INTELLIGENCE UPDATE

Managing server performance for power: a missed opportunity


Daniel Bizo

9 Feb 2024

A recent Uptime Intelligence report discussed the characteristics of processor power management (known as C-states) and explained how they can reduce server energy consumption to make substantial contributions to the overall energy performance and sustainability of data center infrastructure (see *Understanding how server power management works*). During periods of low activity, such features can potentially lower the server power requirements by more than 20% in return for prolonging the time it takes to respond to requests.

But there is more to managing server power than just conserving energy when the machine is not busy — setting processor performance levels that are appropriate for the application is another way to optimize energy performance. This is the crux of the issue: there is often a mismatch between the performance delivered and the performance required for a good quality of service (QoS).

When the performance is too low, the consequences are often clear: employees lose productivity, customers leave. But when application performance exceeds needs, the cost remains hidden: excessive power use.

Server power management: enter P-states

Uptime Intelligence survey data indicates that power management remains an underused feature — most servers do not have it enabled (see *Tools to watch and improve power use by IT are underused*). The extra power use may appear small at first, amounting to only tens of watts per server. But when scaled to larger facilities or to the global data center footprint, they add up to a huge waste of power and money.

The potential to improve the energy performance of data center infrastructure is material, but the variables involved in adopting server power management mean it is not a trivial task. Modern chip design is what creates this potential. All server processors in operation today are equipped with mechanisms to change their clock frequency and supply voltage in preordained pairs of steps (called dynamic voltage-frequency scaling). Initially, these techniques were

devised to lower energy use in laptops and other low-power systems when running code that does not fully utilize resources. Known as P-states, these are in addition to C-states (low-power modes during idle time).

Later, mechanisms were added to do the opposite: increase clock speeds and voltages beyond nominal rates as long as the processor stays within hard limits for power, temperature and frequency. The effect of this approach, known as turbo mode, has gradually become more pronounced with ever-higher core counts, particularly in servers (see *Cooling to play a more active role in IT performance and efficiency*). As processors dynamically reallocate the power budget from lowly utilized or idling cores to highly stressed ones, clock speeds can well exceed nominal ratings — often close to doubling. In recent CPUs, even the power budget can be calibrated higher than the factory default.

As a result, server-processor behavior has become increasingly opportunistic in the past decade. When allowed, processors will dynamically seek out the electrical configuration that yields maximum performance if the software (signaled by the operating system, detected by the hardware mechanisms, or both) requests it. Such behavior is generally great for performance, particularly in a highly mixed application environment where certain software benefits from running on many cores in parallel while others prefer fewer but faster ones.

The unquantified costs of high performance

Ensuring top server performance comes at the cost of using more energy. For performance-critical applications such as technical computing, financial transactions, high-speed analytics and real-time operating systems, the use and cost of energy is often not a concern.

But for a large array of workloads, this will result in a considerable amount of energy waste. There are two main components to this waste. First, the energy consumption curve for semiconductors gets steeper the closer the chip pushes to the top of its performance envelope because both dynamic (switching) and static (leakage) power increase exponentially. All the while, the performance gains diminish because the rest of the system, including the memory, storage and network subsystems, will be unable to keep up with the processor's race pace. This increases the amount of time that the processor needs to wait for data or instructions.

Second, energy waste originates from a mismatch between performance and QoS. Select applications and systems, such as transaction processing and storage servers, tend to have defined QoS policies for performance (e.g., responding to 99% of queries within a second). QoS is typically about setting a floor below which performance should not drop — it is rarely about ensuring systems do not overperform, for example, by processing transactions or responding to queries unnecessarily fast.

If a second for a database query is still within tolerance, there is, by definition, limited value to having a response under one-tenth of a second just because the server can process a query that fast when the load is light. And yet, it happens all the time. For many, if not most, workloads, however, this level of overperformance is neither defined nor tracked, which invites an

exploration of acceptable QoS.

Governing P-states for energy efficiency

At its core, the governance of P-states is like managing idle power through C-states, except with many more options, which adds complexity through choice. This report does not discuss the number of P-states because this would be highly dependent on the processor used. Similarly to C-states, a higher number denotes a higher potential energy saving; for example, P2 consumes less power than P1. P0 is the highest-performance state a processor can select.

- **No P-state control**. This option tends to result in aggressive processor behavior that pushes for the maximum speeds available (electronically and thermally) for any and all of its cores. While this will result in the most energy consumed, it is preferable for high-performance applications, particularly latency-sensitive applications where every microsecond counts. If this level of performance is not justified by the workload, it can be an exceedingly wasteful control mode.
- **Hardware control.** Also called the autonomous mode, this leaves P-state management for the processor to decide based on detected activity. While this mode allows for very fast transitions between states, it lacks the runtime information gathered by the operating system; hence, it will likely result in only marginal energy savings. On the other hand, this approach is agnostic of the operating system or hypervisor. The expected savings compared with no P-state control are up to around 10%, depending on the load and server configuration.
- **Software-hardware cooperation.** In this mode, the operating system provider gives the processor hints on selecting the appropriate P-states. The theory is that this enables the processor control logic to make better decisions than pure hardware control while retaining the benefit of fast transitions between states to maintain system responsiveness. Power consumption reductions here can be as high as 15% to 20% at low to moderate utilization.
- **Software control**. In this mode, the operating system uses a governor (a control mechanism that regulates a function, in this case speed) to make the performance decisions executed by the processor if the electrical and thermal conditions (supply voltage and current, clock frequency and silicon temperature) allow it. This mode typically carries the biggest energy-saving potential when a sophisticated software governor is used. Both Windows and Linux operating systems offer predefined plans that let the system administrator prioritize performance, balance or lower energy use.
  The trade-off here is additional latency: whenever the processor is in a low-performance state and transitions to a higher-performance state (e.g., P0) in response to a bout of compute or interrupt activity, it takes material time. Highly latency-sensitive and bursty workloads may see substantial impact.
  Power reductions can be outsized across most of the system load curve. Depending on the sophistication of the operating system governor and the selected power plan, energy savings can reach between 25% and 50%.

While there are inevitable trade-offs between performance and efficiency, in all the control scenarios, the impact on performance is often negligible. This is true for users of business and web applications, or for the total runtime of technical computing jobs. High-performance requirements alone do not preempt the use of P-state control: once the processor selects P0,

there is no difference between a system with controls and no controls.

Applications that do not tolerate dynamic P-state controls well are often the already-suspected exceptions. This is due to their latency-sensitive, bursty nature, where the processor is unable to match the change in performance needs (scaling voltages and frequency) fast enough, even though it takes microseconds.

Arguably, for most use cases, the main concern should be power consumption, not performance. Server efficiency benchmarking data, such as that published by the Standard Performance Evaluation Corporation and The Green Grid, indicates that modern servers achieve the best energy efficiency when their performance envelope is limited (e.g., to P2) because it prevents the chip from aggressively seeking the highest clock rates across many cores. This would result in disproportionately higher power use for little in return.

Upcoming Uptime Intelligence reports will identify the software tools that data center operators can use to monitor and manage the power and performance settings of server fleets.

The Uptime Intelligence View

Server power management in its multiple forms offers data center operators easy wins in IT efficiency and opportunities to lower operational expenditure. It will be especially attractive to small- and medium-sized enterprises that run mixed workloads with low criticality, yet the effects of server power management will be much more impressive when implemented at scale and for the right workloads.

ABOUT THE AUTHOR

Daniel Bizo

Daniel serves as Research Director at Uptime Institute. Over the past 15 years, he has covered the business and technology of enterprise IT and infrastructure in various roles, the past ten years as an industry analyst and advisor. His research includes sustainability, operations, and energy efficiency within the data center, on topics like emerging battery technologies, thermal operation guidelines, and processor chip technology.

dbizo@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.