

PMBus Takes Command of Data Center Power Issues

By Brian Griffith,
Server Power Delivery Architect,
Intel, DuPont, Wash.



PMBus-capable server power supplies, in conjunction with system-level and data center-level management applications, can lower system operating costs, improve data center efficiency and increase system densities.

Solving data center power problems is an important issue for all manufacturers of servers today. System density (the growing server blades market, for example) and the cost of energy consumption are on everyone's mind. So, it is no surprise the development of products that solve these problems is paramount for any system manufacturer.

To get the highest-possible density and lowest energy consumption in any data center, it is critical to make the most use of all cooling and power-delivery resources. Cooling must be optimized by effectively managing airflow. System power consumption must be controlled to maximize performance without exceeding power-delivery systems. All this must be done while maintaining high availability. To support this demand, tools for monitoring and controlling power at the system, rack and data center levels are being increasingly introduced in today's servers.

PMBus is an open industry-standard communications protocol that is a key component in creating server features to help the data center manager address power issues. Understanding where and how much actual power is consumed at the data center, rack, system and subsys-

tem levels can help optimize data center power-delivery and cooling systems. PMBus power supplies and voltage regulators play a major role in the power monitoring and optimization architecture.

The **figure** shows where these PMBus-enabled power converters are located in the systems throughout the data center. Baseboard PMBus voltage regulators (VRs) that are powering processors and memory offer the ability to monitor power at the subsystem level. System power supplies with PMBus provide system-, rack- and data center-monitoring capabilities.

New system capabilities allow users to control power consumption by limiting performance states of components, such as Advance Configuration and Power Interface (ACPI) performance states of processors. Processor performance states are used to make tradeoffs between performance and power consumption. Details can be found in the ACPI specification.^[1]

Power control capabilities, along with virtualization, can allow data center managers to optimize thermals by controlling hot spots, shifting work loads and maximizing the available resources of systems, power distribution units

(PDUs), cooling units and uninterruptible power-supply (UPS) systems. PMBus power sensors located in the power supply, and on the motherboard, VRs are a key resource used in implementing these capabilities. Standardization around PMBus allows suppliers to focus on improving capabilities and cost, not on making new flavors of identical features.

PMBus commands for use in server power converters can be categorized into three areas: power sensing, thermal management and diagnostics.

Power Sensing

The PMBus commands critical for sensing power consumption in the system power converters are READ_PIN, READ_IIN, READ_POUT and READ_IOUT. The PAGE command in conjunction with the READ_IOUT command is used for converters with multiple outputs.

Each output of a power converter is assigned a PAGE number. **Table 1** summarizes these commands and their associated system management bus (SMBus) protocols. SMBus is used in servers for low-speed communication. PMBus is based on the SMBus revision 2.0 specification.^[2]

Data Formats

The PMBus data format most preferred by system manufacturers is the linear format. This format provides the easiest method for system usage with good enough resolution for any of the power sensors mentioned previously. The PMBus linear format is as follows:

$X = Y \times 2^N$, where Y equals an 11-bit two's complement integer (least significant bits), which is the mantissa; N equals a 5-bit two's complement integer (most significant bits), which is the exponent defining the scaling; and X equals a real-world value.

The system reads both Y and N to determine the real-world value of X. The N value is used to scale the real-world value and should be a constant for any given power supply. Using a constant value for N allows for easier calculations inside the power supply and the system.

The 11-bit Y value provides enough resolution for any system-usage model. **Table 2** shows an example for an 800-W single 12-V output power supply with N scaling values and real-world value range.

Averaging and Accuracy

Accuracy of the PMBus sensors is critical. The power converters are becoming a more integral part of the system, and poor accuracy can have a negative impact on system performance. Because the system operates over a 20% to 100% load range, depending on the system operating state, accuracy over this entire range is

important. Understanding how to sense ac input current and voltage to determine ac input wattage is needed before starting the design.

Sampling the 50/60-Hz ac voltage and current at a high enough frequency to account for distortion is important. All server ac-dc power supplies have active power factor correction (PFC); therefore, this sampling rate can be limited to approximately 5 kHz.

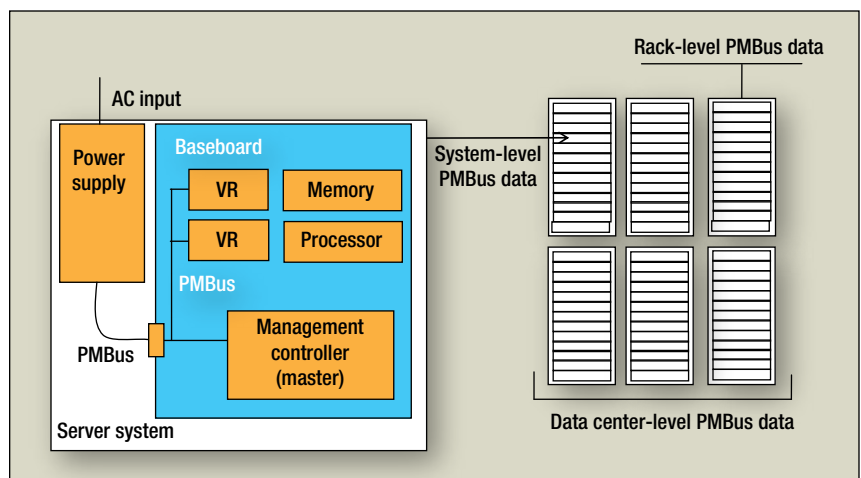
Each PFC implementation is different in how well it limits current harmonic, so each design should measure this and set the appropriate current sampling rate. The loading condition where it is hardest to achieve good accuracy is at high input voltage (240 Vac) and at light loads (20% of maximum). This is the point where the power factor is the worst and where the input current sensor has the lowest signal-to-noise ratio.

The goal of most system manufacturers is to achieve better than $\pm 5\%$ accuracy over 20% to 100% of the converters' operating range. If the best accuracy is required, some manufacturing line calibration may be needed.

Because most server computer systems will poll the converters at a rate equal to about once per second, considerations must be taken to prevent aliasing of the data. The power converter must act as the anti-aliasing filter by averaging the data inside the power supply to prevent error due to slow system-polling rates.

The period of this averaging must be ≥ 1 second to prevent aliasing. This creates a challenge for the power converter. Increasing the system polling rates of the power converters can help to alleviate this challenge. However, it will create its own challenges in the server management design. There are basically three options for averaging sensor data inside the power supply:

- *Infinite impulse response (IIR) filter.* The IIR filter can be implemented with a standard first-order difference equation. Difficulties arise due to the large difference between the sampling rate and the bandwidth. The sampling rate



The ability of ac-dc power supplies and dc-dc converters to communicate using PMBus commands makes it possible to monitor power consumption at the server system, rack and data center levels.

PMBus command	SMBus protocol	PMBus data format	Description
READ_PIN	Read word	Linear	Input power to the power supply in watts. This should be monitored as close to the power supply's input as possible.
READ_IIN	Read word	Linear	Input current to the power supply in amps rms.
READ_POUT	Read word	Linear	Total output power of the main outputs of the power supply.
READ_IOUT	Read word	Linear	Output current in amps for each output. For multi-output power supplies, the PAGE command is used to read the output currents. Smaller outputs like standby outputs may not have a sensor.
PAGE	Read/write byte	—	Used by multiple output power supplies to switch between different outputs.

Table 1. PMBus sensor table.

Sensor	N value (MSB)	2 ^N scaling value	X real-world value	Y value (LSB)
READ_PIN	00100b	2 W	1000 W	001 11110100b
READ_IIN	11010b	0.016 A	11.1 A	010 10110110b

Table 2. Example scaling values for an 800-W 12-V output power supply. Note that the N scaling value should be selected to limit contribution to error due to poor resolution at lighter loads.

is more than 1000 times the filter bandwidth. This creates coefficients in the difference equation very close to 1 and 0, which requires more than a single-byte calculation method. However, this can be managed with the proper coding. The advantage of this method is that it generates a true running average of the sampled data.

- **Arithmetic average.** By simply summing the sensor data over the averaging period and dividing by the averaging period, the average over that period can be determined. Then repeating this for the next period will generate another average value and so on. This will not give a true running average and, therefore, will add some amount of error if the load is very dynamic. The advantage of this method is its simplicity.

- **Running arithmetic average.** By saving all the sensor data points over the average period, newer data can be added to the average and the old data can be dropped from the average. The problem with this method is the need to save large quantities of data: 1000 data points if sampling at 1 kHz. The advantage is the accuracy of a true running average combined with an arithmetic average.

Isolation of Sensors

Since some of the power sensors are located on the primary side of the ac-dc power supply, these ac current, voltage and power sensors must pass information to the secondary-side PMBus microcontroller and still meet

safety regulator requirements. Depending on which method is used to average the data, the speed at which the primary-side sensor data is communicated to the secondary side may be critical.

Standard optocoupler devices used to isolate primary- from secondary-side signals in a power supply are too slow for any of the schemes requiring secondary-side calculations. There are three methods for implementing primary-side sensors:

- **Primary-side calculation.** In this method the input voltage and current are sampled with a primary-side referenced microcontroller. This same microcontroller processes the data to generate the root-mean-square (rms) voltage, rms current and power in watts. The primary-side microcontroller then communicates the resulting data to the secondary. This means there is no need for a fast primary-to-secondary-side communication path. The drawback is the need for a primary-side microcontroller that can perform the needed calculations.

- **Secondary-side sensing and calculation.** In this method, there is no primary-side microcontroller. The primary voltage and current are either passed to the secondary side via isolated sensors, current-sensor transformers or pulse-width-modulation signals over optocouplers. This eliminates the primary-side microcontroller; however, it requires more complicated isolated sensing circuits.

- **Primary-side sampling/secondary-side calculation.** In this method, the primary-side microcontroller mainly serves as the analog-to-digital converter. The raw data is then passed to the primary side for averaging. This keeps the primary-side microcontroller simple; however, it requires a fast, isolated communication method.

READ_PIN for AC-DC Supplies

The capability of reading the ac input power to the power supply (which is also the ac input power to the system) is the most important PMBus sensor to the data center manager and is used in many of the new rack-mounted systems for power management. The ac input power can be used as described in the following three READ_PIN usage models:

- **Control of system power.** By monitoring the total system power, setting limits that the system will not exceed and using the system's power-control capabilities, the data center manager can guarantee that the system, rack and data center

power will not exceed a predefined threshold. This protects data centers from developing hot spots.

- *Optimization of power delivery and cooling equipment.* By monitoring the power consumed by each system, the data center manager can better manage the other resources in the data center to make them more efficient, like computer-room air conditioners used to cool rows of racks and UPSs/PDUs used to deliver power to racks. This helps reduce the cooling system's energy consumption and improve system density.

- *Charge back.* In co-located environments, data center managers may want to bill end users by the system energy consumption. The ac input power monitoring can be used for this billing.

In all these cases, the accuracy of the input power (watts) is critical to ensure there is limited performance impact, reliable protection and proper billing. Good accuracy of the power supply's ac input wattage sensor is a valuable feature to the data center manager.

Calculating the input wattage to the ac-dc power supply is pretty simple. Multiply instantaneous voltage and instantaneous current on the ac wave shapes, then average these values over the needed period. This results in an average of the real input power (watts) to the power supply. The different averaging methods are described in the previous Averaging and Accuracy section on page 15.

READ_IIN, READ_POUT and READ_IOUT

READ_IIN is used by the system to monitor loading levels on PDUs and associated circuits. Rack power is sometimes limited by the circuits powering the rack. By monitoring rms input current, the data center manager can maximize the number of systems in the rack while still protecting against overload conditions on the circuits. This helps to increase system density in the data center.

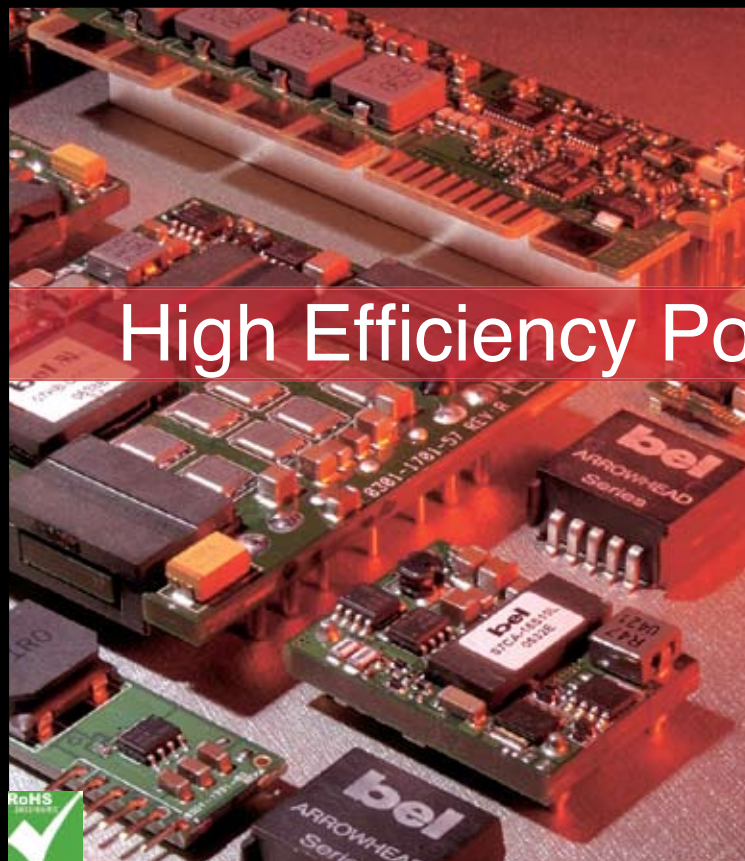
Calculating the rms input current (and rms input voltage) can be done in two ways:

- *True rms calculation.* By saving the data points over at least one cycle of the ac input, the rms current can be calculated by $I_{\text{rms}} = \sqrt{[(X_1 + X_2 + \dots + X_n)/n]}$.

The disadvantage of this method is that more difficult calculations may require a more-expensive primary-side microcontroller. The advantage is accuracy; it generates an accurate rms result independent of distortion on the input current.

- *1.11 × rectified average.* If the input wave shape is a true sine wave, then the ratio of the rms value to the rectified average value equals 1.11. Using this ratio, a simple average value can be calculated and multiplied by 1.11 to generate an rms value. This is a simpler calculation than a true rms calculation. However, with a distorted input current, errors will be introduced.

The READ_POUT and READ_IOUT commands can



Take a look at our broad offering of quality power modules and find out why Bel is now the preferred source for dc-dc converters. Finally, you can get cost effective products in industry standard form factors without sacrificing performance. To learn more about how Bel can help you power your next system, visit us at www.belpower.com.

High Efficiency Power Modules

Isolated Converters - Single & Dual Output

- 1/16, 1/8, 1/4, 1/2 Bricks up to 120A

Bus Converters - 4:1, 5:1, 6:1 Fixed Ratios

- 1/16, 1/8, 1/4 Bricks up to 500W

VRMs - Solutions for most Microprocessors

- Up to 150A Output; Goldfinger and TH

Non-Isolated POL Modules - Boost, Buck and Inverting

- 1A to 150A Output; Vertical Mount or SMT



www.belpower.com • 1-800-BELFUSE

Command	Meaning
IOUT_OC_WARN_LIMIT	Output overcurrent warning limit
OT_WARN_LIMIT_1, _2, _3, _4	Overtemperature warning limit for temperature sensor
IIN_OC_WARN_LIMIT	Input overcurrent warning limit
POUT_OP_WARN_LIMIT	Output overpower warning limit
PIN_OP_WARN_LIMIT	Input overpower warning limit

Table 3. Some SMBus commands that can be set to assert SMBAlert#.

be used by the system in a few ways. The resulting data can be used to determine percent loading status of the system power supplies and motherboard converters. This will let users know whether they have power available to fit more memory or an additional adapter card.

Another use is to tell the power-management agent in the system how much power is going to different parts of the system. The power converters that power memory will tell users the total memory power consumption. The converters that power processors will tell users the processor power consumption. This allows performance/power optimizations by the system power-management agent.

Side-Band Interrupt

Since the typical server system is polling the power converters at such a slow rate, there needs to be a side-band interrupt method if a quick response is needed by the system to react to a condition in the power converter. The SMBAlert# signal is used for this.

The protocol details of the SMBAlert# signal are described in the SMBus specification.^[2] A power converter asserts SMBAlert# by pulling it low after one of the limit thresholds has been exceeded. The system responds by sending the Alert Response Address to all the power converters on that SMBus branch.

The power converter that asserted the SMBAlert# line responds with its address and releases the SMBAlert# line. **Table 3** lists some of the SMBus commands that can be set to assert SMBAlert#.

Thermal Management

PMBus has commands that are useful for server system thermal management. These include fan monitoring, fan control and temperature sensing. In some cases, fans inside power supplies often cool other components in the system. In order to optimize the power consumed by these fans, the system needs a means to control the fans based on the temperature of the system components they are cooling.

FAN_COMMAND_1 is one of four commands used to control up to four fans in a power supply. In other systems, fans in the system cool the power supply. In these cases, the system needs to know the power supply's temperature to allow for optimized power-supply cooling and lower system fan power. READ_TEMPERATURE_1 is one of three commands used to read temperatures inside

the power supply. In both sets of commands, PMBus can be used to help reduce power consumed by system fans or power-supply fans.

Diagnostics

PMBus also has fault-related commands to standardize the diagnosis of power-supply problems in the system. Power supplies are among the highest-failure-rate components in the data center. The more information available on these power supplies, the better a user can determine the root cause of a failure and verify that it is an actual power-supply problem. PMBus also has warning-related commands to head off power-supply problems before the system is affected by a fault.

The STATUS_WORD command is used by the system as the first level of diagnostic information containing a high-level summary of all fault and warning conditions. This command is the first to be used by the system to quickly read the status of the power supply. If more detailed information is needed by the system, the second level of status commands are used such as STATUS_INPUT and STATUS_TEMPERATURE. The sensors used in the power supplies affected by the various status commands are:

- *Input voltage.* This provides warnings if the input voltage drops too low and provides a fault if the input voltage drops lower than the minimum operating point.
- *Output voltage.* This provides a warning and a fault if the output voltage is sensed to be out of operating range.
- *Temperature.* This provides a warning if the temperature is about to exceed its maximum operating point and provides a fault if an excessive temperature causes the power supply to shut down.
- *Fan speed.* This provides a warning if a fan is wearing out and running too slow and provides a fault if the slow-running fan has caused the power supply to shut down.
- *Output current.* This provides a warning if the system loading is at the maximum rating of a power supply and provides a fault if the system has overloaded the power supply causing it to shut down.

New Features and Higher Accuracy

As servers add new power-management capabilities, PMBus will play an important role in these new capabilities. The monitoring of real power consumption at the rack, system and subsystem levels is beginning to be used to help make data centers more efficient, increase data center densities and calculate system energy consumption for charge back. PMBus will help make these new features standard capabilities of all server power supplies and allow for improvements in capabilities such as accuracy. **PETech**

References

1. Advance Configuration and Power Interface specification, www.acpi.info.
2. System management bus specification, <http://smbus.org/specs>.