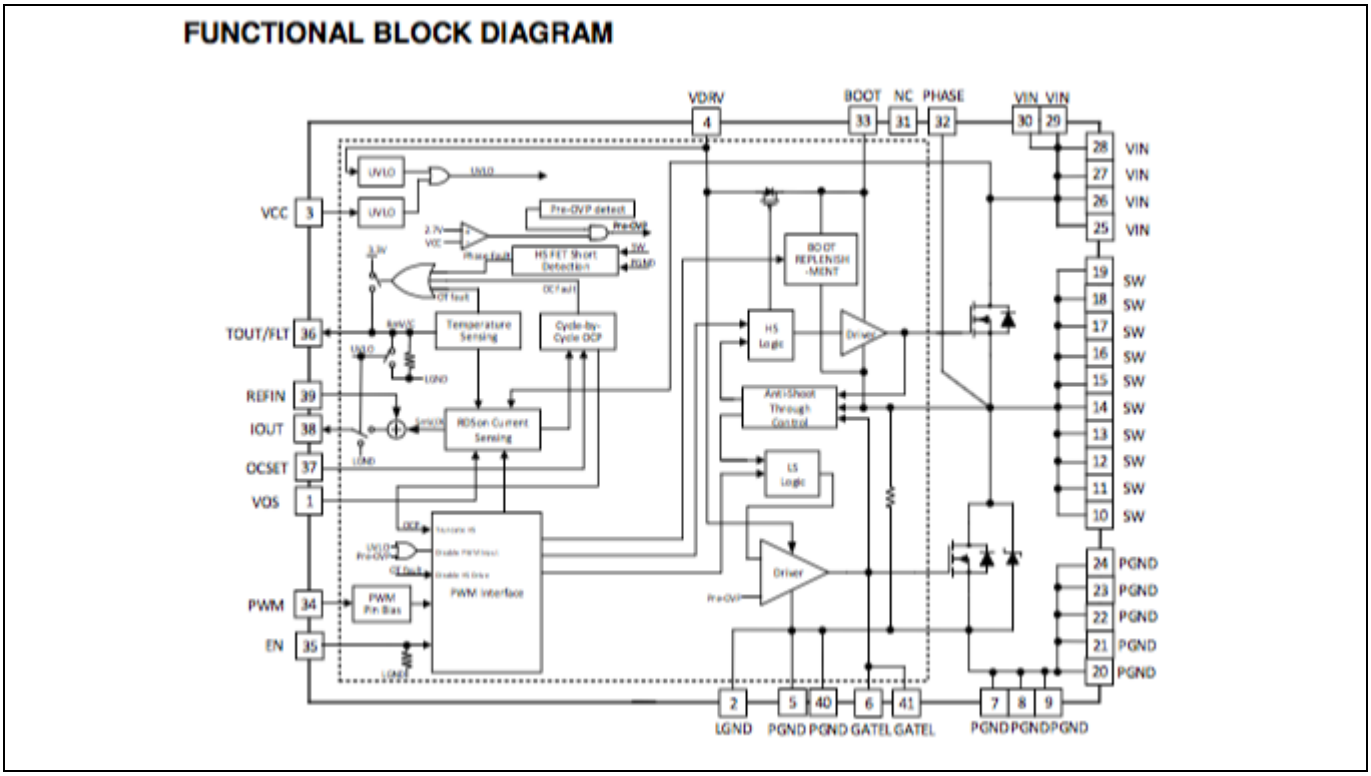

Figure 4 IR35411 block diagram

The IR35411 features deep-sleep power saving mode, which greatly reduces the power consumption when the multiphase system enters PS3/PS4 mode.

Operation of up to 1.5 MHz switching frequency enables high performance transient response, allowing miniaturization of output inductors, as well as input and output capacitors while maintaining industry-leading efficiency.

When combined with Infineon's digital controllers, the IR35411 incorporates the Body-Braking™ feature through PWM tri-state that enables reduction of output capacitors. This quickly disables both MOSFETs in order to enhance transient performance or provide a high impedance output. The IR35411 is optimized for CPU core and memory power delivery in server applications.

The IR35411 is an ideal companion to the IR35215 multi-phase controller. Figure 5 below shows how the IR35215 combines with four IR35411s to create a VR power stage in a 6+1 configuration.

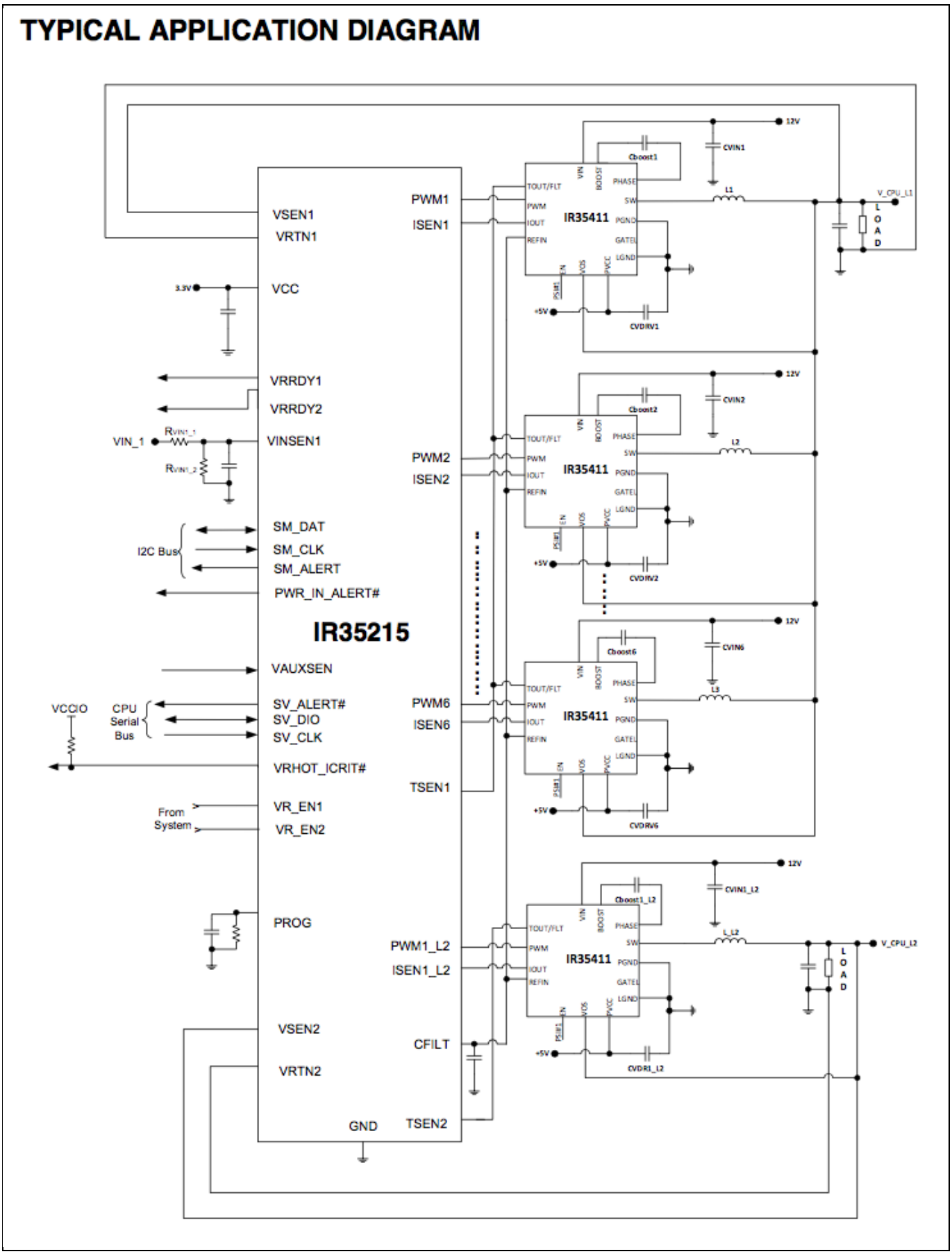

Figure 5 VR using IR35215 controller and IR35411 power stage in 6+1 configuration

5 Summary

While AI is still early in its development, it is already being recognized to be an important and rapidly growing application with expected substantial impacts on our societies. These pioneering AI algorithms are enabled through several high performance computer systems that are challenging designers on many fronts. The traditional datacenter designs are rapidly migrating from general-purpose CPU-only solutions towards combinations of CPUs and GPUs or TPUs, bringing new and more stringent demands on design of server power solutions.

Infineon offers industry's highest efficiency power stages that utilize Infineon's market leading OptiMOS™ MOSFET technology. Through continued Infineon advancements in its power semiconductor technology, our devices are becoming increasingly efficient resulting in continued power loss reductions while increasing our solution densities.

Infineon digital controllers bring unprecedented flexibility and adaptability as well as precise control, telemetry and protection features. As a leader in this AI power delivery market, Infineon offers a broad range of controllers and OptiMOS™ power stages that can support all known AI hardware platforms and their demanding current levels. Infineon enables designers to create state-of-the-art power solutions with highest efficiency and power density for today's high power AI applications as well as future needs.

Published by
Infineon Technologies AG
85579 Neubiberg, Germany

Order Number: B117-I0459-V1-7600-EU-EC