

Energy Efficient High Performance Computing Working Group 12/9/14 Meeting Report

INTRODUCTION

The EE HPC WG held a meeting on 12/9/14. This Working Group is composed of members representing major Federal departments and independent agencies, private sector representatives, and members of the academic community. More information can be found at the working group's website, <http://eehpcwg.lbl.gov>.

NEXT MEETING: February 10th, 9:00-10:00AM Pacific Time

Introductions and Announcements: Dale Sartor, LBNL and Anna Maria Bailey, LLNL

Dale reports, "I am very pleased to announce that Anna Maria Bailey from LLNL will replace me as the Co-Chair of the Energy Efficient HPC Working Group in conjunction with Natalie Bates who will continue on. I am very proud of the accomplishments we have made together. While we are an informal group our reach and impact has been significant. Anna Maria has been an active member of the EE HPC WG from its very early days. She has demonstrated leadership in team activities for conferences, infrastructure and systems. She has also supported the working group administration. Her experience within the HPC community, coupled with her expertise in facilities engineering and management, make her an excellent partner for Natalie, whose background is more focused on computer systems. Together with the help of many volunteers, we will continue to effectively drive implementation of energy efficiency operations and energy efficient design in high performance computing."

Anna Maria responds, "Thank you Dale. I appreciate your kind words of support. I also want to recognize the fantastic results of your work with the EE HPC WG. Your contribution in creating, shaping and growing the EE HPC WG has been instrumental to its success. You brought a strong sense of community and advocacy for energy efficiency across the board as well as a deep background and experience in programs that encourage positive change. I hope and know that you will stay involved with the EE HPC WG and continue to provide leadership and influence. I will miss the most your sharp mind, quick wit and fearless perseverance. I won't fill those shoes in quite the same way (and I won't have the hairstyle either). But, I do thank you and appreciate your support."

Other Announcements:

- The 2014 Procurement Document is out for review with feedback requested no later than 16 January.
- There is a webinar scheduled for 28 January. John Gustafson from Ceranovo will talk about computer arithmetic and energy efficiency. This topic is very 'system'-centric, but promises to be very interesting according to folks like John Shalf and Steve Poole.
- Membership is at 530 – up from just under 500 primarily because of the interest created by SC14 activities.
- The A*STAR Computational Resource Centre in Singapore will host a conference in March called Supercomputing Frontiers 2015. Natalie Bates is participating with them to organize the first Asian HPC Infrastructure workshop as part of the conference.

Conferences Sub-group Update: *Marriann Silveira, LLNL (with help from Tom Scogland, LLNL, Bill Tschudi, LBNL, Steve Martin, Cray, Chung-Hsing Hsu, ORNL and Bob Conroy, OSIssoft)*

SC14:

The EE HPC WG presence at SC14 included a workshop, 4 birds of feather sessions, a panel and an exhibitor booth. Participation was strong in all of these sessions. Thanks to all who participated. For those who couldn't make it, presentations are posted (or in some cases will be posted) on the EE HPC WG website. Also, some of the EE HPC WG organized events will be repeated as webinars in upcoming months.

5th Annual EE HPC WG Workshop

- Dona Crawford is the Associate Director for Computation at LLNL and she opened the workshop. Dona is well known in the Supercomputing community and she helps champion and support the EE HPC WG. Her opening remarks stated that the WG is well aligned with LLNL's goal to improve efficiencies. These glowing remarks set a positive stage for the rest of the workshop.
- Torsten Wilde (LRZ) led a session on *system power measurement methodologies while running a workload* with speakers Daniel Hackenberg (University of Dresden), Robin Goldstone (LLNL) and Tom Scogland (LLNL and Green500). They discussed motivations for adjusting the Level 1 measurement methodology to improve accuracy and reproducibility of results. Daniel discussed the power variability across nodes, suggesting a need to increase the fraction of the compute subsystem that must be measured. Robin described variation in the contribution of the network components across different systems; suggesting a need to include the network in the system power. Tom showed that the measurement should require measuring the entire core phase of the run rather than a subset thereof. All of these have to do with variability and especially changes in system behavior since we originally put the specification together. Each of the presentations seemed to be well received. There were some comments, especially along the lines of raising the requirements yet farther.
- Steve Poole (DoD) led a session on *recommendations for considering energy efficiency during procurement*. The speakers were Jim Laros (SNL), Thomas Ilsche (University of Dresden), Chung-Hsing Hsu (ORNL) and Tom Durbin (NCSA). Each of these presenters covered recommendations that were new to the 2014 version of this document that was just sent out to the EE HPC WG membership for review and feedback. The session also had presentations from Mike Patterson (Intel) Steve Martin (Cray) and Greg Rogers (AMD) that were a commentary on the recommendations. One interesting discussion explored the question of the size of data storage required to support the power and energy measurement defined in the document.
- Anna Maria Bailey, LLNL moderated a session focused on *Exploring opportunities for tighter integration of Supercomputing Centers and Electricity Service Providers*. The speakers included Herbert Huber (LRZ), Rick Wagner (San Diego Supercomputer Center at UC San Diego) and Deva Bodas (Intel). We learned that communicating major power fluctuations to their electricity service providers is required for both LLNL and LRZ.
- The next session was led by Bob Conroy, OSIssoft on *Control System Challenges and Best Practices*. This was organized as a panel with panelists answering pre-prepared questions. Some of the questions were: 1) Why have a control system at all for HPC? 2) Are any feed forward control strategies in use to compensate for load variations due to scheduled or signaled HPC load changes?
- The next session was Mike Patterson (Intel), Ghaleb Abdulla (LLNL) and Chung-Hsing Hsu (ORNL) giving an update on iTUE and TUE. Reporting on this session stimulated a lot of

discussion during the general membership meeting. Dale Sartor, LBNL asked if the EE HPC WG couldn't accelerate the adoption of iTUE and TUE. Anna Maria Bailey, LLNL emphasized that we needed to ask the vendors in the Procurement Considerations document for the ability to measure iTUE. Steve Martin, Cray, chimed in that continuous iTUE measurement capabilities would generate a tremendous amount of data. He further suggested that at least some of the iTUE measurements might be relatively static and only need measuring once – or infrequently.

- The last session of the day was an invited guest speaker, Charlie Manese from Facebook. Charlie discussed how Facebook designs for efficiency and scale and how Facebook contributes those designs to the Open Compute Project.
- There were also “round tables” at lunch that were organized around 5 interest areas. This was new for the workshop and had mixed results. Some of the round tables had lively discussion and were well attended and others, less so. We may try this again next year, but not without tweaking the process.

Liquid Cooling Birds of Feather and Panel

Bill Tschudi reported that the most notable thing was the interest in liquid cooling, especially compared to 3-4 years ago.

- The Liquid Cooling Birds of Feather speakers included Michael Patterson (Intel) Josip Loncaric (LANL), Lynn Parnell (NASA), Tommy Minyard (TACC), and Bruce Myatt (Critical Facilities Solutions and Round Table). The presentations covered liquid cooling infrastructure design, commissioning and controls.
- Michael Patterson reports that the organizers were very pleased with the Liquid Cooling Panel... all 5 presenters did a great job, they all were very open about their designs and the +/- of the different concepts. The presenters were Nic Dube (HP), Ingmar Meijer (IBM), Jean-Pierre Panziera (Bull), Paul Arts (Eurotech) and Thomas Blum (Megware). The organizers had several prepared questions that got good discussion going. Then, the audience began asking questions and they didn't get anywhere near through the prepared questions (that's a good thing!). There was a nice range of questions from the audience. Thursday night, which is very late in the SC week, and there were ~200 attendees. Wow! It's a hot (or maybe a cool) topic.

System and Data Center Metrics and Workloads Birds of Feather

Chung-Hsing Hsu reported on a Metrics and Workload BoF he moderated that was organized by Daniel Hackenberg (TU Dresden), Robin Goldstone (LLNL), and Nicolas Dubé (HP).

Motivated by the limited utility of the PUE metric, the BoF focused on what is needed beyond PUE. Specifically, it had the following goals:

- Having open community dialogue on metrics and workloads for driving the next level of improvement beyond PUE and MFLOPS/W,
- Identifying the right questions to ask, and
- Setting up plans to tackle these problems.

The BoF started with a short introduction of the background, followed by the panel's responses to the questions from the organizers and the audience. The panel consists of four experts: Michael Patterson (Intel), Kevin Regimbal (NREL), Erich Strohmaier (LBL), and Thomas Schulthess (SCS). Chung-Hsing Hsu (ORNL) moderated the panel discussion. There were about 50 attendees in the BoF.

The pre-prepared questions from the organizers are listed as follows:

1. Looking for energy efficient HPC system metrics and workloads, how useful are HPL-based FLOPS/Watt, Graph500-based GTEPS/Watt, and STREAM-based GB/Watt? What else do you deem important?
2. How can the HPC community agree on a (set of) system metric(s) and workload(s) to move from analysis towards optimization?
3. Should the HPC community settle on TUE to overcome the shortcomings of the PUE metric? What do you think is still missing and needs to be improved in datacenter metrics?
4. While throughput/Watt is a system metric, PUE is a datacenter metric. How can we bridge the gap between the two?
5. Are there more important questions we fail to ask?

What constitutes a good metric? The panel agrees that a good metric needs to be simple, easy to measure, and actionable. It has to matter, too. In other words, a good metric should link strongly with a clear goal. For an energy-efficiency metric, the goal is related to energy to solution.

How to agree on metrics and workloads? Some panelists warned that, in practice, important decisions are never made based on benchmarks alone. Each site needs to define metrics and workloads that are meaningful to them. A panelist argued that it is a mistake to consider HPC systems as general purpose. Nowhere else in science do we build general-purpose instruments.

How about TUE? Some panelists like TUE because it fixes a potential problem of PUE. However, the panel agrees that the focus cannot just be on compute. Data intensive workloads are also very important. In addition, energy recovery needs to be factored into the metric. There is a question of whether PUE/TUE points to the most profitable place for investment in energy-efficiency improvement.

How to bridge the gap between system metric and datacenter metric? Only one panelist suggested bolt incompatible numbers together but stopped short on how to do that.

Are there other more important questions that have been overlooked? One panelist thought it a good idea to open up conversation to broader audience. Another panelist concerns about the complexity of the question and suggested not to do so. The panel also expressed the desire to measure productivity (or science output) but acknowledged that it is not measurable.

Finally, the panel provides several suggestions on what to proceed after the BoF. On the outreach front, compiling a list of top PUE sites, providing the how-to on TUE measurement, and informing the governing body best practices are some of the directions. On the research front, clarifying the goals, evolving existing metrics, and finding ways to discourage gaming the system are some possible directions. The panel agrees that the Energy Efficient HPC Working Group is a good place for the community to continue these efforts.

Evolution of the Green500 Birds of Feather

The EE HPC WG – along with the Top500- participated in the Green500 Birds of Feather as collaborators on the EE HPC WG power measurement methodology. Natalie Bates reported that Cray is the first vendor to submit a higher level power measurement- they are to be commended.

Steve Martin, Cray, made a few comments on their Storm system L3 submission:

- 1) Measurement were done on a single cabinet on the manufacturing/test/integration floor.
- 2) Measurements were taken with a high-quality portable Fluke meter with current clamps at the wall panel.

- 3) The biggest obstacle was that the time base of the meter used to take the measurement was not synchronized with the time on the compute nodes. In hindsight a lot of time could have been saved by synchronizing the meters clock before making the HPL runs.
- 4) That fact that all of the equipment was in a single rack, made collecting that data easy, compared to collecting the same level of data on a large multi-rack system... 7.

Steve also reported that there was some amount of confusion with new people reading the power measurement methodology document, understanding the requirements of the various levels, getting the good data that they need.

He continued- doing this on a relatively small system is do-able and affordable. When you start getting multiple racks, multiple rows, may be quite a burden. If the accuracy were less strict, that would allow a lot more participation.

Tom Scogland, LLNL and Green500, replied that was a topic that did come up specifically by the Helmholtz Center system that became the #1 on the Green500 list this year. The methodology actually doesn't specify an accuracy requirement for metering. It references a specification for information on revenue grade meters and that specification requires .2 % accuracy. We are working on refining that to be a little more explicit about what each level requires. What exactly that is going to look like isn't nailed down, but it is definitely in the works. They could have done a level2 if the accuracy was 1% instead of 0.2%. Aside from that, their submission also showed variation across the time of the run depending on which portion of the core phase they took (as much as 50% variation). Instead, they took the average of the entire phase of the core phase. Requiring the entire core phase has been planned already, but it has become abundantly clear.

Piz Dora from CSCS was also a L3 submission.

Dynamic Power Management For MW-Sized Supercomputers Birds of Feather

Bob Conroy, OSIssoft, reported on the Dynamic Power Management BoF. Speakers were Axel Auweter (LRZ), Terry Hewitt (STFC), Tapasya Patki (LLNL and University of Arizona), and Akhil Langer (University of Illinois). There were 75-80 attendees, so there was interest in the topic. We did an informal survey at the BoF to see how many people were aware of the importance of dynamic power management due to Mega-Watt inter-hour power fluctuations. There were some people already paying attention and others interested in understanding more about measuring the variability of power in the HPC environment and how it plays into performance and scheduling. Balancing efficiency with performance; power provisioning and planning, power capping. All of these were included in the presentations which will be posted on the EE HPC WG website.

EE HPC WG Booth

Bob Conroy also reported on the EE HPC WG Exhibitor Floor booth. For the second year in a row we did the 10'*10' booth. Signed up a few dozen new members. This year we solicited support from the membership to help with staffing the booth. Someone in the WG was in the booth at all times. Good time for networking, better awareness for what we're trying to achieve. Looking forward to doing it again in 2015.

Other Conferences:

- The Demand Response Team presented a paper that will be published as part of a Smart Grid Energy Informatics conference in Zurich, Switzerland. The paper was presented by Bo Nørregaard Jørgensen from the University of Southern Denmark.
- The EE HPC WG website lists many upcoming Conferences and Workshops that have an HPC Energy Efficiency Focus

Future Conferences: (more details at <http://ehpcwg.lbl.gov/events-and-links>)

Infrastructure Sub-Group Update: *William Tschudi, LBNL*

LIQUID COOLED COMMISSIONING TEAM:

The Liquid Cooling Commissioning Team has been working with ASHRAE to have them publish an updated version of the EE HPC WG Liquid Cooling Commissioning Guidelines. This will first be published as a whitepaper and then included in the next edition of ASHRAE's Liquid Cooling Guidelines for Datacom Equipment Centers.

ASHRAE TC9.9 and EE HPC WG

The EE HPC WG Infrastructure Sub-group has been working with ASHRAE TC9.9 over the past few years. It started when we made recommendations about inlet water temperatures set-points and ranges for liquid cooling infrastructure, which resulted in the inclusion of these recommendations in ASHRAE's updated Liquid Cooling Guidelines. We're now working with ASHRAE on the commissioning guidelines and we've opened a discussion about working with ASHRAE on a controls document. The relationship is informal and valuable to both groups. The EE HPC WG develops content in areas that are important to HPC and energy efficiency and ASHRAE more broadly disseminates the content in their publications. The ASHRAE TC9.9 Committee has a meeting planned for January and we're working with them to get an EE HPC WG update on the agenda.

CONTROLS TEAM:

There are lessons learned and best practices evolving from implementing and operating supercomputer centers with complex infrastructure systems and the highly variable demands placed upon these systems with today's supercomputers. This team will focus on sharing designs, challenges and best practices for integrated control systems in order to determine if there are universal learnings.

The Team has been meeting regularly with strong participation. They have been sharing controls designs as well as issues and concerns. Bruce Myatt from the Critical Facilities Round Table has taken the lead to outline a whitepaper on HPC controls systems and energy efficiency. This whitepaper is intended to be more of a performance guideline than a technical specification. The whitepaper outline has been reviewed and revised by the Controls Team. The next step is to flesh out the next level of detail for 2-3 of the items on the outline.

TUE TEAM:

As mentioned earlier, both LLNL and ORNL have been testing the iTUE and TUE metrics. iTUE and TUE [Total Power Usage Effectiveness (TUE) and IT Power Usage Effectiveness (iTUE)] account for infrastructure elements that are a part of the HPC system (like cooling and power distribution).

ORNL's test results were interesting. With Titan, ORNL put new, more efficient compute in their racks and did not change the rack/chassis level power and cooling. This led to the question: For the same workload, what is the expectation of the ITUE trend if the original compute part of the system is replaced by a newer system with less energy consumption, but the infrastructure part of the system remains the same? The answer is: this is the same thing as if you put new, more efficient servers in your data center and did not upgrade your room level power and cooling; PUE goes up. ORNL put new, more efficient compute in their racks and did not change the rack/chassis level power and cooling. Their iTUE went up.

The TUE team is seeking more sites to test iTUE and TUE. Anyone interested should contact Natalie.

ENERGY REUSE EFFECTIVENESS:

The Energy Re-use Effectiveness Team in collaboration with The Green Grid has developed a standard metric for measuring the contribution of re-using heat generated by HPC systems for other useful purposes. Florent Parent, Calcul Quebec/Compute Canada is interested in testing this metric at his site. Anyone else interested in sharing your experiences or testing the ERE metric should contact Natalie.

Systems Sub-group Update: *Natalie Bates, EE HPC WG*

SYSTEM WORKLOAD POWER MEASUREMENT METHODOLOGY:

The EE HPC WG along with the Green500, Top500 and Green Grid have developed a standard methodology for measuring system power while running a workload. The ultimate goal is to have broad use of the highest quality energy and power measurement methodology for all of their system workload energy efficiency benchmarking activities.

HPC AND GRID INTEGRATION:

The Demand Response Team is investigating how HPC centers have, can and should engage more actively with the Grid electricity providers. This is an investigative activity with the ultimate goal of educating the HPC DOE Facility and Operations Managers about HPC and grid integration opportunities and challenges.

The Team has written a paper analyzing data collected from 11 US-based Supercomputing Center sites that are on the Top100 list.

The team is now focused on extending this work to European-based SC sites that are on the Top50 list. Nine sites have provided data: CEA and EDF from France, LRZ, Juelich and Stuttgart from Germany, CSCS from Switzerland, ECMWF from the UK, Cineca from Italy and KTH from Sweden. The team has developed another questionnaire that will be used to collect data from Electricity Service Providers in these countries. It focuses on their interests and involvements (both current and planned) in grid integration. This data will allow for comparing and contrasting electricity markets between countries in Europe and between the US and Europe.

One of the key learnings from this analysis is that some of the tools that are being developed to help with energy efficiency may also be used in the future for electricity grid integration. An example of such a tool is dynamic power management, which was the subject of discussion for the SC14 BoF that Bob Conroy reported on earlier.

PROCUREMENT CONSIDERATIONS:

The RFP Team has written a whitepaper that recommends procurement document requirements that target more energy efficient HPC systems. The intention is to raise the bar and extend the requirements with a yearly update of the whitepaper. The first year was 2013.

The Team completed a draft of the 2014 update and sent it out the EE HPC WG membership for review and feedback. The 2014 version has four new or enhanced sections: 1) enhanced measurement section with more detailed definitions and descriptions, 2) new section on timestamping and clocks, 3) new section on temperature measurements and 4) enhanced section on air and liquid cooling. Intel, AMD and Cray all reviewed and gave feedback on a draft of this document.

SW UPDATE:

Three efforts continue to develop momentum; these are 1) creating an on-line annotated list with links of energy efficiency workloads/benchmarks 2) promote development of a power measurement API, such as the work being led by Jim Laros from Sandia National Laboratory and 3) share best practices for dynamic power management.

PARTICIPANTS INCLUDED

Name	Organization
Anna Maria Bailey	LLNL
Natalie Bates	EE HPC WG
Andrea Bartolini	ETH
Anita Cocilova	LLNL
Lloyd Brown	BYU
Andrew Cassidy	IBM
Hsu Chung-Hsing	ORNL
Bob Conroy	OSIsoft
Mike Ellsworth	IBM
Jeff Filliez	SDSC
Mike Garceau	3M
Nathan Gregg	
Kim Griger	
Eric Grunebaum	Teracool
Paul Henderson	Princeton University
Jacob Jenson	SchedMC/SLURM
Detlef Labrenz	LRZ
Thomas Leung	GE Global Research
Steven Martin	Cray Inc.
Dave Martinez	SNL

Dale Sartor	LBNL
Marriann Silveira	LLNL
Mike Thomas	ESD Global
Bill Tschudi	LBNL
Philip Tuma	3M
Graeme Walker	
Ralph Westcott	PNNL

EEHPC WG