Industry Insights: What You Should Know about Power and Performance Efficiency

Our latest Q&A session takes a look at an area of HPC that is weighing on many people's minds as datacenter environments face increasing pressure to expand capabilities while controlling costs. We've gathered industry experts to share their unique perspectives and to take a look at what we can expect to see coming up on the horizon in the next few years. Vice presidents, chief technology officers, computer scientists and engineers have come together to provide a range of high-level viewpoints that will, hopefully, get you started with your own specific challenges.

— Suzanne Tracy, Editor in Chief

Q1: What are the major components impacting power consumption for today's HPC data centers? Does any one stand out above the others?

Processor design seems to be the one area with the most design flexibility.

– Michael Wolfe
The Portland Group

Michael Wolfe: Processors, memory, disk, cooling. Processor design seems to be the one area with the most design flexibility related to power.

Andrew Jones: Components driving power consumption fall into two categories — those that, as consumers, we cannot control, and those we can. Power consumed by server hardware is increasing and is beyond our direct control as buyers (although manufacturers are working to optimize power efficiency). The biggest factors we can influence are design and deployment of HPC systems as a whole (datacenter included) and recognizing total cost of ownership (including power) when procuring.

Steve Kinney: The business requirement for additional compute capacity is having a big impact on power consumption for HPC data centers. With the higher capacity, comes the requirement for more cooling. System providers are placing more and more compute systems in each cabinet, which may require a redesign of the datacenter to support the higher capacity.

Ed Turkel: Apart from the IT equipment, the major power consumers are cooling systems and power distribution systems (losses). Power consumption for the cooling systems still leads power system losses. Power systems are making significant gains in reducing distribution losses (more efficient power supplies, more efficient UPS, possibility of high voltage AC power delivery to the rack, etcetera). For the IT equipment, the compute equipment still leads in absolute power consumption, but significant gains are being made in performance/watt. Storage and network equipment are increasing in absolute power consumption, but do not seem to be making the same gains in performance/watt that the compute equipment is making.

Blake Gonzales: Power in HPC environments is generally consumed by the same components and infrastructure that you would typically find in enterprise environments, with one very large caveat: You have to scale up!

► The first classification of major power consumers in HPC environments is well-known system components, such as processors, memory and, to some degree, disk storage and interconnect hardware. The ratio of compute nodes to other components is usually high. So, you will likely find that most of your power is consumed by your computational resources. Take a walk around any HPC environment and find the components that are generating heat; these hot spots will be your major power consumers.

▶ This leads us to a second classification of major power consumers: cooling infrastructure. For every watt consumed in a component, there is a cooling power component that must be consumed as well. Powered cooling infrastructure includes items such as chillers, air handlers and fans. When you add a few more servers in an enterprise environment, it is easy to overlook the power consumed by the incremental cooling needed. This is not so with HPC. Careful planning for HPC power and cooling resources is as critical as the design of the computational components.

Steve Scott: There are several factors that contribute:

▶ First is the power usage effectiveness (PUE) which is the ratio of facility power to the power that actually gets delivered to the compute cabinets. Improvements in delivery and cooling infrastructure are moving this from historical levels of close to 2, down to 1.4, 1.2 or lower, leaving only modest additional gains to be had.

▶ Second is the power conversion inside the box, where high voltage AC is stepped down to low voltage DC that the chips can use. This typically causes inefficiencies (power losses) in the tens of percent, and could potentially be reduced to less than 20 percent with additional work.

▶ Lastly, and most importantly, there is the power consumed by the processors, memory and interconnect. The vast majority of this power goes into control overhead and data movement, rather than actual computation in the ALUs of the processors. Additionally, many HPC applications sustain small fractions of peak performance. A system with a PUE of 1.2 has only 20 percent upside by improving the cooling and power delivery. An application sustaining five percent of peak performance has an upside of 20x by increasing its efficiency.

Every watt required to power a server nominally requires another watt to cool it.

— Bob Masson
Convey Computer

Bob Masson: In the continuing effort to get more performance into a smaller space, server and storage manufacturers are increasing component densities to the limits of cooling technology. For example, multiple vendors (e.g. Dell, HP, IBM) have compact blades that contain twin, dual-socket server motherboards in a small (~310 cubic inch) enclosure. However, because of the heat they generate, commodity processors have effectively reached their limit of performance (at least that can be gained by increasing the clock rate). To add performance, IT managers have no choice but to add more and more racks of off-the-shelf servers to the datacenter.

As a result, the major components that impact power consumption are the rows and rows of racks that are being added to datacenters in an attempt to get the performance HPC users require to run their applications. To illustrate: a modern datacenter today can cost up to $200 million to build, with power and cooling costs making up half of the annual operating costs. Every watt required to power a server nominally requires another watt to cool it.

Q: Q2: *What do you see as the key areas in developing an effective power strategy that allows for maintaining or increasing performance?*

Michael Wolfe: All power-efficient designs trade lower single-thread performance for higher multi-thread efficiency. We need to effectively manage even higher levels of parallelism than we might have predicted, because we're going to need additional parallelism just to maintain parity with fast single cores. We're also going to need novel architectures that can effectively balance memory bandwidth with multicore performance.

The primary strategy for optimizing power is to ensure proper total cost of ownership as the driver of procurement.

— Andrew Jones
NAG

Andrew Jones: The primary strategy for optimizing power is to ensure proper total cost of ownership (including power) as the driver of procurement, not purely peak performance and initial capital cost. This enables the evolutions of datacenter optimization (e.g. run warm, "free-cooling," hot aisles) and choices of power-efficient HPC system designs (e.g. more parallelism, lower power processors, etcetera) to be correctly attributed as delivering increased performance against cost.

Steve Kinney: A redesign of certain areas of the data center to allow higher floor ratings for power and cooling is an important part of an effective strategy. In HPC data centers, the floor rating of 50 to 80 watts per square foot will not support today's 15 kWatt and up cabinets. Compute systems should start having a performance rating of FLOPS per watt to give IT teams the ability to review total cost of ownership of certain brands.

Ed Turkel: Clearly, the deployment of systems with higher-efficiency components is key, including very high-efficiency power supplies, low-power processors and memories; and SSDs, instead of rotating disks, will drive strategies for high-efficiency systems. At the datacenter level, the introduction of monitoring and management systems that allow management of the IT in datacenters versus the power and cooling systems in those datacenters allows for better management of PUE. Finally, applying cooling systems from the datacenter level to the rack, or collection of racks, down to more aggressive cooling strategies with the racks, will lead to greater efficiencies.

You may find that replacing your infrastructure will save dollars in the long run.

— Blake Gonzales
Dell High Performance Computing

Blake Gonzales: HPC hardware power efficiencies continue to improve as time passes. For instance, Intel and AMD are delivering faster processors with more cores, while reducing the required power envelopes. For organizations with hardware three years old or beyond, it is especially prudent to analyze your power bill. Compare your current HPC power draw to that of a new replacement system with similar performance characteristics. You may find that replacing your infrastructure will save dollars in the long run. Of course, results may vary depending upon in what part of the country your HPC system makes your power meter spin. For many organizations, the power bill over the life of an HPC system exceeds the cost of the HPC system itself!

Sometimes, quantifying power draw and costs is not a very simple exercise. If you don't have a good handle on how much power you are consuming, it makes it very hard to make economical decisions about power. The first step to an effective power strategy is quantifying your usage. Do whatever it takes to meter and monitor your power usage. Only then will you be able to be smart about making your HPC system green.

Steve Scott: System designers must attack all areas of power inefficiency simultaneously. However, the biggest gains will come in developing new, more efficient microarchitectures and execution models. Typical multicore CPU processor cores are designed to run single threads fast, not run threads power-efficiently.

We need to transition to heterogeneous microarchitectures, where a small number of latency-optimized cores are coupled to a large number of power-optimized cores. Most of the heavy lifting must be done by the power-efficient cores, which will spend relatively more of their transistor and power budget on computation, and less on orchestration and latency reduction.

Processors will need to aggressively exploit locality to reduce data movement, which is actually much more energy-intensive than computation. The programming system (compilers, runtime and libraries) will have to manage this exposed heterogeneity and memory system hierarchy efficiently while shielding the user from the underlying complexity — a tall order!

Bob Masson: The key area of development for increasing performance is to drastically increase the efficiency of the hardware used in computing. In other words, given a fixed number of transistors, the key is to make those transistors hundreds or thousands of times more efficient (in terms of performance/watt) for a given HPC application.

The only way to substantially increase efficiency is to employ hardware (e.g. field programmable gate arrays, or FPGAs) specifically tailored to an individual application. If a given number of transistors can be arranged to compute a specific algorithm, they'll be much more efficient. (In this case, "efficient" means it takes less cycles and/or less gates to implement than a general-purpose counterpart). In other words, the use of heterogeneous computing (employing different types of processors for different application kernels) reduces the amount of logic needed to solve a specific problem.

Q3: How critical is optimizing HPC applications and their algorithms for low power to improving the performance efficiency of HPC systems?

Michael Wolfe: Explicitly optimizing for low power is the wrong paradigm; applications don't control the power envelope, and decisions, such as what hardware components to power off are better made at the system level. Instead, we should optimize applications for efficient and effective use of the available resources. The challenge is expressing this in a way that delivers performance across the wide range of potential architectures

PARTICIPANTS

Bob Masson
Marketing
Convey Computer

Steve Scott
Senior VP and CTO
Cray

Blake Gonzales
HPC Computer Scientist
Dell HPC

Ed Turkel
Manager, Business Development
Scalable Computing & Infrastructure
Hewlett-Packard

Steve Kinney
Global Solution Manager
Computing on Demand
IBM

Andrew Jones
VP HPC Consulting
NAG

Michael Wolfe
Compiler Engineer
The Portland Group

that will arise in the coming decade, such as using low-power processors, novel accelerators and wide vector instruction sets.

Andrew Jones: Optimizing software and algorithms is a key opportunity to dramatically improve the total cost of ownership of HPC solutions. By optimizing applications, fewer resources are required to deliver the results, thus reducing the power required. Equally, innovations in algorithms can deliver applications that are power-aware — that is, they recognize the energy consumed and the user can balance energy-cost against time-to-solution when selecting algorithms for a given simulation.

Steve Kinney: Some applications in use today were developed in a time when space/power/cooling and, in many cases, overall performance of the application were not an issue. The overall "rating" of each application should be evaluated to see where design improvements can be made and then weighed against the investment of redesign and expected return on investment.

Power efficiency is one of the rationales for the increasing focus on application accelerators, such as GPGPUs.

— Ed Turkel
HP

Ed Turkel: An area of investigation is linking the monitoring information on power consumption and thermals within the datacenter to workload management of HPC applications. This allows for strategies to distribute work in a way that maximizes energy efficiency across the datacenter. Power efficiency is also one of the rationales for the increasing focus on application accelerators such as GPGPUs, where the presumption is that at least some portions of applications that lend themselves to particular accelerator technologies will run at higher performance/watt versus on mainstream processors. This heterogeneous processing model might allow for finer tuning of power consumption versus performance.

Blake Gonzales: Let me start out by asking two questions:

1. Have you ever seen a "power threshold" mentioned in the requirement specification for a particular algorithm?
2. When was the last time you noticed "wattage" listed in the requirements to run an application?

I suspect the answer for most will be never. Applications and algorithms are not thought of like light bulbs requiring a minimum power draw. That isn't to say they should not be viewed this way; my point here is to illustrate the wide gap between application development and energy usage. In the HPC world, a small energy optimization on one thread can lead to a large power savings if we multiply that optimization across thousands of threads across a cluster. Since the operating system and the hardware are more tightly coupled, I think we will start to see more progress in the Linux kernel before we see major enhancements in application code.

Users should ... write their code such that parallelism is exposed to the system software at as high a level as possible.

— Steve Scott
Cray

Steve Scott: The challenge for HPC application developers will be to map their applications onto the new generation of heterogeneous processors with exposed memory hierarchies. This must be done in a way that is portable across machine types and forward in time. We should not ask users to write their codes to map onto a particular machine. Rather, they should write their code such that parallelism is exposed to the system software at as high a level as possible. Within each node of a distributed application (e.g.: one using MPI), the code should be written so that loops at the highest level in the call tree can be safely performed in parallel. Programmers should also think about locality, meaning that computations can work on data "near" them, and re-use data multiple times once it is referenced from memory. Sophisticated system software can then map the problem onto the low-level hardware in a way to exploit maximal parallelism and reduce data movement.

Bob Masson: It isn't critical to optimize HPC applications and algorithms — it's mandatory! There is no other way to substantially increase performance/watt than to employ application-specific hardware to each problem. Sure, processor vendors (i.e. Intel and AMD) are developing multiple technologies to reduce energy consumption. But these are incremental changes, and don't address the fundamental issue — that is, for a given application, a given sequence of general-purpose instructions must be executed.

Why not tackle the problem from a different viewpoint? Instead of executing a series of general-purpose instructions, why not translate the algorithm directly into hardware? Almost by definition, gates arranged to perform a specific function will be more efficient than a general-purpose solution. For example, the SSE (streaming SIMD extensions) instructions in x86_64 instruction set were added by Intel and AMD specifically to carry out long sequences of floating-point operations on matrices to increase computing efficiency.

Q4: Do you expect any "green" technology breakthroughs in the next 3 to 5 years? What are they?

Michael Wolfe: We can predict the benefits from transistor feature scaling and some architectural innovations, given past history; it will be interesting to see what effect new packaging technologies will have, like 3-D stacking. On the other hand, HPC has adopted commodity parts (processors in particular) as a cost-effective alternative to custom design; when the world of innovation is driven by embedded computing (cell phones and smaller), will we be able to build systems from those parts?

Andrew Jones: The primary breakthrough will be the recognition of the role software (both implementation efficiency and algorithm design) has to play in delivering cost savings related to power efficiency. Beyond that, the key hardware technologies will be increased use of power switching across the system — while many modern processors will reduce power when not fully utilized, the ability to gate specific parts of the chip will improve, and the same capability will work into other parts of the system — memory, interconnect (maybe balancing power against bandwidth on a job-by-job basis), I/O, etcetera.

> **There are major cost savings to be had from an open loop cooling system that uses outside air during colder days instead of the traditional closed loop cooling models.**
>
> *— Steve Kinney*
> IBM Computing on Demand

Steve Kinney: Chip designers are designing with performance and power consumption in mind. We are seeing lower wattage processor chips on the market today. The next big item will be the improved design of the memory chips to make them faster, cheaper and more cost effective. Also, above-cabinet cooling is being deployed in more data centers, which are moving away from the traditional raised floor. There are major cost savings to be had from an open loop cooling system that uses outside air during colder days instead of the traditional closed loop cooling models.

Ed Turkel: Without discussing HP futures, some areas of interest include:

1. New, efficient, components under development. (An example is the memristor, a new electrical device that might provide significant gains in power efficiency, which is under development by HP-Labs.)
2. Hyper-efficient power distribution systems within systems, racks, across multiple racks and spanning the datacenter
3. Overall sustainability and dematerialization focus for systems design, from areas as simple as re-usable packaging to systems designed for recyclability or green disposition

Blake Gonzales: Several energy-efficient technologies are primed for entrance into datacenters.

▶ The first is “chiller-less” datacenters, currently being tested by Google. Unchilled air from outside the datacenter is forced across components and then vented outside as it heats up. Cooler climates are great for this. Servers are racked and stacked with minimalistic designs that allow forced air to pass across components easily.

▶ The other is understanding how hot we can run CPUs without failure, or at least understanding the CPU failure rate. Once we have this data in hand, one can design applications to tolerate occasional failures.

Most datacenters deliver AC power to individual components. Typically, power is delivered via AC power lines. Once inside, it’s converted back and forth between AC and DC until ultimately delivered as AC to the server power supply. Unfortunately, every time a conversion is made, energy is lost. On the other hand, DC datacenters convert outside AC power into DC once, which is then delivered as DC throughout the datacenter. Individual server power supplies are no longer needed, as DC is delivered directly to board-level components. The biggest barrier I see to DC power adoption is the ubiquity of AC power paradigm. It will take some time to start to think differently about power delivery.

Steve Scott: Data movement (between nodes, from main memory into the processor, and inside the processor) is the primary culprit when it comes to power usage. Technologies that reduce the cost of data movement will, thus, have the highest impact. 3-D stacked local memory has great potential to reduce the power for main memory bandwidth. Similarly, the integration of optical signaling into the package to enable low-power, long distance communication between chips could have a big impact on the energy cost of network communication. The transition to energy-efficient, heterogeneous processor architectures, and the development of effective and easy-to-use programming software to use these architectures will be the biggest win, however, and perhaps the most challenging.

Bob Masson: Heterogeneous computing is arguably the next green technology. No other technique or technology exists that can provide hundreds or thousands of times improvement in performance/watt for HPC applications. Some form of heterogeneous computing is inevitable. The laws of physics dictate that processor clock rates cannot substantially increase (because it is physically impossible to dissipate the heat generated beyond a certain point). As we’ve heard from John Shalf, head of Berkeley Lab’s Science-driven Systems Architecture team, “*Energy efficiency has become a first-order design constraint for future systems. We really don’t see the current path of scaling up conventional hardware as sustainable either in terms of the initial hardware cost or the price of powering such systems over its lifetime.*”

Performance must be obtained by re-appling a given number of transistors (and hence a given number of watts) to be more efficient. What remains is to make heterogeneous computing usable and available to “the masses” without undue reprogramming or drastic changes to the application development ecosystem. SC