Power Management Framework for Extreme-Scale Computing

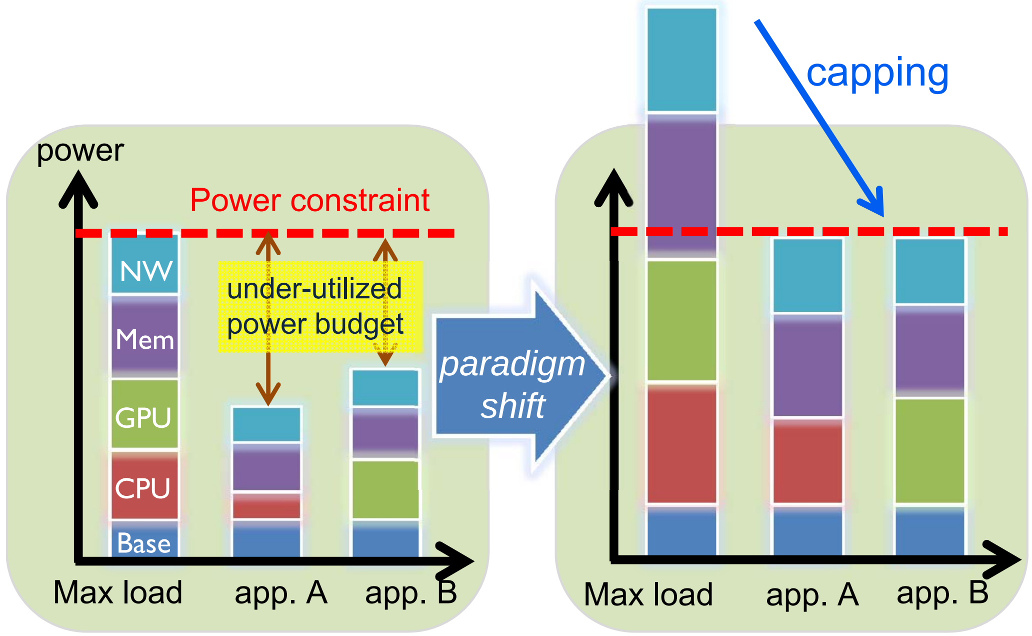
Masaaki Kondo

Graduate School of Information Science and Technology, The University of Tokyo.
Information Technology Center, The University of Tokyo.

- Power: A first class design constraint in Extreme scale systems
  - 10-20PFLOPS with about 10MW electricity in today's top supercomputers
  - Practical range of power budget : 20 - 30MW
  - About 50x improvement in power-efficiency towards Extreme-scale systems
- Needs paradigm shift to power-constraint adaptive system design
- Key challenges
  1. Framework to maximize application performance under a given power constraint
  2. Power aware job scheduling to maximize total system throughput and to minimize under-utilized power budget
  3. Power-performance simulation and analysis framework
  4. Standardized API for power monitoring and control

Paradigm Shift to Power Constrained Systems

Power-Constrained Adaptive System



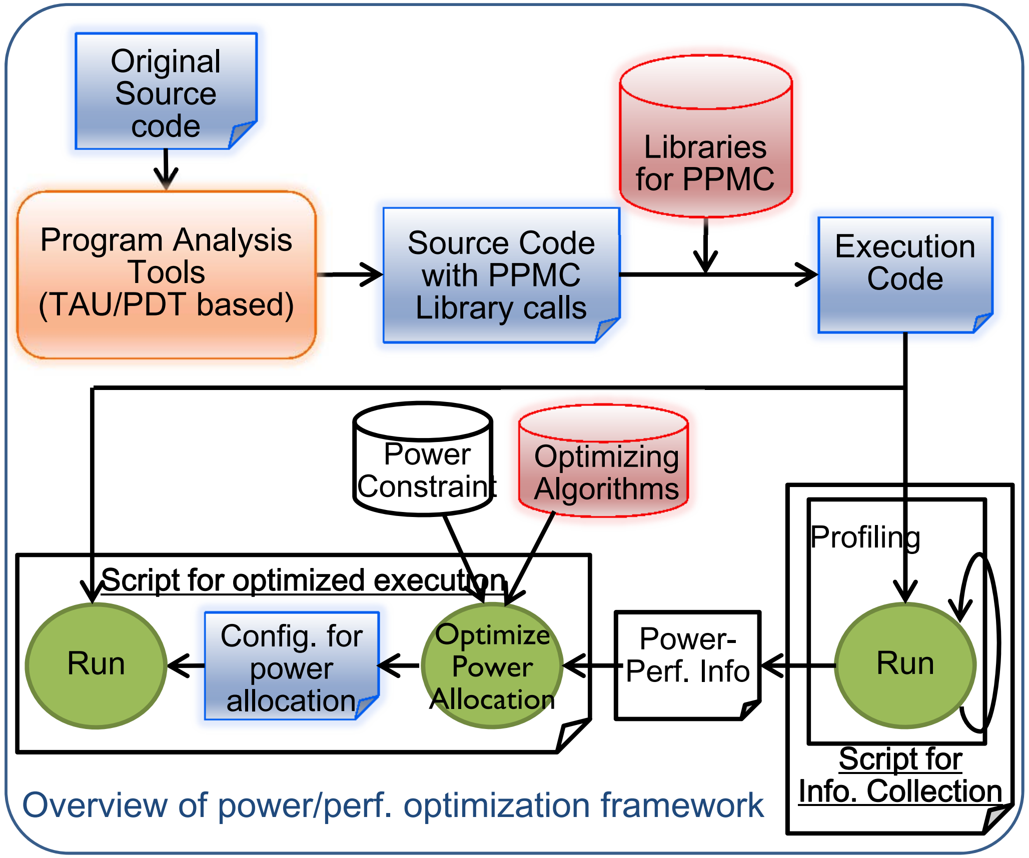
conventional

Extreme-scale

- Allows peak power to exceed the constraint (HW over-provisioning)
- Controls power-knobs to make effective power below the constraint
- Improves performance by allocating power budget to each component

Power-Perf. Optimization Framework

- helps optimize performance within a power constraint
- Integrated framework of compiler, profiler, and runtime tools

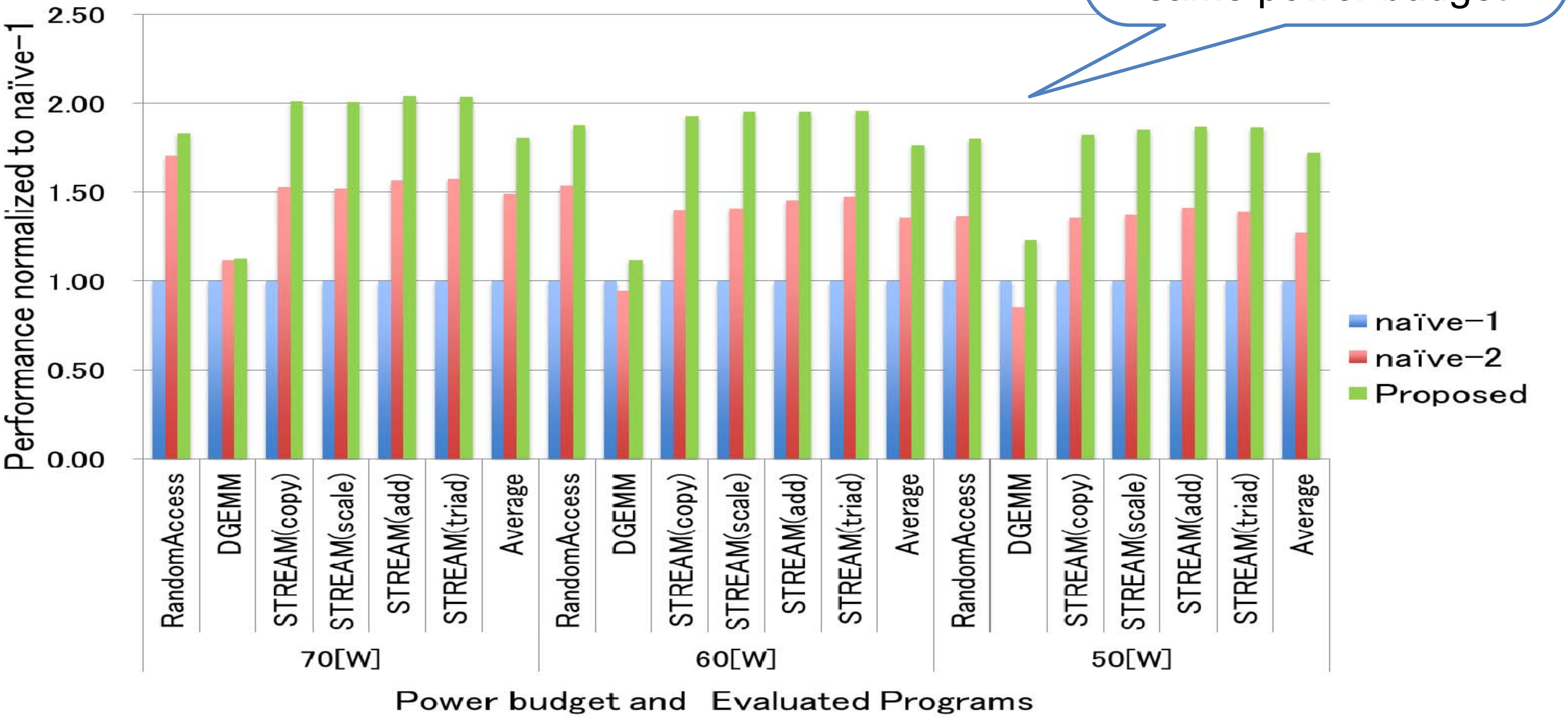

Original Source code
Libraries for PPMC
Program Analysis Tools (TAU/PDT based)
Source Code with PPMC Library calls
Execution Code
Power Constraint
Optimizing Algorithms
Profiling
Script for optimized execution
Run
Config. for power allocation
Optimize Power Allocation
Power-Perf. Info
Run
Script for Info. Collection

Overview of power/perf. optimization framework

Example: Optimizing CPU-Memory Power Allocation

- Evaluation on an Intel Xeon E5-2620 system
  - RAPL for power-cap control
  - naïve-1: allocate power with same ratio as TDP
  - naïve-2: power model based on effective power

Fine tuning based on detailed power model
↓
2x performance with the same power budget

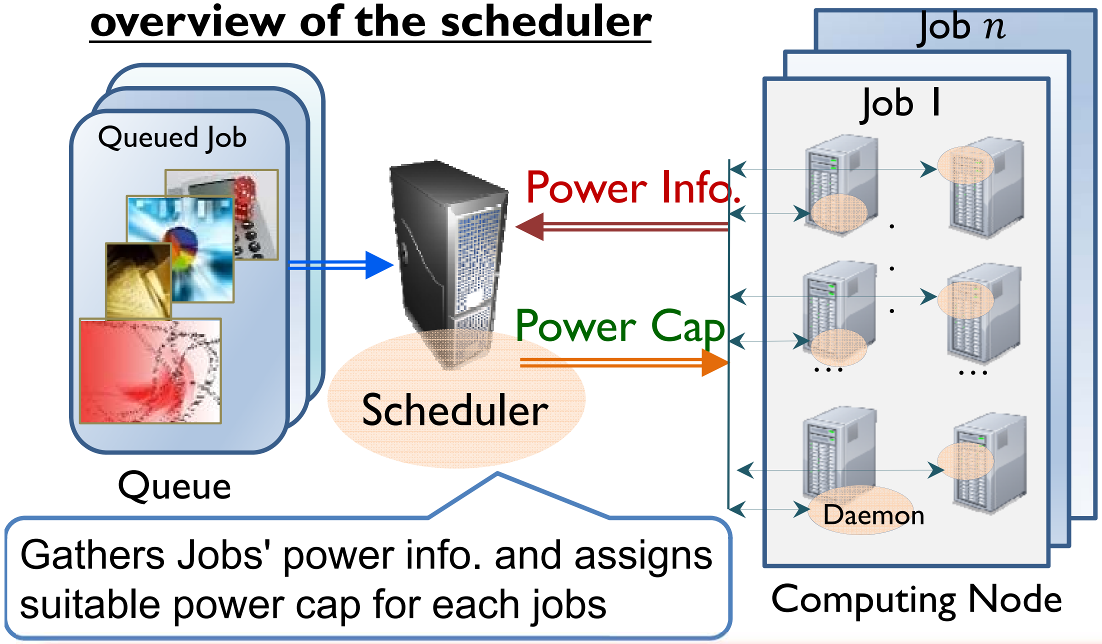


Job Scheduling with Power Allocation

Power aware Job Scheduling

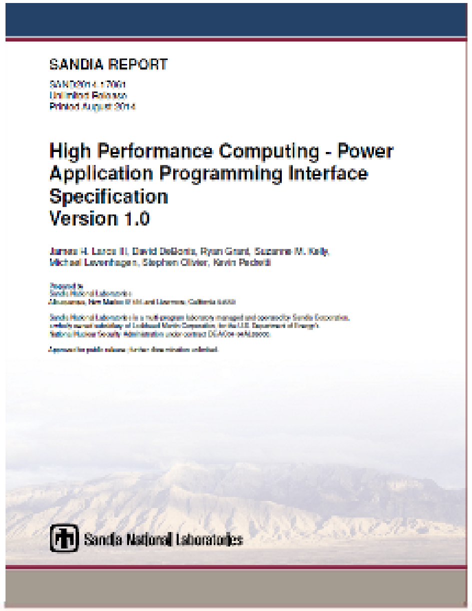
- When, which and where job should be executed to optimize total job throughput under power constraint
- Dynamically allocate power-cap to each job based on it’s priority

overview of the scheduler



Standardized API

- Need easy to use, machine/host independent, eternally available API for HPC eco-system
- Recent effort in SNL
- Community wide discussion is indispensable



SANDIA REPORT

SAND2014-17061
Unlimited Release
Printed August 2014

High Performance Computing - Power Application Programming Interface Specification
Version 1.0

James H. Laros III, David DeBonis, Ryan Grant, Suzanne M. Kelly, Michael Levenhagen, Stephen Olivier, Kevin Pedretti

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

All these need international collaboration and discussion!