Global Survey of Energy and Power-aware Job Scheduling and Resource Management in Supercomputing Centers

**System Design Challenges:**
- Building systems for HPC under a Power Budget
- Peak power demands for future Exascale systems > 20MW
- Instantaneous power fluctuations: 8MW
- Microarchitecture improvements and high degree of parallelization not sufficient

**Eight Survey Questions for the sites:**
1. Motivation behind investing in Energy and Power Aware Job Scheduling and Runtime Management (EPA-JSRM)
2. Target infrastructure (e.g. site-wide power budget, cooling capacity, etc.)
3. Workload characteristics
4. Adopted design for EPA-JSRM
5. Implementation details for EPA-JSRM
6. Application/task level and topology-aware solutions
7. Results and challenges
8. Next steps including system procurement

**Participating Sites:**
- CEA (Alternative Energies and Atomic Energy Commission), France
- Cineca, Italy
- KAUST (King Abdullah University of Science and Technology), Saudi Arabia
- LRZ (Leibniz Supercomputing Centre), Germany
- Riken, Japan
- STFC (Science and Technology Facilities Council), United Kingdom
- Tokyo Institute of Technology, Japan
- University of Tokyo and University of Tsukuba (JCAHPC), Japan
- Los Alamos and Sandia National Laboratories (Trinity), United States

**EPA-JSRM solutions depicted on the right have already been adopted - at least, in parts by these sites:**
- Automated reduction of node availability by the resource manager
- Leveraging power-aware design of portable APIs
- Design of system-wide frameworks (like job schedulers) that use static-based metrics
- Use of power-capping mechanisms supported by CPU and system vendors
- Ongoing design of high-level APIs for end-users and resource managers
- Implementing of statistical approaches for prediction

**ISRM solutions adopted:**
- Dynamic shutdown of jobs in response to limited power budget (Reactive approach)
- Job selection based on job size, job length, etc. to shut down
- Automated reduction of node availability by the resource manager (Proactive approach)
- Reduces the theoretical maximum power that can be consumed
- Drop in system utilization
- Use of power-capping mechanisms supported by CPU and system vendors
- Attempts to keep total power consumed below a specific limit
- Power cap applied over a specific time-window (in order of minutes)
- Use system interface to trigger specific power capping mechanisms
- Design of portable APIs
- Design of system-wide frameworks (like job schedulers) that use static prediction models
- “Tagging” applications based on their power usage characteristics (feedback-driven approach)
- Mapping of “tags” to performance metrics
- Storage of historical records attained over past job runs
- Use of tag-values for future budget assignment

**Telemetry Monitoring solutions adopted:**
- Sensors for monitoring energy and power
- Both in-band as well as out-of-band
- Direct real-time measurements
- Thermal-based sensors coupled with prediction models
- Model to indirectly derive power-based metrics
- Ongoing design of high-level APIs for end-users and resource managers
- Energy and power monitoring
- Feedback mechanisms

**Next phase of JSRM roadmap:**
- Continue working on more stable designs of system-wide frameworks (e.g. job schedulers) that allocate resources in a power-aware manner
- Invest in robust energy/power predictors that rely on statistical modeling
- Leverage power-capping mechanisms exposed by vendors

**System Characteristics**

<table>
<thead>
<tr>
<th>Organization</th>
<th>Site Power Budget</th>
<th>Site Cooling capacity</th>
<th>Major HPC System</th>
<th>System Power Draw</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIKEN</td>
<td>Up to 25 MW</td>
<td>Up to 40 MW</td>
<td>K computer (83,944 nodes)</td>
<td>Up to 15 MW</td>
</tr>
<tr>
<td>Tokyo Tech</td>
<td>Up to 5 MW</td>
<td>Up to 5 MW</td>
<td>TSUBAME2.5 (1400 nodes)</td>
<td>Up to 5 MW</td>
</tr>
<tr>
<td>CEA</td>
<td>Up to 10 MW</td>
<td>Up to 10 MW</td>
<td>Anticipated ZSPF System in 2017</td>
<td>Up to 5 MW</td>
</tr>
<tr>
<td>KAUST</td>
<td>Up to 5 MW</td>
<td>Up to 5 MW</td>
<td>Shaheen 2 (6174 nodes)</td>
<td>Up to 5 MW</td>
</tr>
<tr>
<td>LRZ</td>
<td>Up to 10 MW</td>
<td>Up to 15 MW</td>
<td>SuperMUC Phase 1/2</td>
<td>Up to 5 MW</td>
</tr>
<tr>
<td>STFC</td>
<td>Up to 5 MW</td>
<td>Up to 5 MW</td>
<td>846 x dual SKX (128GB), 840 x KNL (96GB), 24x dual SKX (1TB)</td>
<td>**</td>
</tr>
<tr>
<td>LANL + SNL</td>
<td>Up to 20 MW</td>
<td>Up to 30 MW</td>
<td>** Trinity (9436 HSW nodes + 9984 KNL nodes)</td>
<td>Up to 10 MW</td>
</tr>
<tr>
<td>Cineca</td>
<td>Up to 10 MW</td>
<td>**</td>
<td>Marconi (7500 nodes)</td>
<td>Up to 4 MW</td>
</tr>
<tr>
<td>JCAHPC</td>
<td>Up to 10 MW</td>
<td>**</td>
<td>Oakforest PACS (8208 nodes)</td>
<td>Up to 5 MW</td>
</tr>
</tbody>
</table>

**Notes:**
- **Information unavailable as of Oct 2017**

---

**The Efficient HPC Working Group (EEHPC-WG) invites other supercomputing sites to participate in enhancing this survey. It welcomes questions, feedback, and comments from the entire HPC community.**

This QR code alongside, points to the EEHPC-WG webpage (https://eehpcwg.lnl.gov/) that contains additional links to the white paper related to this poster, the feedback form, and other information on ways you can participate in this era of Energy and Power-aware computing.