Power Delivery in a Modern Processor

Power delivery is one of the most significant challenges in modern processors. The power delivery network (PDN) must meet the demanding requirements of modern CMOS technology, supply power with excellent efficiency, and swiftly respond to changes in power draw.

These problems hold for 1W smartphones up to 200W server processors and massive machine learning accelerators like Cerebras' 15kW wafer-scale CS-1. To run at the target clock frequency every transistor and circuit in a modern chip requires power supplied at the right voltage. If the supply voltage is too low the circuits will switch slowly and fail – producing incorrect results, stability problems, and other peculiar failures .

Generally, CMOS logic tends to operate at around 1V due to the physics of silicon. However, modern process technologies using FinFET transistors and other techniques have nominal voltages between 0.65V and 1.2V. Innovative circuit designs can use a supply voltage that is close to the transistor threshold voltage, a technique demonstrated by Intel's near-threshold voltage research. While processors using NTV (e.g., Ambiq Micro) recently entered production, it is generally quite novel. The power consumption of a switching circuit (such as a processor) is proportional to the square of the voltage, so reducing the voltage is critical for boosting efficiency. For chip designers this is a classic Goldilocks problem: the voltage should be just high enough to avoid errors and no greater.

However, low voltage operation is a challenge for power delivery because it requires delivering large currents to the processor. Take a modern server processor like Intel's 14nm Cascade Lake Xeons. High-end Xeons have a TDP of 205W, which conceptually translates into a current of 205A at 1V. In reality processors are far more complex and have many different voltage domains and power supplies, but a simpler example is useful. Keeping the power consumption the same and reducing the voltage to 0.75V would increase the current required to 274A . While Intel's high-end server processors are fairly power-hungry, they are easier to deal with than some accelerators. For example, Nvidia's Volta V100 is rated for 450W, some future processors are expected to hit 600W, and as mentioned, Cerebras' CS-1 is an astounding 15kW.

Generally, when transmitting power it is much more efficient to use a high voltage and a low current. At higher voltage the current is lower and fewer wires are needed to transmit power, which reduces costs. Additionally, the resistive losses are proportional to the square of the current, so increasing the voltage and reducing the current will reduce resistive losses and improve energy efficiency. This is the reason why long distance power lines typically operate above 110kV, and the same general concepts apply within a server or data center. While some servers use traditional 12V power supply, some newer designs use 48V for efficiency – and this is particularly common for high-power accelerators that draw over 350W.

Putting these factors together, the conceptual goal of power delivery is to move power throughout the system at as a high a voltage as realistic for transmission efficiency, and then convert down to a very low and stable voltage for efficient and correct computation.

Anatomy of a Power Delivery Network

As Figure 1 illustrates, power delivery is a full system problem that starts all the way at the main power supply and extends to a power grid in the processor that ultimately reaches the transistors that perform computations on die. For a desktop, the power supply will convert from 110V or 220V AC to a 12V DC current that is distributed across the motherboard to the processor and other components. In a notebook or smartphone, things will be a bit different; typical lithium-ion batteries output a 3.7V DC current so there is no AC to DC power conversion and the conversion ratio is much lower.

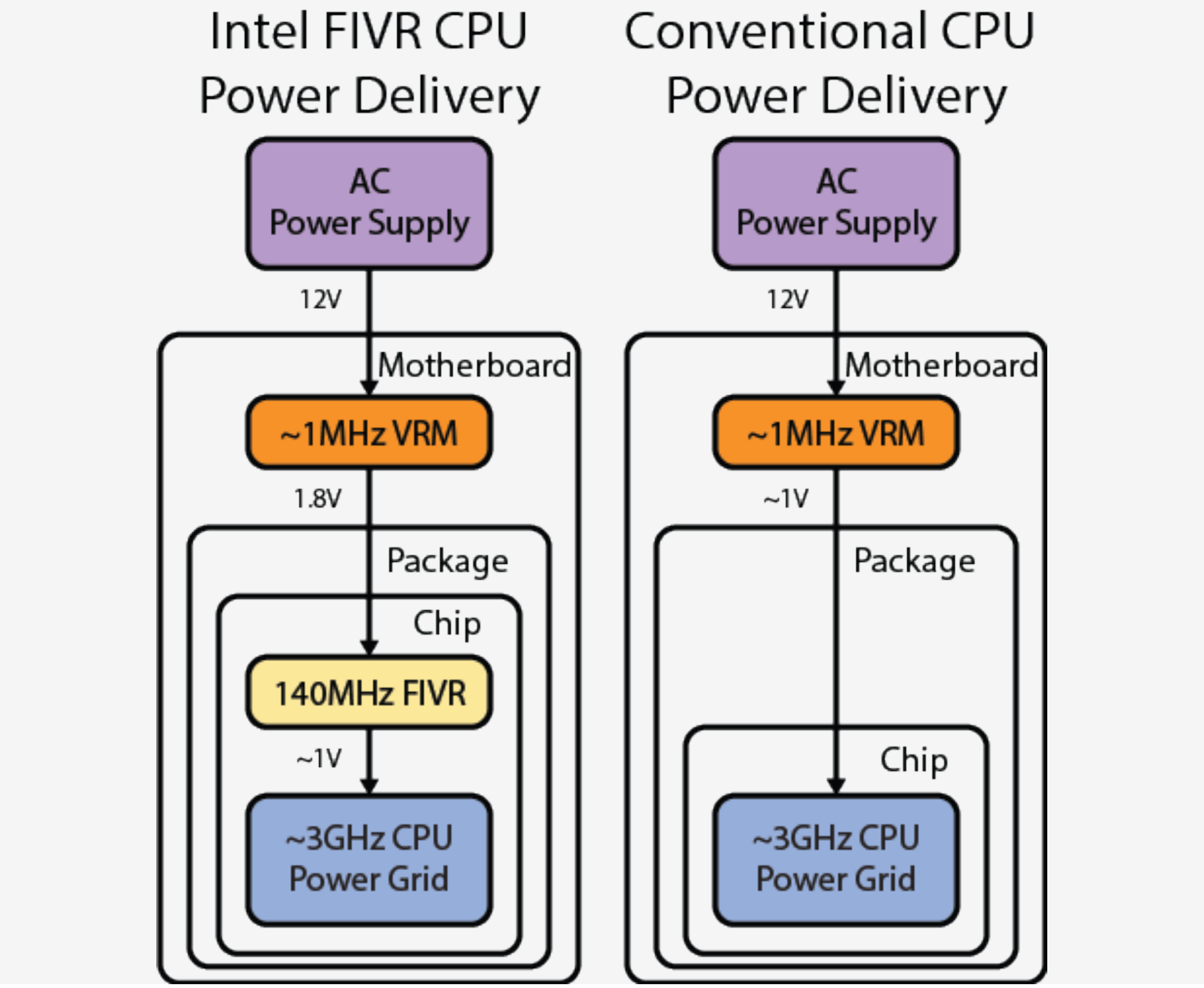

Figure 1. Power delivery in modern systems using Intel's FIVR (left) and conventional VRMs (right).

For standard processors, such as those from AMD, the voltage regulation modules (VRMs) convert down to about 1V. Generally, the VRMs are placed in close proximity to the processor so that most of the distance that the power travels will be using the 12V signals on the motherboard. The 1V power supply is transmitted a short distance across the motherboard, through the processor package and into the processor itself through a set of bumps. The processor contains a power grid that fans out from the bumps and uses the various metal interconnect layers to deliver power to the transistors on die. Motherboard voltage regulators are fairly slow and operate at around 1MHz, meaning that the VRM can only adjust the output voltage every microsecond.

Many Intel-based systems follow the same principles, but with an extra stage in the power delivery. The FIVR or fully-integrated voltage regulator is integrated into the processor die itself and fans out to dozens of power rails on different blocks (e.g., CPU cores, CPU L2 caches, different GPU blocks, etc.). The FIVR is used in most server processors starting with the Haswell generation. It is also used in the Haswell and Broadwell client processors and now Ice Lake and Tiger Lake client. Note that the Skylake client family (including Coffee Lake, Comet Lake, etc.) do not use the FIVR. In these systems, the motherboard VRMs convert the 12V (or 48V) signal down to about 1.8V, which is transmitted from the VRMs, across the motherboard, the package, and the processor pins and into the FIVR. The FIVR is responsible for the last stage of power conversion and finally converts the voltage from around 1.8V down to around 1V, depending on the needs of the particular power rail. One advantage of the FIVR is that the voltage delivered from the motherboard VRMs to the processor is about twice as high as in a conventional system. Using a higher voltage reduces the necessary current by a similar factor of two, reduces the number of power pins, and boosts efficiency. The downside is that voltage conversion is never 100% efficient and the FIVR does lose some efficiency as a result. The relationship between the transmission efficiency gains and conversion efficiency losses is highly dependent on the exact situation. Overall for high-power processors, the FIVR appears to be a win. Additionally, the FIVR is amazingly fast – it operates at 140MHz, two orders of magnitude faster than a motherboard VRM.

Swift Responses Needed to Dynamic Conditions

The speed of the FIVR hints at one of the biggest challenges in power delivery for modern processors. Focusing on steady-state power and thermal characteristics (e.g., TDP) understates the magnitude of the power delivery problem. Modern processors are extremely dynamic and their behavior changes based on the workload. When a transistor switches, it requires a relatively small amount of current. However, if many transistors switch simultaneously the total current draw can become significant and create noise on the on-chip power supply. In high-speed designs like a CPU or GPU, the number of transistors switching can change dramatically from cycle to cycle. For example, when a CPU core starts executing AVX512 multiply-accumulate operations, the power draw is much greater than simply executing integer arithmetic. Similarly, dynamic voltage and frequency scaling systems (DVFS) will change the processor frequency and voltage on the fly in response to changes in the workload or operating conditions. These sudden spikes in current draw can cause the voltage to temporarily dip (referred to as a droop).

Two examples help to illustrate this challenge. Most data centers are optimized for efficiency and high utilization – which translates into 40-60% CPU utilization for the processor, with bursts that go even higher. Returning to the 205W TDP Intel Xeon datasheet, the processor is rated for a maximum current draw of 273.75A across the major voltage rails and an incredible 413W power draw at the package level to deal with peak demand.

Client processors, especially for notebooks and smartphones are an entirely different story, and even more challenging. They are typically optimized for very burst behavior and must deliver maximum performance for short periods of time (e.g., loading a webpage), while drawing practically no power during idle (e.g., waiting for user input). A notebook that is operating at 40-60% CPU utilization would have a tremendously short battery life and a client processor probably spends about 90% of its time at idle. The overall result is that client processors have an even greater discrepancy between the TDP and maximum power and current draw. The most recent Ice Lake U-series and Y-series processors are rated for 15W and 9W TDP. To deliver greater performance, system vendors can configure the TDP as high as 25W and 12W respectively. However, the maximum rated current draw for the CPU and GPU is dramatically higher – up to 70A and 49A respectively, and this excludes the power for the memory controller and system agent.

The crux of the challenge is that voltage regulators, whether motherboard VRMs or Intel's FIVR, are much slower than the transient current spikes caused by switching activity. The Haswell FIVR can ramp up a power rail (e.g., to a core) from 0V to 0.8V in about 0.32 microseconds. However, for a modern 3GHz design, that translates into roughly 1000 clock cycles. Slower conventional VRMs can increase the voltage at about 10-25mV per microsecond, and would take around 100X longer to make a similar ramp from 0V to 0.8V, or about 100K clock cycles. Without proper design, these transient spikes can effectively cause a voltage droop and brown-out, conceptually similar to how in older homes the lights dim when a microwave or hair-dryer starts. As an aside, Skylake client processors and AMD processors use an on-die low-dropout (LDO) voltage regulator that is also very fast. However, LDOs act as a variable strength resistor and only reduce voltage to a power rail. Because LDOs operate through resistance, they tend to be inefficient for larger changes in voltage (E.g., >10% voltage reduction).

As discussed previously, if a processor is operating at 3GHz and the voltage suddenly drops then the transistors may no longer correctly function, so either the voltage must be kept constant or the frequency must drop. In practice, most companies employ a combination of several techniques. For example, AMD developed an adaptive clocking technique that reduces the frequency during some voltage droops.

Decoupling Capacitors Smoothly Deliver Power

To address the mismatch between nearly instantaneous transient current spikes and the latency of voltage regulators, modern systems rely on decoupling or bypass capacitors. These capacitors store energy and can quickly provide power to ensure a consistent voltage while the regulators are just beginning to respond. Returning to Figure 1, systems include decoupling capacitors (or decaps) at every step of the power delivery network. Motherboards typically include capacitors in many places, but especially around the socket as illustrated in Figure 2. Processor packages also incorporate decoupling capacitors, typically around the edges and on the underside. Lastly, processors use a variety of on-die capacitors; these are the closest to the active circuits and provide the fastest response times to transients.

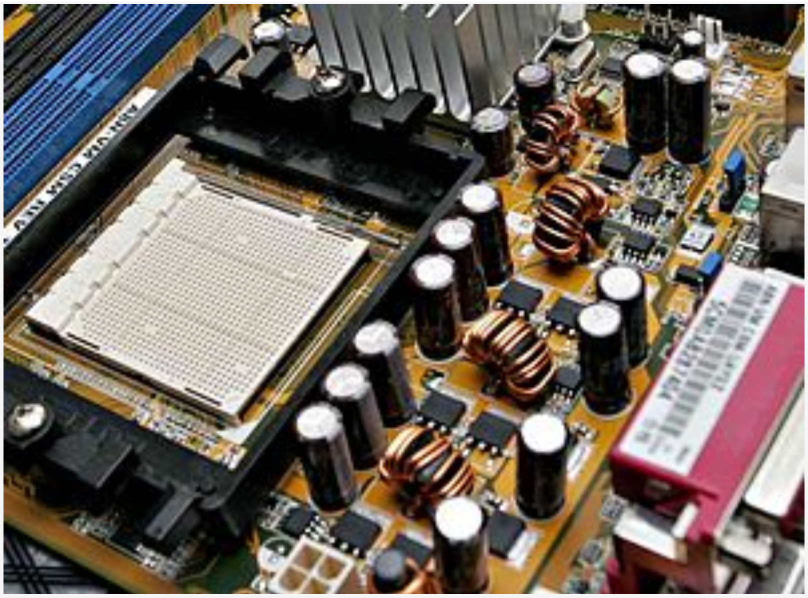

Figure 2. Decoupling capacitors around processor socket.

On-die capacitors are available in a variety of flavors. The simplest type of on-die capacitor is a regular transistor, sometimes called a MOSFET capacitor. These decaps can be easily inserted in standard cells very close to the crucial areas that are expected to have high switching noise. Since the decaps are close to the switching activity, they can easily absorb the noise and quickly supply additional current.

Additionally, chips that are designed using automated tools tend to have whitespace, or areas that are empty due to the imperfect nature of the tools and constraints on placing differently shaped blocks close together. Filling up whitespace with decaps is a fairly common technique, since it is conceptually free. While MOSFET capacitors are available in any digital process technology and are simple to place, they are not ideal capacitors. Like any other transistors, they leak and also can be challenging to fit in areas of a chip that are very congested. Another approach is to modify the process technology and create more specialized structures such as metal-insulator-metal (MIM) capacitors, metal-oxide-metal (MOM) capacitors, or deep trench capacitors.

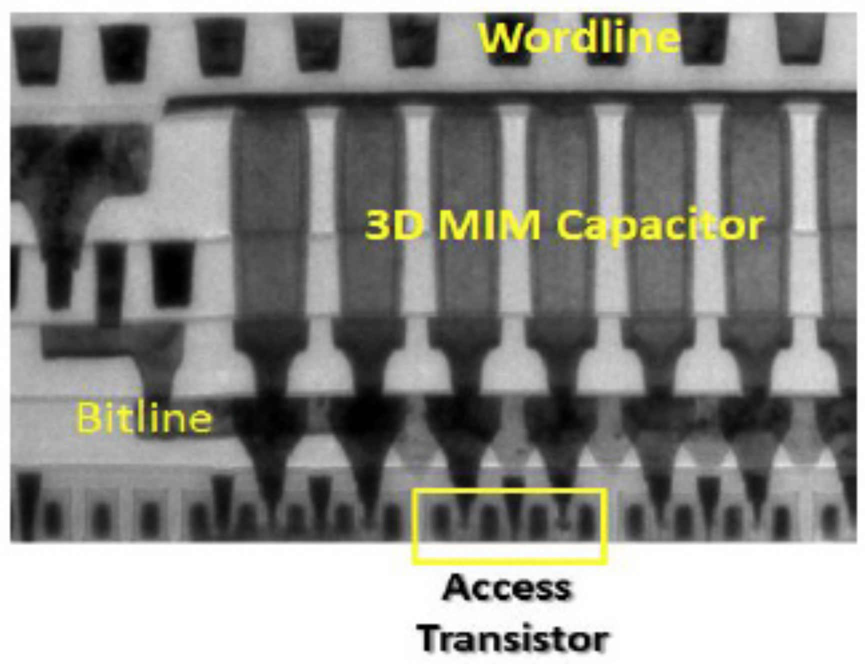

Figure 3. Intel 22nm MIM capacitors for eDRAM.

As the name implies, a MIM capacitor is formed by two parallel metal layers with an insulating high-k dielectric between them. Intel's 22nm process family includes two different types of MIM capacitors. As illustrated in Figure 3, the first type of MIM capacitor is used for a bit-cell in eDRAM and is formed in the lower M2-M4 metal layers. The second is featured in Intel's 22FFL process and uses the extremely thick 4μm top layers for the parallel metal layers. Intel is hardly unique here – other manufacturers also offer MIM capacitors. For example, AMD employed an upper level MIM capacitor in the Zen CCX for decoupling to reduce voltage droop. MIM capacitors are generally superior to MOSFET decaps, but they are slightly farther away because they are in upper metal layers and the extra formation steps increase the cost of manufacturing a bit. MOM capacitors use the same general idea of parallel metal lines, but turned 90°. The metal lines are formed horizontally within two adjacent vertical metal layers (e.g., M3 and M4), with the interlayer dielectric oxide acting as the insulator.

Deep trench capacitors are another option, but are fairly uncommon in a logic process, because etching out the high aspect ratio trenches adds significant cost to the process. Deep trench capacitors were used for several generations of processor technology starting with IBM's 32nm SOI logic process and continuing down to the 14nm SOI process. IBM's deep trench capacitors are used for large eDRAM arrays, which implement various L2, L3, and L4 caches of for POWER and zArch processors, and they are also used for decoupling as well. As an example, IBM claims they were able to eliminate all package capacitors in the 32nm IBM mainframe z12 processor, replacing them with on-die deep trench capacitors. More recently at IEDM 2019, TSMC described forming deep trench capacitors in a silicon interposer. This is a clever and elegant approach, although these capacitors are not as close to the active logic as an on-die solution and therefore cannot completely replace on-die decoupling capacitors.

System Power Delivery Balances Performance, Efficiency and Cost

Power delivery for high-performance processors must navigate myriad challenges. Ideally, the power delivery network primarily operates at very high voltage for transmission efficiency, but ultimately provides a low and stable supply voltage to the CMOS logic that implements the processor. The power conversion, from AC to DC and from high-voltage to low-voltage should be as efficient as possible.

At the same time, the current required by the processor is changing constantly in response to dynamic conditions such as the mix of instructions or the DVFS system. To buffer against these near instantaneous changes and reduce power supply noise, modern designs use decoupling capacitors at nearly every level of the power delivery network from the motherboard to the processor die itself. Faster and more responsive power delivery networks require less decoupling capacitance. Zooming into the processor itself, there are a number of on-die capacitor options. The simplest is using regular transistors, which are easy to place and work on any process technology but are fairly inefficient. However, many semiconductor manufacturers also offer superior capacitors built with special process technologies or circuit designs such as MIM capacitors in the metal layers and less commonly, deep trench capacitors in the silicon or interposer substrate.

These variables are all related: process technology, decoupling capacitors, DVFS systems, voltage regulators, etc. and processor designers must account for all of them to deliver the best performance, efficiency, and cost.