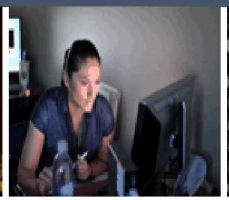
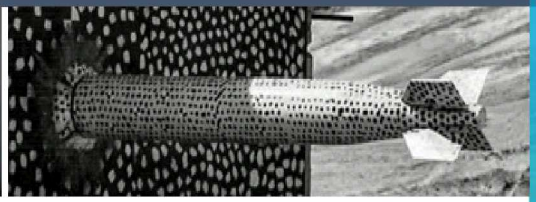
SAND2020-6206C

PowerAPI: A Standardized Interface to Power/Energy Monitoring and Control





Ryan E. Grant
Center for Computing Research
Sandia National Laboratories
regrant@sandia.gov

PRESENTED BY

CCR
Center for Computing Research

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Power API Survey

- Goal: Clarify exactly what it means to set power/energy on nodes
- Multi-vendor
- Concentrating on HPC systems

- Describe the high level architecture as it pertains to compute resources (processors, memory, IO, accelerators) and how the power consumption, temperature, and frequency/Voltage are set and controlled.
- Does setting a frequency lock the clock to that specific frequency?
  - If not, what bounds are provided on the frequency provided?
  - What throttling limits/actors exist in your system?
- Does your system provide hard guarantees on power and/or frequency?
- Does your system provide soft guarantees on power and/or frequency?
- Can frequencies be set independently on discrete cores?
  - What power domains exist in your designs that limit frequency and/or power settings?
- Can CPU frequency be set independently from memory frequency?
  - Is CPU frequency tied to any other system bus or component frequency?
- Does your system provide abstractions of power/performance states in any way?
  - If so, are these states linear or non-linear in terms of performance?
- Is frequency setting reproducible? That is, if frequency is changed at a given point in an application, is the change observable in the same manner each time? This assumes the application behavior itself is reproducible.
  - Does there exist a setting or a range of settings where reproducibility is provided or not?
  - If results are reproducible normally, are there any situations in which this may not be true.
- Do you have states/modes that are not tied directly to a frequency voltage pair?

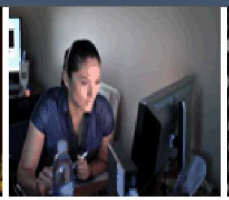
Survey Lessons Learned

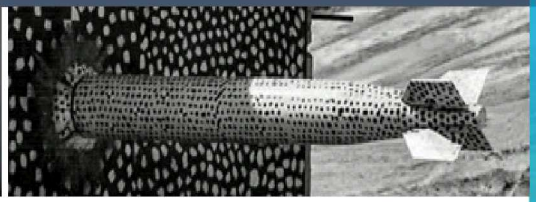
- Caveat: The survey is still in progress
- Some of the insights so far have been represented in questions in the survey
  - For example: reproducibility questions
- Key takeaway: Setting a frequency does not always "set" the frequency
- Ramifications: Power API metadata improvement
  - Emphasized that rich metadata is required to understand a system

Takeaways continued

- Important to have multi-node interface
  - Other components of system (e.g. RM) need to find info on what it means to set states
- Power capping systems is harder than it seems at first
- Abstraction of system with interrelated components is important
  - Power API system model is sufficient for this today

PowerAPI: A Standardized Interface to Power/Energy Monitoring and Control





Ryan E. Grant
Center for Computing Research
Sandia National Laboratories
regrant@sandia.gov

PRESENTED BY

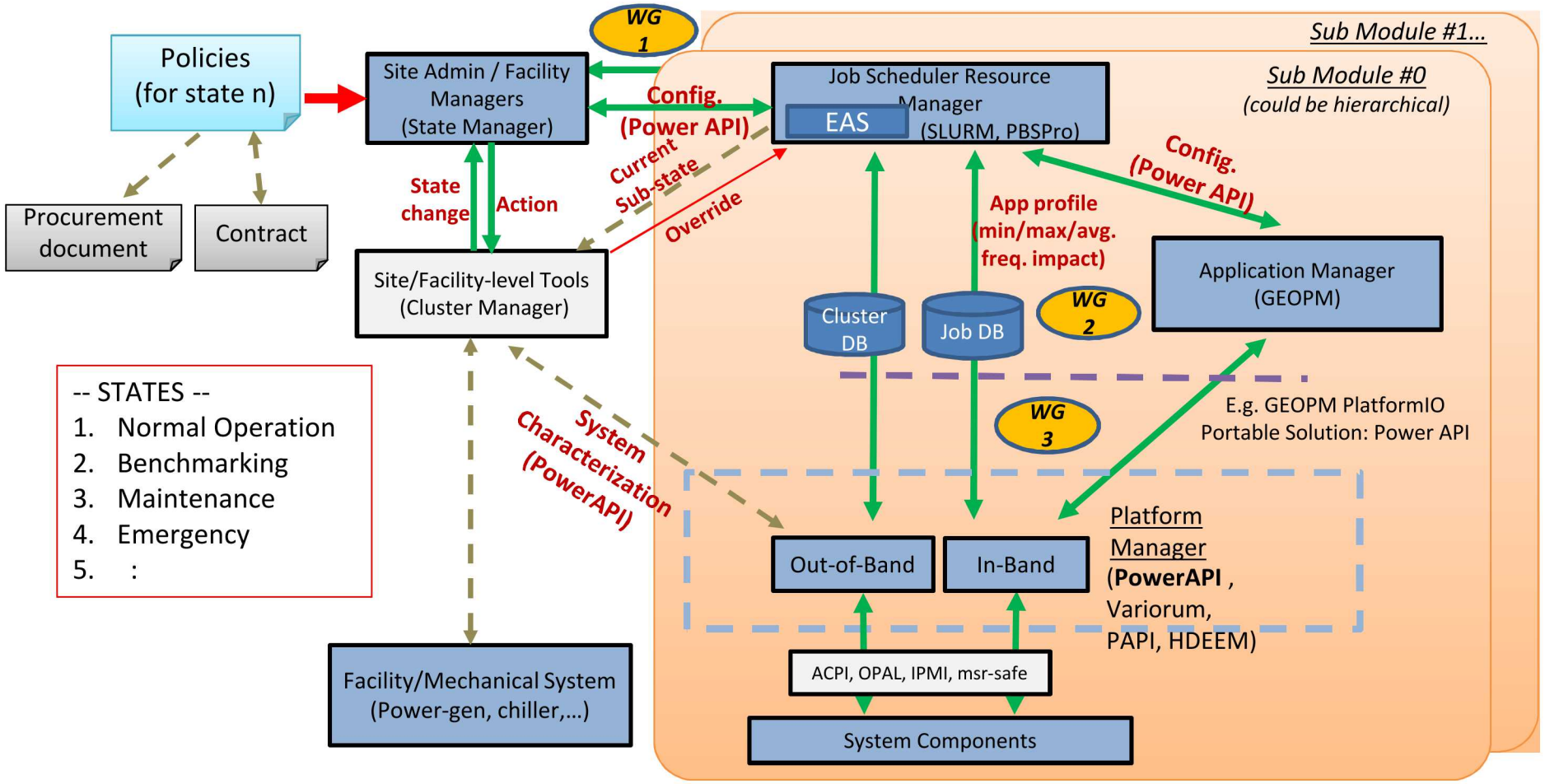
CCR
Center for Computing Research

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

What is the PowerAPI?

- The PowerAPI is a comprehensive system software API for interfacing with power measurement and control hardware
- Designed to be comprehensive across many different levels of a data center
- Many different actors can interface with a single API to perform several different roles
- Encompasses facility level concerns down to low level software/hardware interfaces

HPC PowerStack High Level Flow



Updates

Power API Version 1.0 released!
https://github.com/pwrapi/powerapi_spec

Community model:

- New Specifications Document
- Open meetings
- Multi-institution involvement

Power API

- Already have tools, installations and interfaces for many different types of hardware
  - Slurm plugin
  - Redfish interoperability

- Pwr tool – command line easy to use power/energy collection tool
  - Used like date or time with flexible collection options for more advanced use
  - Really useful for batch job scripts
  - Fully portable between systems

Solving Tough Problems

- Multi-actor setting manipulation is hard to solve
  - Multiple ways to interact with power/energy settings
  - Hardware can override user control
  - Hierarchy of permissions – HW, OS, user/apps
  - Middleware/OS/users much more difficult to control multi-actor problem
- Valid use cases for having multiple actors
  - Need to notify actors that want to know if settings have changed

Power API solution

- Allow registering for notification/callback on settings change

- Much easier to capture just Power API interactions
  - Always some elements that are not using Power API (e.g. Hardware)

- Proposal: two different modes
  - 1) just monitor for changes done through Power API
  - 2) Power API daemon launch to observer changes
    - Much more expensive

Power API solution

- Using Power API throughout PowerStack would be useful to protect multiple layers of the stack

- Easy to monitor for changes that are not forced via hardware
  - If hardware is modifying things, there's probably a good reason

- Work on this is ongoing and opportunity to participate: https://eehpcwg.llnl.gov/meetings.html

Thank you

Questions?

regrant@sandia.gov

Acknowledgments:

This work was funded through the Computational Systems and Software Environment sub-program of the Advanced Simulation and Computing Program funded by the National Nuclear Security Administration