

State of the Practice: Energy and Power Aware Job Scheduling and Resource Management (EPA-JSRM)

EEHPC Working Group
EPA-JSRM sub-team

November 15, 2017
SC'17, Denver CO

Focus of this BoF:

State of the Practice: Energy and Power Aware Job Scheduling and Resource Management (EPA-JSRM)

Background of the EPA-JSRM team:

- Sub-team under the Energy Efficient HPC Working Group
- Includes interested members from across the globe
 - HPC Centers/Facilities
 - Researchers
 - Vendors
- Goal: Assess the environmental, computational, and usage drivers motivating power management efforts

Recent work:

- Identify the State of Practice regarding EPA-JSRM
- Survey done in 2016/2017
- White-paper + poster published

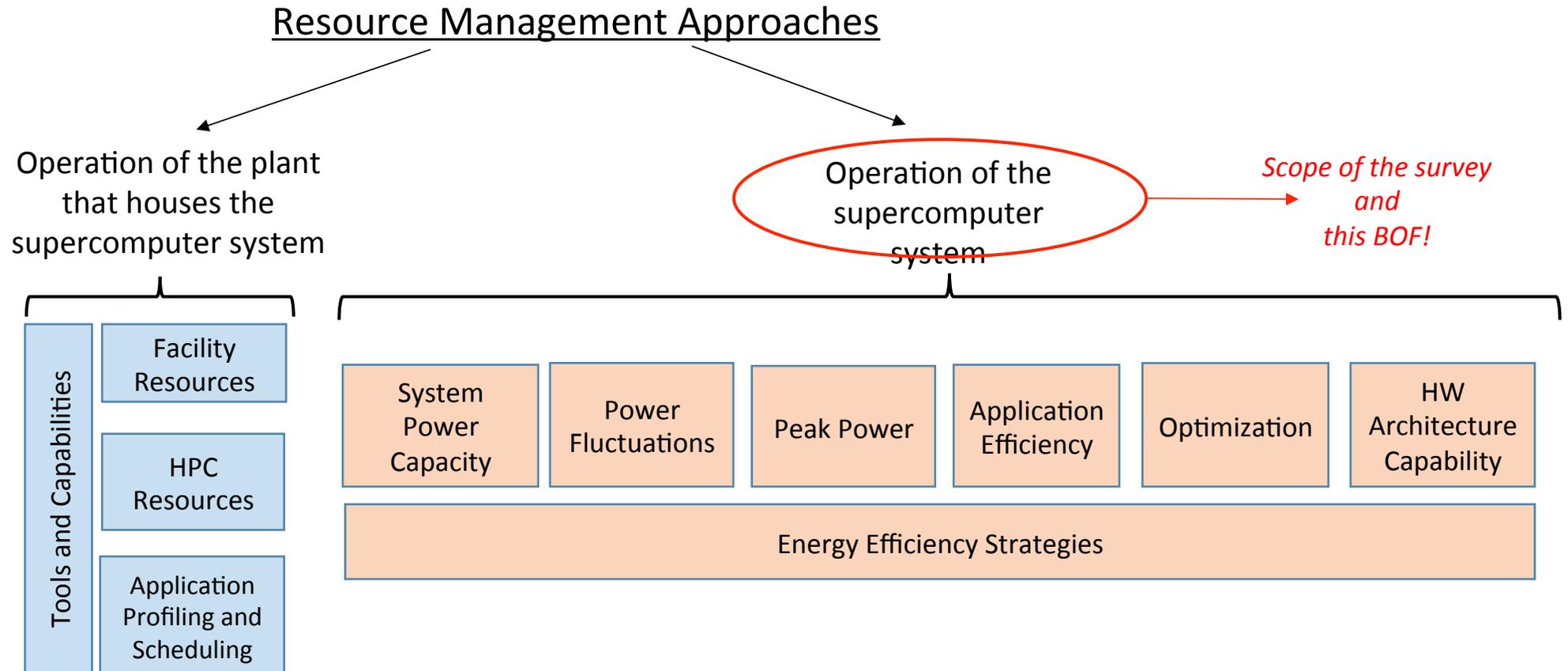
Today's Agenda

- Summarize key takeaways from the white paper
- Get feedback from the audience on the next steps and future roadmaps.

Energy and Power Aware Landscape in System Design

System Design Challenges:

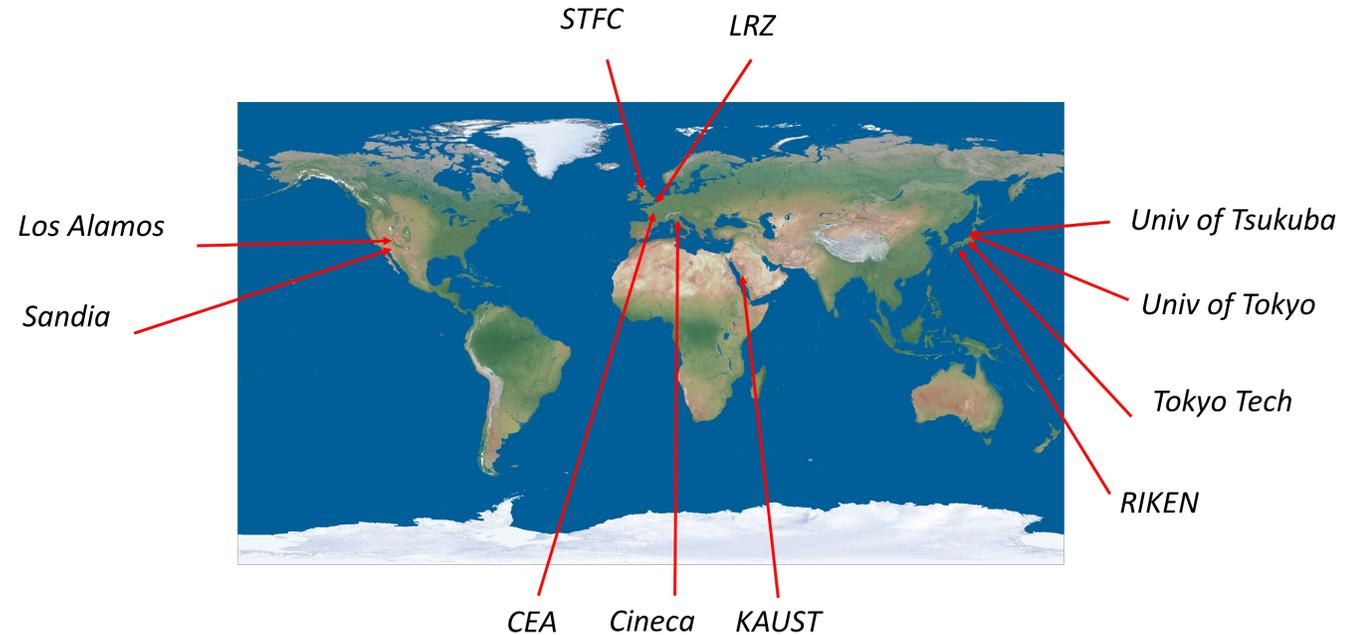
- Peak power demands for future Exascale systems ~20-30MW
- Microarchitecture improvements and high degree of parallelization not sufficient



First global survey of EPA-JSRM capabilities among supercomputing sites

Participating Sites :

- **CEA** (Alternative Energies and Atomic Energy Commission), France
- **CINECA**, Italy
- **JCAHPC** (University of Tsukuba, University of Tokyo), Japan
- **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
- **Trinity** (Los Alamos and Sandia National Laboratories), United States



Criteria for inclusion in the survey:

- Be **actively pursuing** an EPA-JSRM solution, **and**
- Targeting solution on a **large-scale HPC system**, **and**
- Be investing in technology development with the intention of using the EPA-JSRM solution in the site's **production computing environment**.

Survey Questionnaire

Survey Responses to be discussed today...

- Motivation for investing in EPA-JSRM solutions
- Adopted design and implementation details
- Results and challenges
- Next steps

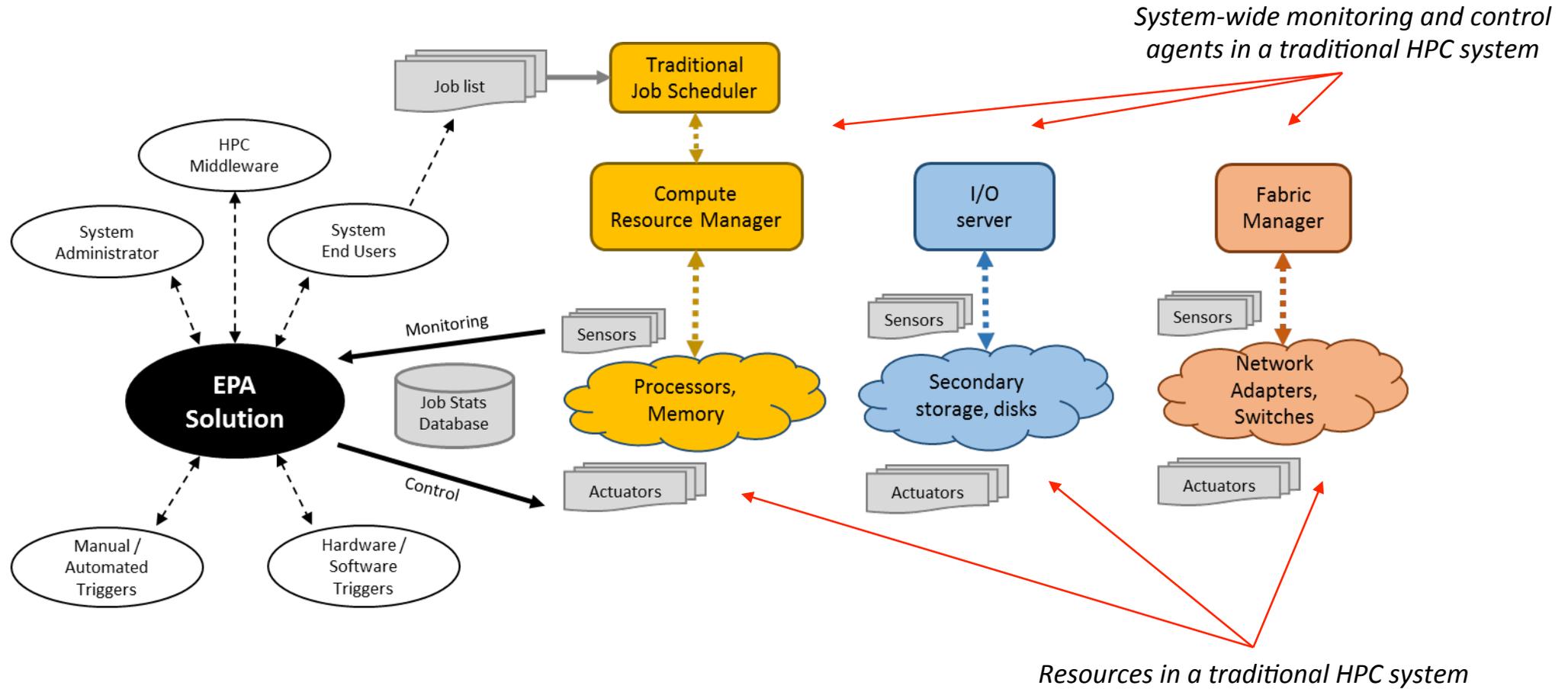
Remaining questions addressed in more detail in the white paper:

- Target infrastructure & workload characteristics/
- https://eehpcwg.llnl.gov/pages/conf_sc17a.htm

Motivation for investing in JSRM solutions

- Power constraints due to external factors
 - Natural disasters, shortage of electricity
 - Government mandates, limits to operation costs
- Power limits imposed due to internal infrastructure limitations
- Motivation for staying "ahead of the game" while dealing with power constraints
 - Investments in predictability and stability of power consumptions of future systems
- Prioritization of higher compute power by limiting secondary infrastructure costs like cooling, etc.
- Education and evaluation of end users
- Ecological responsibility: desire to "be green"

EPA-JSRM design overview



Measurement and Modeling solutions

- 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

EPA-JSRM Solutions Adopted

#	<u>Approach</u>	<u>Challenges</u>
A.	Dynamic termination of jobs <ul style="list-style-type: none"> • Job selection based on job size, job length, etc. to shut down • RIKEN 	<ul style="list-style-type: none"> • Choosing the right metric for terminating jobs
B.	Automated reduction of node availability <ul style="list-style-type: none"> • Reduces the theoretical maximum power that can be consumed • Resource manager and job scheduler play an important role • Tokyo Tech 	<ul style="list-style-type: none"> • Drop in system availability • Already shut-down nodes take time to boot up. This increases queue wait-time for jobs that are waiting for those nodes.
C.	Use of power-capping mechanisms <ul style="list-style-type: none"> • Attempts to keep total power consumed below a specific limit. • Power cap applied over a specific time-window, e.g. Intel RAPL • SLURM – SDPM, Cray’s CAPMC • KAUST, Tokyo Tech, JCAHPC, Cineca 	<ul style="list-style-type: none"> • Out of band / SLURM SDPM: High performance variability in performance has been observed low queue wait times, coarse grained power limiting.
D.	Leveraging p-states (for specific jobs); c-states + s-states (for idle nodes) <ul style="list-style-type: none"> • SLURM, IBM Load Leveler -- Platform LSF • CEA, LRZ, STFC 	<ul style="list-style-type: none"> • Design of standardized user interfaces / portable APIs • Granularity of p-state change – per process v/s job • Platform specific
E.	Static prediction models <ul style="list-style-type: none"> • System-wide control • Implemented within job schedulers, and used history of past runs • IBM Load leveler • LRZ, STFC 	<ul style="list-style-type: none"> • Selection of input parameters for the static model
F.	Tagging applications based on power characteristics <ul style="list-style-type: none"> • Mapping of “tags” to performance metrics • Tag-values for future budget assignment • LRZ, STFC 	<ul style="list-style-type: none"> • Dependent on user input • Need to maintain historical records

Short and Long Term Goals

- To be used in procurement documents
- Strong interest in continuing development and deployment

Long term goals

- Implement power estimator for the jobs
- Invest in extending power capping mechanisms to multiple systems within the same site
- Incorporate facility power and cooling information within the JSRM solution

Next Steps...

<u>For sites</u>	<u>For vendors</u>	<u>For the community</u>
<ul style="list-style-type: none">• What granularity do you see your site adopting EPA-JSRM solutions – job, node, socket, core, memory, network ?• What opportunity analysis data is needed to encourage adoption of fine-grained control?	<ul style="list-style-type: none">• (Schedulers)<ul style="list-style-type: none">• Interfaces for collecting power/energy constraints from users / sys admins• Topology-aware placement• (OS) Kernel modules for controlling power/ energy• (System components)<ul style="list-style-type: none">• Interfaces for setting/ reading power/energy control knobs, MSRs, CSRs	<ul style="list-style-type: none">• Standardization of Interfaces across components• Need for additional surveys

Next Steps...

(input from audience)

<u>For sites</u>	<u>For vendors</u>	<u>For the community</u>
<ul style="list-style-type: none"> • What granularity do you see your site adopting EPA-JSRM solutions – job, node, socket, core, memory, network ? <ul style="list-style-type: none"> -- <i>Expose user priority(account for turnaround time)</i> -- <i>incentive for users (e.g. CPU allocation time, tie that with power mgmt solutions, constrained resource)</i> -- <i>shortlist metrics</i> • What opportunity analysis data is needed to encourage adoption of fine-grained control? <ul style="list-style-type: none"> --> <i>LRZ has research, not deployment ready</i> --> <i>KAUST: two job queues (high/low limit), design different versions of the same code -- user's responsibility, instrumenting codes</i> --> <i>CRAY: profiling tools (e.g. craypat, perftools)</i> --> <i>STFC: visuals for load imbalance</i> <i>ORNL's challenges:</i> <ul style="list-style-type: none"> -- <i>reliability of profiling tools?</i> -- <i>tools: prefer tools delivered by vendors</i> <i>DOD: main concern:</i> <ul style="list-style-type: none"> <i>Total cost of ownership, operation costs</i> <i>7:1 = (operation costs: energy costs)</i> 	<ul style="list-style-type: none"> • (Schedulers) <ul style="list-style-type: none"> • Interfaces for collecting power/energy constraints from users / sys admins • Topology-aware placement <ul style="list-style-type: none"> -- <i>relying on user-input -- skepticism</i> -- <i>machine learning, data analytics</i> -- <i>topology awareness: expose topology in some format, n/w related placement (performance --> energy efficiency), balancing nodes (monitoring interfaces)</i> • (OS) Kernel modules for controlling power/ energy • (System components) <ul style="list-style-type: none"> • Interfaces for setting/ reading power/energy control knobs, MSRs, CSRs 	<ul style="list-style-type: none"> • Standardization of Interfaces across components • Need for additional surveys

Closing thoughts...

- Do get in touch if you would like to participate!
- Contact Info:
 - Natalie Bates
 - natalie.jean.bates@gmail.com
 - Subscribe to EPA-JSRM mailing list:
 - Google group: epa-jsrm@googlegroups.com