HOW2POWER TODAY
Your Power Design Newsletter

Exclusive Technology Feature

ISSUE: October 2018

Energy-Based Efficiency Metric Helps To Optimize Server Power Delivery For Dynamic Workloads

by Viktor Vogman, Power Conversion Consulting, Olympia, Wash.

Server workloads are usually highly dynamic and often spikey. Designing a server power train with continuous power ratings equal to or even exceeding workload peak power levels is not always energy efficient and cost efficient. This article studies opportunities for power architecture optimization based on an efficiency metric that accounts for dynamic energy usage, introduces potential power delivery solutions, and discusses tradeoffs for corner cases.

Background On Server PSU Energy Efficiency

Energy conversion efficiency is defined as the ratio of useful energy transferred by a conversion device to total energy supplied to the device. When output power and input power for a device are constant the efficiency can be defined as the ratio of output power to input power.

For devices in which input and output power are variable, e.g. dc-dc converters and power supplies used in computer/server systems, energy efficiency is not typically represented by a single dimensionless number. This complicates their relative performance assessment.

In an effort to promote energy efficiency, the voluntary certification program 80Plus was established to certify computer and server system power supply units (PSUs) that have more than 80% energy efficiency at certain specified percentages of rated loads.[1] 80Plus certified PSUs have become the market (and industry) standards and the 80Plus certifications are now being widely used as reference efficiency levels along with the more detailed static efficiency curves characterizing the power ratio as a function of output power.

Despite the great progress made by the 80Plus program, comparison of different energy converting devices for variable power cases remains ambiguous. For example, if one converter or PSU is expected to have higher efficiency at low power levels and the other at high power levels, which one can be considered more energy efficient and recommended as a superior design?

Dynamic Efficiency Metric

Conversion power losses are not a linear function of output power, therefore PSU efficiencies in dynamic modes (at average power) significantly differ from the levels provided by a static efficiency curve. In cases of variable power consumption, typical for datacenter applications, the classic energy-based definition helps to resolve the ambiguity. Using transferred (E_o) and supplied (E_{in}) energies in the following efficiency equation allows us to characterize efficiency *Eff* in such cases as one dimensionless number.

$$Eff = \frac{E_o}{E_{in}} = \frac{\int_0^T P_o(t)dt}{\int_0^T P_{in}(t)dt},$$

where $P_o(t)$ and $P_{in}(t)$ are continuous output and input power signals respectively, and T is operation time.

The process of characterizing conversion efficiency for such applications is illustrated in Fig. 1. It starts with acquiring the typical (benchmark) workload power profile, which is essentially a continuous output power signal $P_o(t)$ over some validation time interval T_v sufficient for the workload characterization. Then the digitized power signal gets processed and rearranged into a histogram $P_{o.eqv}(t)$ (Fig. 1)—a level-varying step function representing portions of energy supplied to the load at given rates (power levels P_{oi}).

Besides the power levels, the power histogram in Fig. 1 displays the time periods T_i over the course of which the system operates at power level P_{oi}. This data array can be used to compute the output energy E_{oi} transferred at each power level.

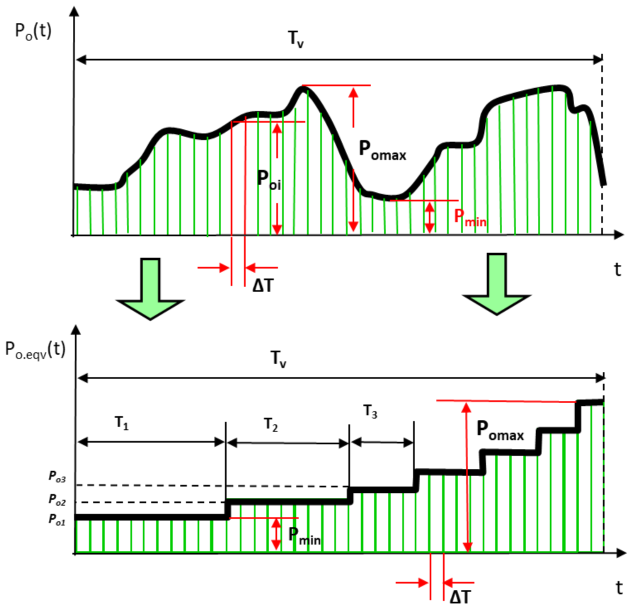

Fig. 1. A digitized continuous power subsystem output power signal $P_o(t)$ over validation time interval T_v can be used to create a power histogram $P_{o.eqv}(t)$ displaying the time periods T_i over which the system operates at each power level P_{oi}. This data can be used for accurate energy-based efficiency computation in a variable-power operating mode.

Since each energy portion gets converted at a different efficiency (depending on the power level and static efficiency curve), the power histogram data can be used to compute corresponding input energies $E_{in.i}$, which facilitates the calculation of a single number for conversion efficiency *Eff* for a given PSU and load profile.

Numerical integration of the digitized power signal over time would yield energy. Thus, the energy-based efficiency for a variable-power operating mode can be computed using the following equation:

$$Eff = \frac{E_o}{E_{in}} = \frac{\sum_{i=1}^{n} E_{oi}}{\sum_{i=1}^{n} E_{in.i}} = \frac{\sum_{i=1}^{n} P_{oi} \cdot T_i}{\sum_{i=1}^{n} P_{oi} \cdot T_i / Eff(P_{oi})} = \frac{\sum_{i=1}^{n} P_{oi} \cdot D_i}{\sum_{i=1}^{n} P_{oi} \cdot D_i / Eff(P_{oi})} \tag{1}$$

where $Eff(P_{oi})$ is the efficiency level array from the static efficiency curve, T_i is the specific power level time interval, P_{oi} is a power reading taken over a short time ΔT (Fig. 1), and D_i is the duty cycle for each power level P_{oi}. ($D_i = T_i/T_v$).

Note that if a system can operate in a mode that does not draw any power from the PSU, i.e. $P_{o1} = 0$, then a first component in the denominator of the above equation takes an indeterminate (zero/zero) form, which can be evaluated by substituting in the expression for efficiency as follows:

$$\frac{P_{o1}}{Eff_1} = \frac{P_{o1}}{P_{o1}/(P_{o1} + P_{loss.1})} = P_{loss.1}$$

where $P_{loss.1}$ is the PSU power loss (self power consumption) at zero load.

Some server PSUs report their real-time power, so in many cases datacenter operators have the ability to collect and evaluate power information without using any external equipment. In case the monitoring accuracy or sampling rate is not adequate, the output power histogram can be obtained by processing input power data from a digital power meter placed at the server power input, and then taking values for power supply static efficiency from the curve that is typically supplied by the PSU manufacturer. For input power data, equation (1) can be rewritten as follows:

$$Eff = \frac{E_o}{E_{in}} = \frac{\sum_{i=1}^{n} E_{oi}}{\sum_{i=1}^{n} E_{in.i}} = \frac{\sum_{i=1}^{n} P_{in.i} \cdot Eff(P_{in.i}) \cdot T_i}{\sum_{i=1}^{n} P_{in.i} \cdot T_i}$$

where *Eff*(*P*in.i) is the efficiency level array from the static efficiency curve plotted vs. input power.

Switching losses in the PSU are proportional to the output voltage of its PFC stage. Higher output PFC voltage increases switching losses, which is most noticeable at light loads. To meet 80Plus light-load efficiency requirements, the PFC voltage typically gets adjusted (reduced) as output power reduces.[3]

With highly dynamic workloads, when load frequencies exceed the PFC stage bandwidth, the PFC voltage control latency may not allow the PSU to achieve static 80Plus efficiency levels P_{oi}. So, in high-frequency load cases or in cases when power supply static efficiency data is not available to the user, the energy-based approach remains the only method for fair comparison of PSU efficiency. In such cases the comparison of different PSU efficiencies inevitably turns into a comparison of energies E_{in} consumed by the PSUs when running identical workloads over the same T_v time interval:

$$E_{in} = \sum_{i=1}^{n} P_{in.i} \cdot T_i$$

How The Dynamic Efficiency Metric Works

Let's see how this dynamic efficiency metric can be applied in real applications. For illustration purposes let’s consider a server system with two redundant PSUs, where the system is fully active 15% of the time and idling 85% of the time. Its idle power is 10% of a single power supply rating, and system peak power is 90% of the PSU rating.

This case is shown in Fig. 2 for redundant 80Plus Titanium efficiency PSUs sharing power equally. Dynamic (actual) efficiency can be computed for this case using equation (1):

$$Eff = \sum_{1}^{2} P_{oi} \cdot D_i / \frac{\sum_{1}^{2} P_{oi} \cdot D_i}{Eff_i} = \frac{0.05 \cdot 0.85 + 0.45 \cdot 0.15}{0.05 \cdot \frac{0.85}{0.59} + 0.45 \cdot \frac{0.15}{0.94}} = 0.765.$$

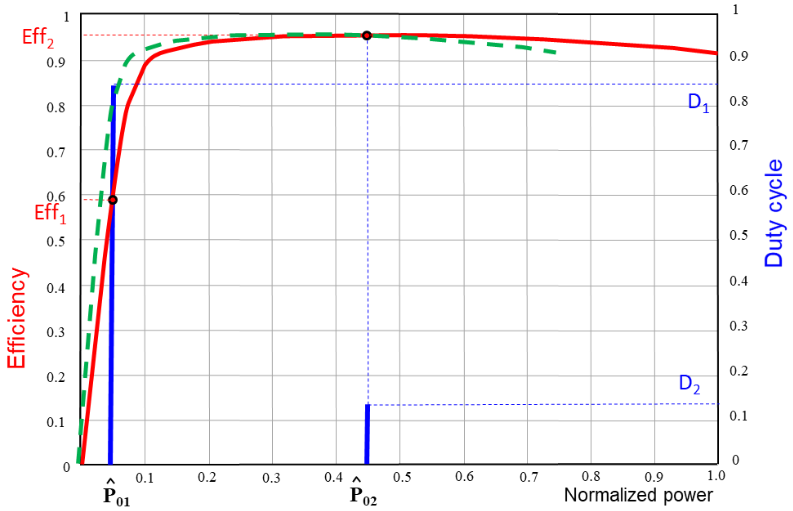

Fig. 2. Illustrative example of a high-efficiency power supply and an associated two-level workload histogram [D_1,D_2,P_{o1}, P_{o2}]. Despite using 80Plus Titanium efficiency PSUs operating in their highest efficiency point (94+%),the actual energy-based efficiency of the power subsystem is only 76.5%. Selecting PSUs with 25% lower continuous rating (this PSU efficiency curve is given by the green dashed line) results in more than 12% higher power-subsystem efficiency.

This example illustrates that despite using power supplies with the highest certified efficiency the actual energy-based efficiency of this power subsystem is relatively low—only 76.5%. Since in redundant mode PSUs share power equally and operate below 50% of their ratings, let's explore what would happen if the PSUs were replaced with lower-rated modules.

If we select PSUs with a 25% lower continuous rating, for which P_{o1} and P_{o2} would be respectively 0.067 and 0.6 of the new PSU rating (its efficiency curve is given by green dashed line in Fig. 2) we would get:

$$Eff = \frac{0.067{\cdot}0.85+0.60{\cdot}0.15}{0.067{\cdot}0.85/0.83+0.60{\cdot}0.15/0.938} = 0.893,$$

or an efficiency gain greater than 12%. This means that for the given workload profile a major gain in efficiency can be achieved with cheaper and smaller-sized power supplies.

This example also shows how the dynamic efficiency metric can be applied to real workloads yielding a single efficiency number for a variable-power operating mode. It also reveals how the metric could be sensitive to small changes in selected power conversion hardware. Similarly, this metric can be used for other power conversion devices, such as intermediate bus converters, CPU/memory voltage regulators, etc.

Optimizing Power Subsystems For Real Workloads

The dynamic efficiency metric supports the intuitively obvious assertion that conversion efficiency gain would be zero for: a PSU having higher efficiency at light loads but used in an application with a workload continuously drawing high power; or a PSU having higher efficiency at heavy loads but used in a system idling all the time. The metric can also show that the optimal solution for a wide-range load operation would be an interleaved load-adaptive PSU with a "flat" efficiency plot. However, in many cases this solution is cost prohibitive, so the metric can be used to quantify alternative (cheaper) hardware options that enable efficiency gains at low power levels.

Let's consider a server system consuming 900 W of peak power and having two redundant power supplies operating in parallel. The workload profile has 10 levels at which the system operates for equal time periods, i.e. duty cycle D_i for each power level equals 0.1 (Fig. 3). Power loss curves for these two PSUs are shown in

Fig. 3 for two cases: 750-W and 1000-W continuous power rating. Power loss is plotted instead of efficiency so as to better show the difference between the two PSUs.

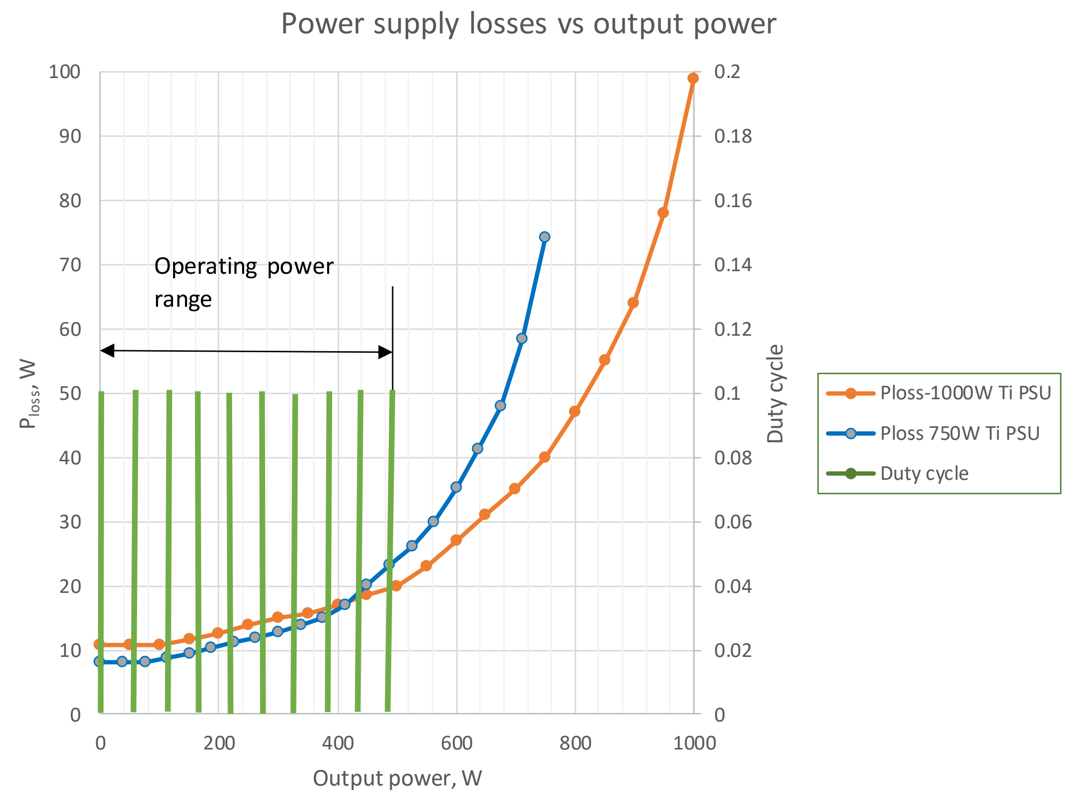

Fig. 3. 750-W and 1000-W 80Plus Titanium PSU P_{loss} curves demonstrate that a cheaper, lower-rated PSU can provide energy savings. Even with a given uniform workload profile (D_1=D_2=…=D_{10}=0.1) the energy-based dynamic efficiency metric computation shows 0.6% conversion efficiency gain with a lower-power-rated PSU.

Even with such a “disadvantageous” workload profile, having comparatively short time intervals of low-power operation, the dynamic energy-based metric shows a 0.6% conversion efficiency gain for the PSU rated for lower power. That efficiency gain would lead to significant energy bill savings at the datacenter level. Although selecting a higher-rated power supply seems to be a safe approach as it covers the full load range with margin, this choice would be associated with a higher total cost of ownership (TCO). This is a result of two factors—about a 30% higher hardware cost (due to higher power rating) and a $6 (at 10₵/kWh) higher annual electricity bill.

These examples show that lower-rated power supplies having the same 80Plus certification level can provide energy savings. At the same time, redundancy can be impacted when using lower-rated power supplies in cases when peak power exceeds the single module rating. When both redundant power supplies are active, a peak power slightly exceeding the continuous rating of one PSU does not present an issue. However, when one PSU fails to claim redundancy the remaining PSU needs to support the peak power level. How can this contradiction be resolved?

Providing The Same Redundancy Level With Lower-Rated PSUs

Essentially there are two operating modes impacting tradeoffs that can be made in selecting more-efficient, lower-rated PSUs: in one mode, peak-power time periods are thermally significant and in the other mode, not.

For thermally insignificant time periods, supporting a higher peak-power level does not have significant impact on PSU component size and cost.

However, the peak-power time period can be considered thermally significant if any component in the PSU can reach its maximum operating temperature and cause tripping of overtemperature protection. Supporting peak power levels for thermally significant time intervals may have major impact on PSU size and cost. To prevent this from happening two additional energy and temperature-based protection levels can be incorporated, using closed loop system throttling technology (CLST).[2]

This control technique was specifically developed for reduction of server power supply size and cost, but it also can be used in other applications, such as dc-dc converters, in which load power can be controlled by alert signals. In a CLST supply, the peak power durations and temperatures of critical components are monitored to determine when system power should be reduced by throttling. As soon as peak power time or a PSU component temperature warning threshold is exceeded, the PSU load gets reduced and the system remains operational. PSU load reduction typically does affect system performance, which is why for thermally significant time periods the system architect may need to consider certain tradeoffs.

These tradeoffs are summarized in the flowchart (Fig. 4).

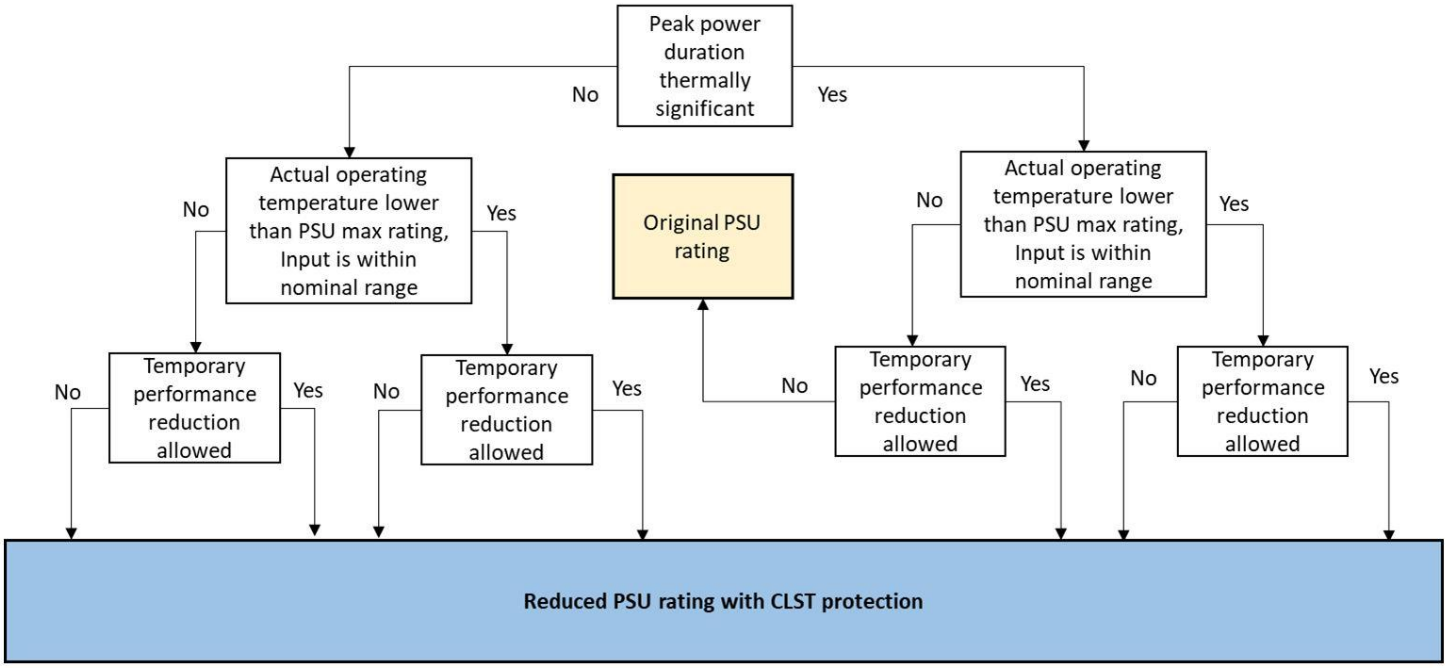
Peak power duration thermally significant
No
Yes
Actual operating temperature lower than PSU max rating, Input is within nominal range
No
Yes
Original PSU rating
Actual operating temperature lower than PSU max rating, Input is within nominal range
No
Yes
Temporary performance reduction allowed
No
Yes
Temporary performance reduction allowed
No
Yes
Temporary performance reduction allowed
No
Yes
Temporary performance reduction allowed
No
Yes
Reduced PSU rating with CLST protection

Fig. 4. CLST tradeoffs.

The flowchart shows that a reduced PSU rating and lower TCO can be achieved in all cases except for the corner case with thermally significant peak-power time periods, highest operating temperature, low input-voltage operation, and in which even a temporary reduction in system performance impact is not allowed.

Dealing With Power Virus Conditions

In real applications a server power train can be overstressed by a power virus—a computer program that executes specific codes to force excessive CPU power consumption, leading to a power supply overload, CPU overheating, or a system crash. Since a power virus is not a useful workload, designing power supplies and on-board voltage regulators to support such a condition is not a cost-effective approach. Therefore, using CLST protection for this case should also be considered justified.

To prevent virus impact on a power supply rating, the PSU needs to have a hard current limit. This limit needs to be set slightly above the real-application peak-power level, P_{max}:

$I_{LIM} \geq P_{max}/V_{omin}$

where V_{omin} is the minimum supplied voltage level.

Since virus detection time and CLST response time are finite, CLST implementation needs a buffer that can supply virus power until CLST is fully activated. In many cases the power supply output or a common bus decoupling capacitor can naturally act as buffers. Since it is a passive element, using the cap as a buffer constitutes the simplest solution.

To determine the required value of this capacitor let's refer to the power delivery equivalent schematic shown in Fig. 5.

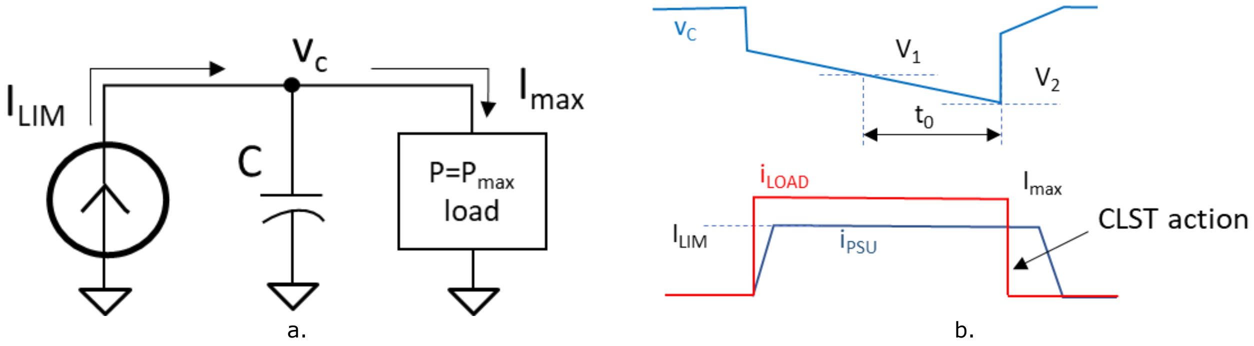

Fig. 5. Equivalent schematic of the power supply in a current limiting mode (a) and a timing diagram illustrating CLST action during a power virus event when the V_1 level is used as a CLST trip threshold (b).

In this diagram a current-limiting PSU is represented as a current source I_{LIM}, connected in parallel with an output cap C that acts as a buffer, and the load operating in a constant-power mode. The energy balance equation for this network can be written as follows:

$$E_{PSU} + E_{CAP} = E_{LOAD}, \tag{2}$$

where E_{PSU} is the energy supplied by the PSU, E_{CAP} is the energy delivered by the buffer cap, E_{LOAD} is the energy consumed by the load. The energy portions provided by the cap and consumed by the load over the CLST response time t_0 can be described by commonly used equations:

$$E_{CAP} = C\frac{V_1^2 - V_2^2}{2} \tag{3}$$

$$E_{LOAD} = P_{max} \cdot t_0 \ , \tag{4}$$

where P_{max} is the virus power level, V_1 is the power supply output voltage or voltage across the cap at normal workload peak current (V_1 is defined by the PSU load line), V_2 is the minimum cap voltage allowed for normal load operation, and t_0 is the CLST response time interval, which includes the virus detection time.

Cap voltage (and its level V_1) can be selected for tripping the CLST. In this case no additional virus detection sensors are needed. To compute the E_{PSU} energy supplied by the PSU during this event, the voltage across the cap as a function of time needs to be determined. It is described by the following equation:

$$v_c(t) = V_1 - \int \frac{(P_{max}/v_c(t) - I_{LIM})dt}{C}$$

This equation can be linearized within the CLST short-response-time interval t_0 during which the V_c change is small as compared to its nominal value and where $V_c(t)$ is replaced with its minimal allowed value V_2. This substitution for $V_c(t)$ in the integral expression yields acceptable accuracy for practical purposes. In this case the expression for energy supplied by the PSU can be written as follows:

$$E_{PSU} = \int_0^{t_0} I_{LIM} V_c(t)dt = \int_0^{t_0} I_{LIM}\left[V_1 - \frac{(P_{max}/V_2 - I_{LIM})\cdot t}{C}\right]dt = I_{LIM}V_1 t_0 - \int_0^{t_0} I_{LIM}\cdot\left[\frac{(P_{max}/V_2 - I_{LIM})\cdot t}{C}\right]dt$$

Multiplying the numerator and denominator in the integrand expression by t_o and noting that $(P_{max}/V_2 - I_{LIM})\cdot t_0/C$ represents a capacitor voltage swing from level V_1 to level V_2, such that $(P_{max}/V_2 - I_{LIM})\cdot \frac{t_0}{C} = V_1 - V_2$, we can simplify the above equation as follows:

$$E_{PSU} = I_{LIM}\cdot\frac{V_1+V_2}{t_0}\cdot\frac{t_0^2}{2} = I_{LIM}\cdot\frac{V_1+V_2}{2}\cdot t_0 \qquad (5)$$

Substituting equations (3), (4) and (5) into the energy balance equation (2) we obtain an expression for the minimum buffer capacitance value needed to keep the output voltage above the allowed minimum V_2:

$$C_{min} = \frac{[2P_{max} - I_{LIM}(V_1+V_2)]t_0}{V_1^2 - V_2^2} \qquad (6)$$

Equation (6) has an obvious physical meaning: it shows that at a given virus power P_{max} and current limit I_{LIM} levels, the lower the minimum allowed voltage V_2 and the shorter the CLST response time t_o, the smaller the buffer capacitance needed.

Example: t_o = 10 µs, V_1=11.8 V, V_2=11.4 V, P_{max} =2.0 kW, I_{LIM} = 100 A: C_{min}=1800 µF

Since in this example virus power exceeds the PSU power limit ($I_{LIM}\cdot V_1$) by almost a factor of two, it proves that with fast CLST response time, output voltage under power virus protection can be still held within spec limits with a reasonable size buffer capacitor.

Conclusion

In many cases a server power delivery architecture can be optimized based on workload profile and an energy-based dynamic efficiency metric. This metric allows accurate comparison of energy efficiency for different power supplies operating in variable-power modes.

The CLST technique for handling various peak-power conditions minimizes the cost and size of a server power subsystem for nearly all applications. The exceptions include a few corner cases for which no performance tradeoffs are acceptable. But with fast CLST response, power virus protection can be provided without any power supply modifications and no impact on power train cost and size.

Future work could focus on developing software for automatic generation of workload power profiles. This would simplify the computation of dynamic efficiency.

References

1. 80 PLUS Certified Power Supplies and Manufacturers.
2. “Closed Loop System Throttling (CLST): A New Power Technology for Improving Server Efficiency” by Brian Griffith, Viktor Vogman, and Andrew Watts, IDF 2010, TMTS001.

3. "Building Variable Output Voltage Boost PFC Converters with the FAN9612 Interleaved BCM PFC Controller," Fairchild Semiconductor application note, AN-8021. Rev.1.0.1, 6/1/10

About The Author

Viktor Vogman currently works at Power Conversion Consulting as an analog design engineer, specializing in the design of various power test tools for ac and dc power delivery applications. Prior to this, he spent over 20 years at Intel, focused on hardware engineering and power delivery architectures. Viktor obtained an MS degree in Radio Communication, Television and Multimedia Technology and a PhD in Power Electronics from the Saint Petersburg University of Telecommunications, Russia. Vogman holds over 50 U.S. and foreign patents and has authored over 20 articles on various aspects of power delivery and analog design.

For more information on power protection techniques, see How2Power's Design Guide, locate the Design Area category and select Power Protection. Also see the Application category and select Data Centers.