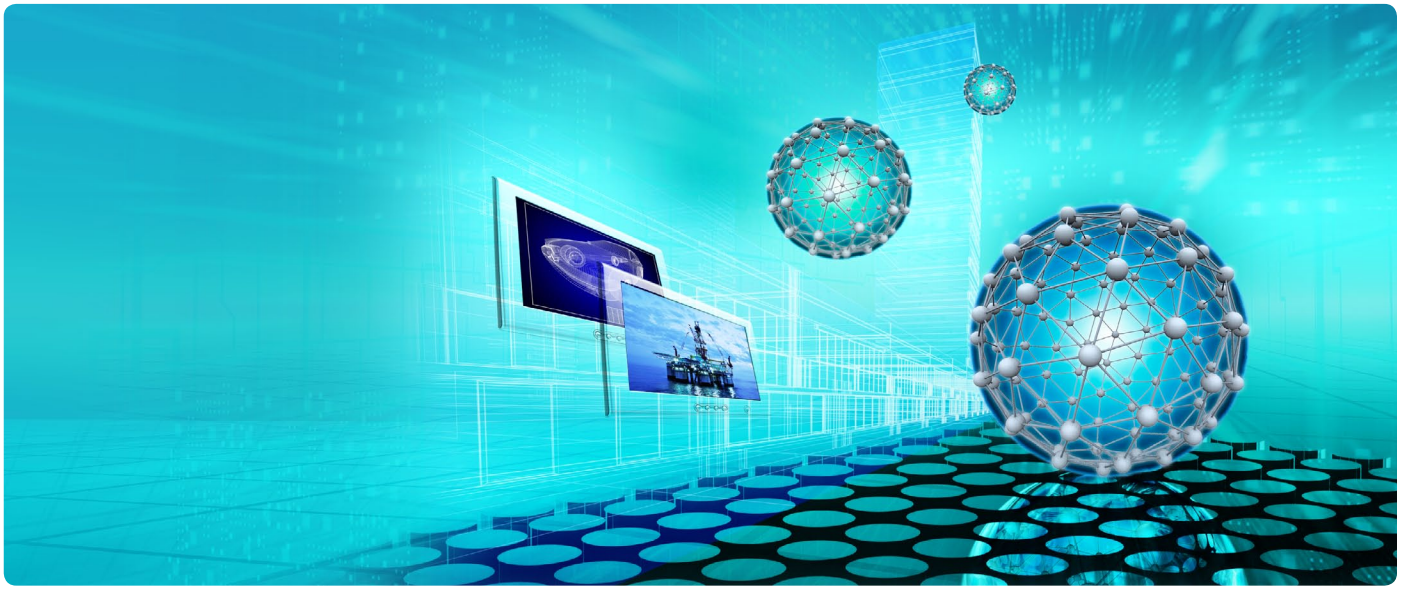
WHITEPAPER



The Growing Demand for Intelligent Power Management Control

Adaptive COMPUTING

Table of contents

3 Introduction
3 No More Reckless Abandon
4 Clock Frequency Control
5 Shut It Down (or Something Like That)
6 Get Green While You're At It
6 Other Parameters
7 An Exciting Future

Introduction

Remember the good old days when the only concern with HPC was making systems bigger and discovering ways to achieve peak performance? Unfortunately, the fruits of those exciting times created a challenge: how do we afford to power these ever-growing massive machines, limit their carbon footprint, stay within imposed power caps, comply with government regulations, and maintain performance?

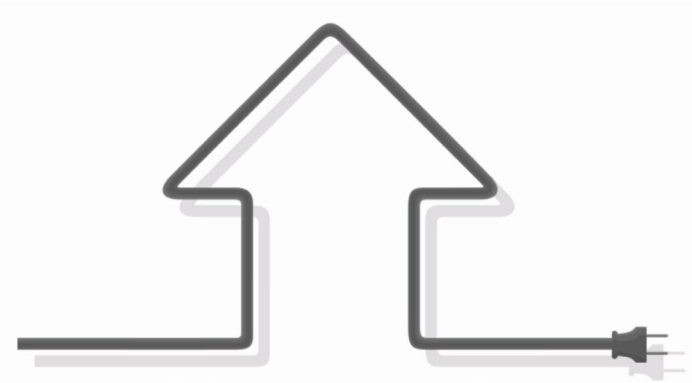
One way is through intelligent power management control. Certainly, the demand for power management (monitoring and control) solutions is growing. It is estimated that U.S. data-center power consumption doubles every five years (2000–2005 period [1], although later studies[2] show slower growth (36 percent in U.S. and 56 percent worldwide) in the 2005–2010 period due to the slower economy and technology such as more efficient computer chips and server virtualization; however, the growth rate has been increasing (7 percent growth in 2013). The march toward exascale computing—which conceivably could reach power consumption of 10 GW[3] (think twice the power needs of New York City)—isn't helping relieve anxiety for how HPC data centers can afford to keep future systems cost-efficient and environmentally friendly.

Unpredictable energy prices don't do much to calm nerves either. Globally, data centers today use about 30 billion watts of electricity annually, comparable to the output of 30 nuclear power plants. In the United States, the average cost of powering an HPC system rated at 1 MW is roughly $1 million per megawatt-year[4] when including power distribution and cooling costs. In the United Kingdom, the costs are about twice[5] that of prices in the United States, while in Germany[6] prices are about 2.5 times U.S. prices. And who knows what future pricing will look like?

"Between 2006 and 2013, the portion of HPC budgets devoted to power and cooling held steady between eight and nine percent," says Steve Conway, Research Vice President, HPC at International Data Corporation (IDC). "Yet as systems expand and it is uncertain when deeper, more integral energy-efficient capabilities will become available, sophisticated power management solutions will be, if they aren't already, in high demand."

So the foreseeable exascale HPC future (still exhilarating to anticipate isn't it?) foreshadows that solutions need to be implemented now. Indeed, the value of developing power management solutions has attracted the attention of data center administrators and operators who are feeling the pressure primarily on the bottom line as they expand. According to worldwide surveys by IDC, HPC data centers rank power and cooling as their No. 2 concern. The No. 1 concern? Surprise, surprise—the need for bigger budgets.

As energy costs have risen to levels approaching capital costs, a gradual shift has occurred away from peak performance and toward power management. Data centers are betting on new cooling technologies and power management software solutions, such as those that are being developed at Adaptive Computing for its Moab HPC Suite, to help ease power consumption burdens.



No More Reckless Abandon

The concern about power consumption is relatively new for data centers. They had the budgets, and they had these massive, amazing systems—they were living computing nirvana. So they giddily sped along until governments, environmental groups, and bottom-line managers began to take notice.

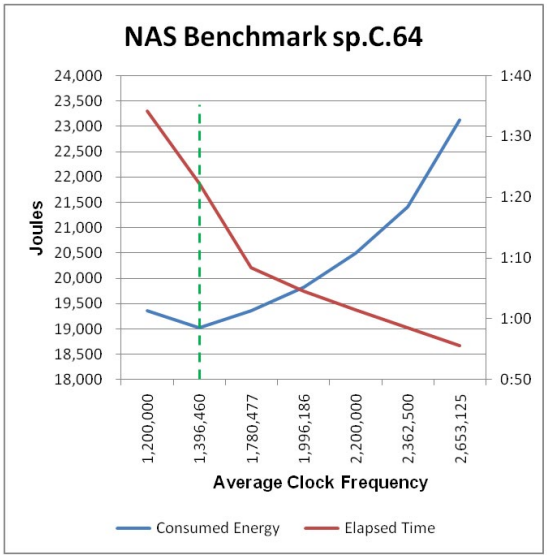
Data centers historically waste 90 percent or more of the electricity they pull off the grid by running their facilities at maximum capacity around the clock, regardless of whether the system is being fully utilized or not. But the days of reckless waste are quickly coming to an end as the realities of cost and government constraints have demanded forward thinking around better utilization of energy.

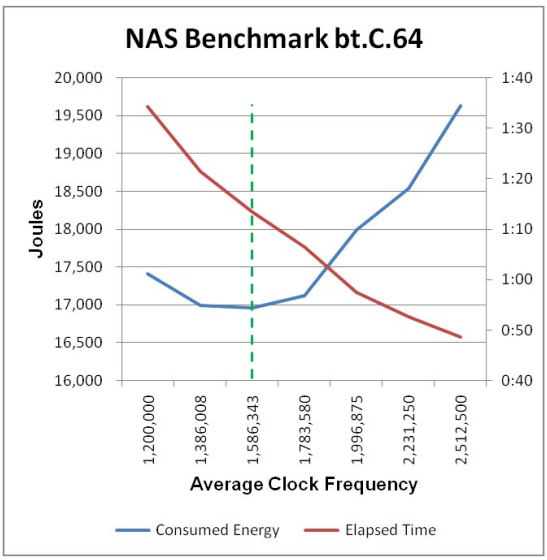
Fortunately, strategic methods are being used or are on the horizon to drastically reduce energy waste through innovative software solutions that can reduce energy usage when the system is running at full capacity or when clusters of compute nodes sit unused. This paper discusses some of the most immediate solutions to drive down HPC power costs.

For jobs Moab decides to start, it then needs to decide where to place them.

Clock Frequency Control

For some time, processors have been able to increase clock frequency to run faster as integrated circuits have become smaller. A faster clock boosts performance, but unfortunately also increases power levels. In power management, ideally you would adjust systems so that power consumption is proportionate to its workload. In essence, the server would consume no power when it is idle, little power when the workload is low, and more power when the workload is increased.





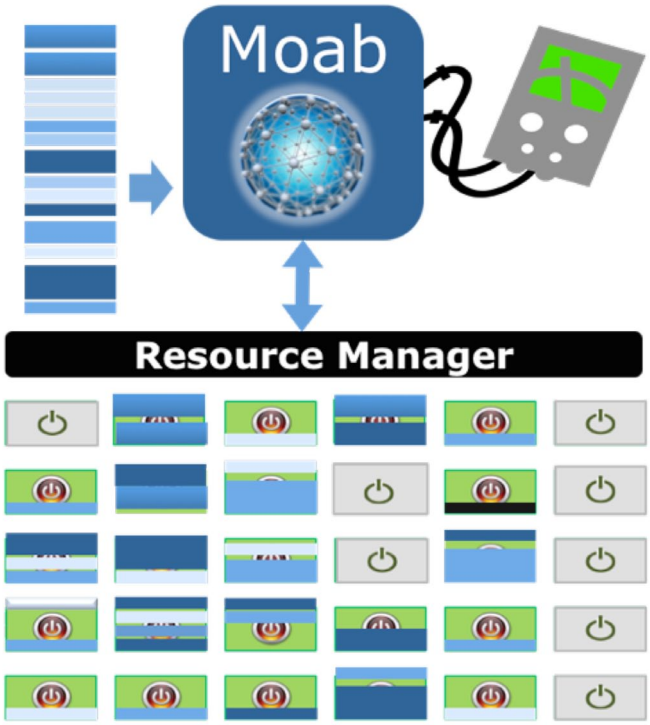
So turning off the node or slowing the clock frequency whenever excess CPU time is available is one sure way to lower energy consumption. Of course, there are power-performance tradeoffs, but certain jobs, such as memory-bound workloads, make this strategy potentially advantageous.

It's no secret that waiting on memory subsystems can account for a significant power drain. Without having to reference memory, a processor could perform several hundred computations in the same time it spent waiting for memory and greedily gobbling electricity. Why run the CPU at a high clock frequency if it just sits consuming valuable power waiting on memory latency? Better to lower the clock frequency and save power, right? Well, it depends.

Some jobs involve computations that require constant memory accesses and therefore incur a lot of memory latency (i.e., they are memory-bound). These jobs can execute in the same time with reduced power consumption if you cut the clock frequency. Other jobs with intense computations and few memory accesses (compute-bound) can, if you reduce the clock frequency too extensively, increase their execution time so much that the job consumes more energy than it would have originally. Still, if you manage it right, you can adjust the clock frequency to account for memory and/or I/O latency, making a difference in power consumption by getting the performance you expect while saving energy.

So, a balance must be achieved. The challenge is finding that sweet spot where processor power consumption and memory workloads are creating an equitable balance that save energy and maintain performance.

To reach that equilibrium, you have to experiment. The ideal method is to run a series of workloads—say 10 jobs—preferably with the same data or a similar workload at different clock frequencies and measure power consumption over time. Then when you run similar workloads, you know the optimal clock frequency that will offer the best bang for your buck. When the job is done, your policy will determine that the nodes be set back to the default clock frequency to accommodate other jobs.



Shut It Down (or Something Like That)

Data centers, particularly the smaller ones, have times when clusters of compute nodes sit idle drawing energy. In fact, idle nodes consume 30–70 percent as much energy as when they are running an application. A solution, of course, is to turn off those nodes and concentrate a workload in fewer nodes. When data center demand increases, additional nodes can be reactivated.

But what kind of "off" do you want? Of course, that depends on how quickly a data center needs nodes to be returned to their active state. One option you have is to completely shut down idle compute nodes. However, switching off all or some compute nodes is often not ideal. For one, powering on and off hosts can produce workload latencies because of the boot time required—which can take up to 45 minutes—causing users to become extremely unsatisfied. Another drawback is that often a percentage of compute nodes need manual interaction when restarted (not making data center operators happy).

A more viable option is to place the compute nodes in low-power suspend or sleep states or even to hibernate a node. When nodes are in the low-power suspend or sleep states they consume 10–50 percent of power compared to the active running state. So while not as great a savings as a complete shutdown, suspend or sleep will offer significant energy savings without the severe latency consequences of a complete shutdown.

In hibernation, the compute node saves the contents of its random access memory to a hard disk or other storage and powers off, saving as much energy as a shutdown. Once reactivated, the computer returns to the same state as it was in before hibernation. On boot, the boot loader detects the presence of the hibernated state file and restores system state, which occurs much more quickly than a cold reboot after a shutdown. Hibernation allows you to avoid the burden of saving unsaved data before shutting down and restarting all running programs after powering back on.

If you or your users are the impatient type (when it comes to performance we all are, aren't we?) then the suspend (Linux) or sleep (Windows) modes are a better alternative than hibernate. These modes offer the advantage of returning to an active running state much quicker than from hibernation. A hibernated system must reboot, then read back data to RAM from disk on resuming, which typically takes much more time than recovering from suspend or sleep, but often much less time than a reboot. A system in suspend or sleep mode only needs to power up the CPU, perhaps some devices, and if present, the display, which depending on the server can range from almost instantaneous to many seconds.

Get Green While You're At It

The energy-saving methods mentioned above can be manually invoked, but their real power lies in their automation via policies.

CPU clock frequency control can be automatically invoked through Moab job templates. When a job template matches a submitted job and the template specifies a clock frequency control option, the job takes on the option specified by the template, regardless what the user has specified at job submission time or in the job script. This allows the administrator to control clock frequency for jobs using specific applications that have an experimentally determined "best" frequency at which the site wants such jobs to execute.

Setting policies through the Moab HPC Suites that implement one of these strategies is a great way to save energy and get some green computing kudos along the way. Green computing, in terms of power management, is looking for ways to responsibly use and limit power consumption automatically, which this strategy does.

For example, let's say you have a green pool of 20 idle nodes. Another 10 nodes have just completed a job and now you have 30 idle nodes. Your green policy states that you can only have 20 idle nodes. If that number exceeds 20 then the leftover nodes are placed in sleep, hibernation or whatever state you want, reducing the number of fully powered idle nodes, thus saving power. While during this time the amount saved may seem minimal, the cumulative effect over a year's time will astound even the most jaded energy accountants.

Some data centers may recognize that clusters are more active at specific times of the day, or that energy costs are variable, depending on the time of day. As such, they may set policies that suspend a set amount of cluster nodes at different times of the day to take advantage of these cost-saving variables.

Another factor to consider if you decide to lower energy consumption by adjusting clock frequency is the capabilities of your nodes. For example, if you have older nodes on which you are unable to adjust clock frequency, you will have to run memory- or I/O-bound jobs on newer nodes on which you can adjust clock frequency to consume less energy. But what do you do with the older nodes? If there aren't other jobs to run, you can choose to put them in a suspended or sleep state or turn them off completely. But if you have jobs that need to run at full speed, such as CPU-bound jobs, you can run them on the older nodes while the newer nodes are consuming less energy on the other jobs.

Other Parameters

Not all systems run just one interface. For example, some run both HP and IBM systems, which have different commands to manage their power sources. So, Moab offers the ability to set parameters to account for these different commands. Other parameters Moab can manage include fairshare scheduling such as within a condo HPC service. In a condo service, a principal (such as the university) buys compute nodes that can be placed into a pool or common HPC resource using funds supplied by various university departments or researchers out of their budgets. For example, the university might have 200 nodes and four departments may each have paid for 50 nodes. They will each expect to get the use of 50 nodes during the year, but can get use of more nodes if the other departments are not using their nodes.

An Exciting Future

The demand for power management will continue to grow, if not simply because energy constraints and rising costs will drive it. Already, forward-thinking data centers are calling for such things as power-aware scheduling and automatic bi-directional communication with their power company. And new chips are being developed with per-core clock frequency control—these are just a few exciting developments coming soon.

We are moving toward a remarkable period as we move closer toward exascale, developing and adopting innovative solutions to help data centers maintain the pace toward growth to meet the high demand of big data, reach peak performance, and maintain energy consumption within manageable levels. Such innovations should keep government agencies, environmentalists, power companies, and bean counters content and satisfied—so much so that even they will sit back, relax, and marvel at the future of HPC.

[1] US EPA. 2007. Report to Congress on Server and Data Center Energy Efficiency, Public Law 109-431. Prepared for the U.S. Environmental Protection Agency, ENERGYSTAR Program, by Lawrence Berkeley National Laboratory. LBNL-363E. August 2. http://www.energystar.gov/datacenters.

[2] Jonathan Koomey. 2011. Growth in Data center electricity use 2005 to 2010. Oakland, CA: Analytics Press. August 1. http://www.analyticspress.com/datacenters.html

[3] DCD Intelligence, DCD Industry Census 2013: Data Center Power, 31 January 2014. http://www.datacenterdynamics.com/focus/archive/2014/01/dcd-industry-census-2013-data-center-power

[4] U.S. Energy Information Administration, Electric Power Monthly, Table 5.3. Average Retail Price of Electricity to Ultimate Customers (Industrial column), March 2014 data released May 21, 2014. http://www.eia.gov/electricity/monthly/epm_table_grapher.cfm?t=epmt_5_03
Note industrial electricity price of $0.07 per kWhr yields $613.62 per kWyr (year = 365.25 with days × 24 hours), which with power distribution and cooling costs yields a "rule-of-thumb" estimate of ~$1,000 per kWyr or $1M per MWyr for an HPC cluster rated at 1 MW.

[5] GOV.UK, International industrial energy prices, Quarterly: Industrial electricity prices in the EU for small, medium, large, and extra large consumers (QEP 5.4.1, 5.4.2, 5.4.3 and 5.4.4), last updated 27 March 2014. Excel spreadsheet https://www.gov.uk/government/statistical-data-sets/international-industrial-energy-prices Using medium customer, including tax, worksheet (~2MW), July-Dec 2013 column, and UK GBP-to-US Dollar exchange rate of 1 GBP = $1.57 yielded an electricity rate of $0.1471 per kWhr or $1,289.48 per kWyr, with rule-of-thumb estimate of ~$2,101 per kWyr or $2.1M per MWyr for an HPC cluster rated at 1 MW.

[6] Same as 5, but used the Germany information in the same spreadsheet. Germany electricity rate of $0.1694 per kWhr yields $1,484.96 per MWyr, which with power distribution and cooling costs yields a "rule-of-thumb" estimate of ~$2,420 per kWyr or $2.42M per MWyr for an HPC cluster rated at 1 MW

Let's talk...Set up a Demonstration...and Test in your Environment

An Adaptive Computing solutions advisor can guide you to the products and services that will best meet your needs and will work with you to set up a live, online demonstration designed specifically for your organization.

Contact a solutions advisor by phone or email, or visit our Web site today

North America, Latin America +1 (801) 717.3700
Europe, Middle East, Africa +44 (0) 1483 243578
Asia, Pacific, Japan, India +65 6597-7053
Email: solutions@adaptivecomputing.com
www.adaptivecomputing.com

Corporate Headquarters
1712 S. East Bay Blvd.
Suite 300
Provo, Utah 84606

Adaptive COMPUTING