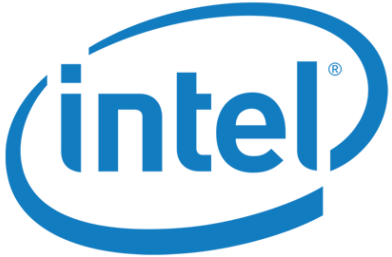
Global Extensible Open Power Manager

http://geopm.github.io/geopm



Matthias Maiterth
[matthias.maiterth@intel.com]

Workshop on Energy Efficiency in HPC (organized by the WG on Energy Efficiency of ETP4HPC) - part of the European HPC Summit Week 2018
30 May 2018, Ljubljana

GEOPM Core-Team:

- Asma Al-Rawi
- Fede Ardanaz
- Brandon Baker
- Chris Cantalupo
- Jonathan Eastep (Lead)
- Brad Geltz
- Diana Guttman
- Siddhartha Jana
- Fuat Keceli
- Kelly Livingston
- Matthias Maiterth

GEOPM Motivation

Performance of future large-scale HPC systems will be limited by power costs.

Today's power management techniques don't manage power optimally:

- Static frequency selection is a suboptimal strategy, since app consist of computational phases with distinct frequency-runtime sensitivity
- Uniform power capping exposes processor performance variation
- Processor locally decides to Turbo, irrespective of critical path

Making wiser use of power requires a breakthrough in power management strategy with much more global, dynamic application awareness!

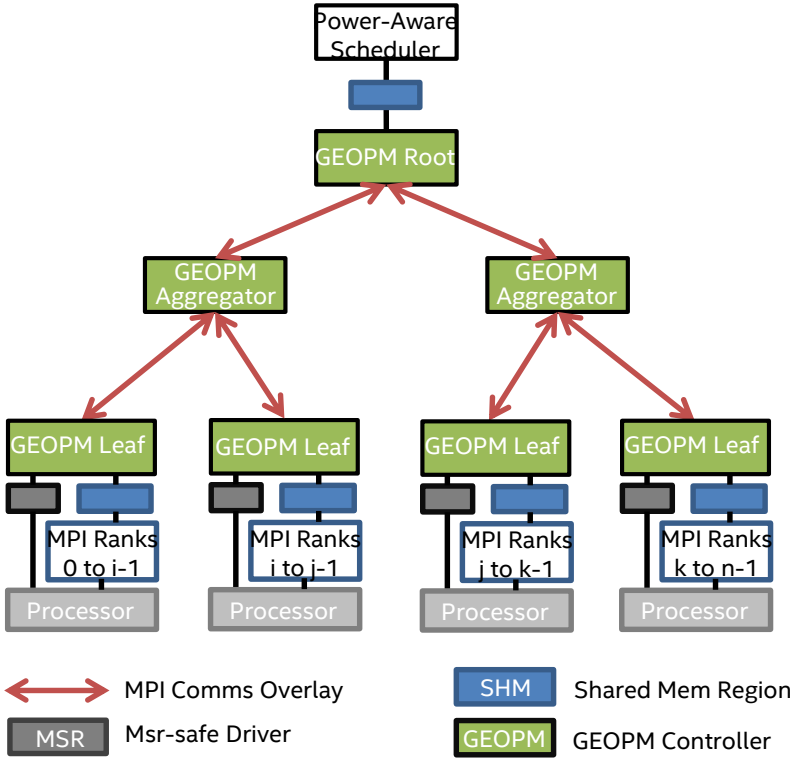
A solid foundation requires collaboration across the HPC community.

Global Extensible Open Power Manager

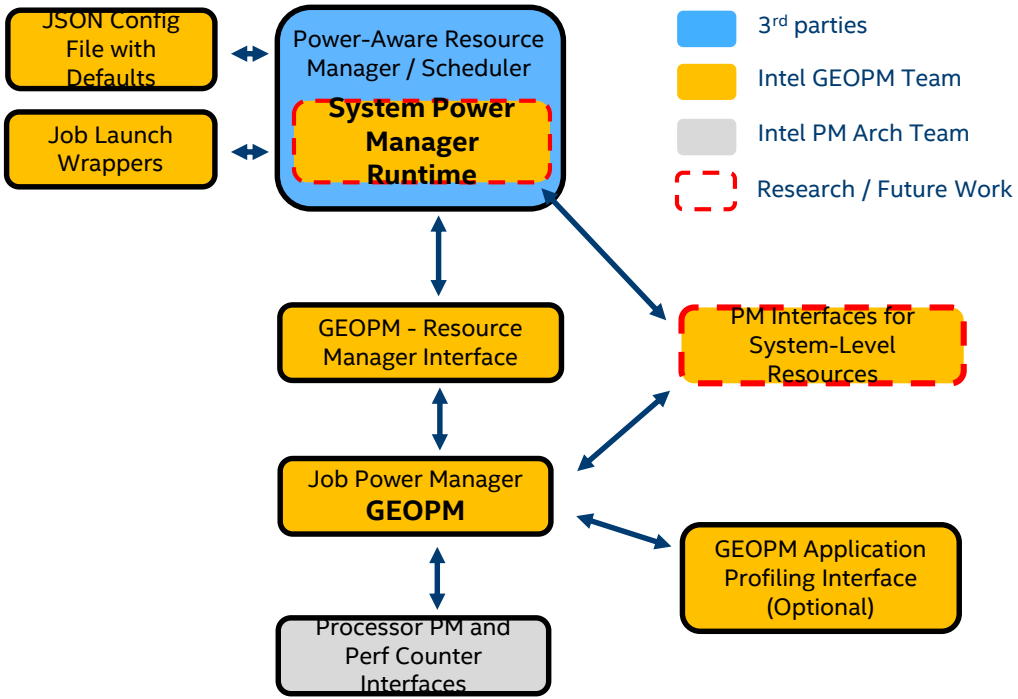
- Introducing GEOPM:
  - Free open source power management runtime and framework
  - Contributed to accelerate community research on power management strategies to overcome Exascale challenges
  - Plug-in architecture for extensibility in two dimensions:
    - control algorithms
    - hardware platform portability
  - Example plug-ins included which significantly improve performance and efficiency via application-awareness

Hierarchical Design and Communications

- Scalable tree-hierarchical design
  - Tree hierarchy of controller agents
  - All agents run in the job compute nodes
  - Each agent runs ctrl algorithm plug-in
  - Recursive control / feedback algorithms
- Flexible tree configuration
  - Tree depth, fan-out, balance, placement optimized via MPI Cartesian grid
  - Tree auto-configured for deployments ranging from Rackscale to Exascale



GEOPM Interfaces and HPC Stack Integration



- Job power manager
  - User-Space Runtime
  - Safe interaction with MSRs via msr-safe (by LLNL)
  - Flexible objective function via plug-ins
  - Globally optimizes HW control knobs across all compute nodes of job (current target: RAPL / DVFS)
- Feedback-guided control system
- Feedback from app / libs via GEOPM APIs
  - OpenMP region detection
  - Automatic detection to be added

GEOPM Project Goals Overview

Managing power:

- Managing power efficiency or max performance under power cap

Managing manufacture variation

- Power / frequency relationship is non-uniform across different chips in the same system

Managing work imbalance:

- Divert power to CPUs with more work

Managing system jitter:

- Divert power to CPUs interrupted or stalled by system noise

Application profiling:

- Report application performance and power metrics

Runtime application tuning:

- Extensible runtime control agent with plug-in architecture

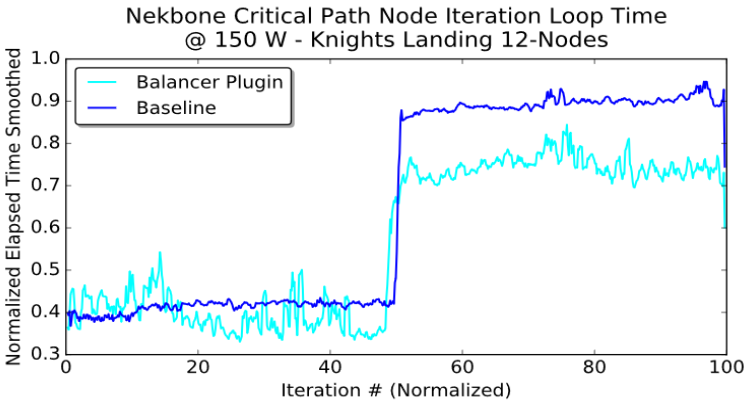
Integration with MPI:

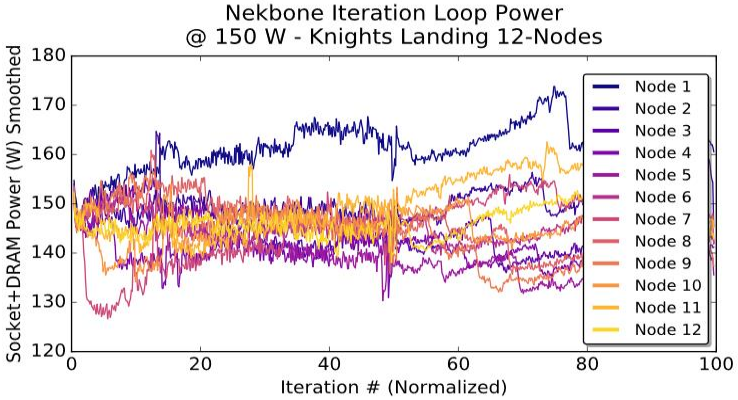
- Automatic integration with MPI runtime through PMPI interface

Integration with OpenMP:

- Automatic integration with OpenMP through OMPT interface

Runtime and Power Allocation Traces

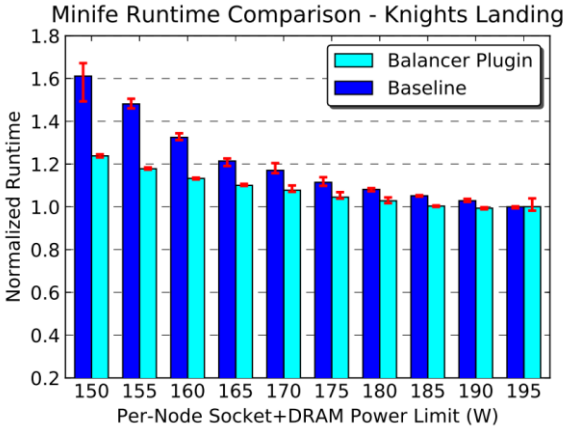


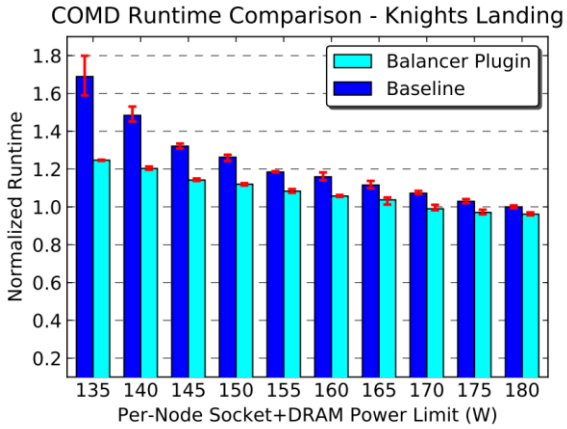


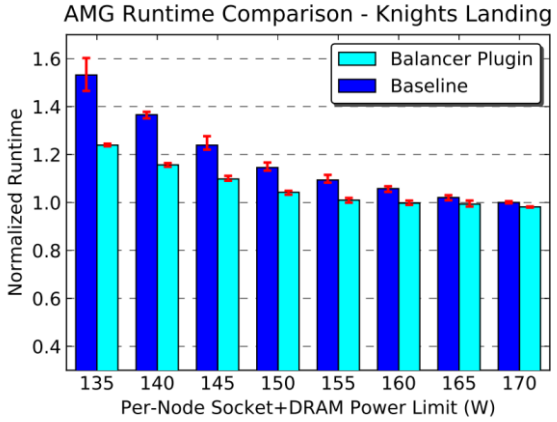
- GEOPM power balancer plug-in speeds up the critical path in Nekbone CORAL workload, by identifying bottlenecks and re-allocating power.
- Nekbone does two CGs with different characteristics leading to re-learning of best power allocation (~iter. #50).

Results: Inter-Node Power Balancing

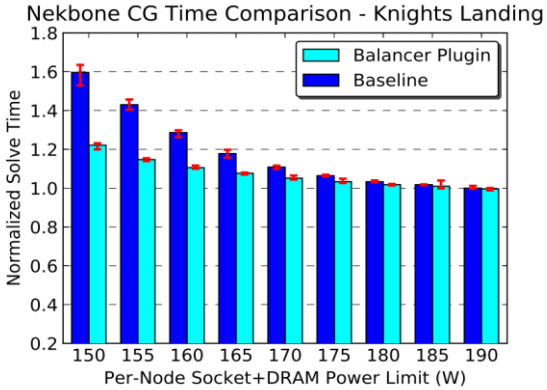
- See GEOPM ISC'17 paper by Eastep et al. for details of experimental setup and further analysis
- Compared overall time-to-solution when capping job power on 12-node KNL cluster with power balancer plug-in vs. static uniform power division (baseline); swept over a range of different job power caps
- Region of interest in job power caps: low-end of job power caps was selected to avoid inefficient clock throttling and the high-end of the job power caps equals the unconstrained power consumption of the workload
- Main result: **up to 30% improvement** in time-to-solution at low end of caps (miniFE, CoMD, AMG), with **up to 9-23% for the rest.** Improvement generally increases as power is more constrained

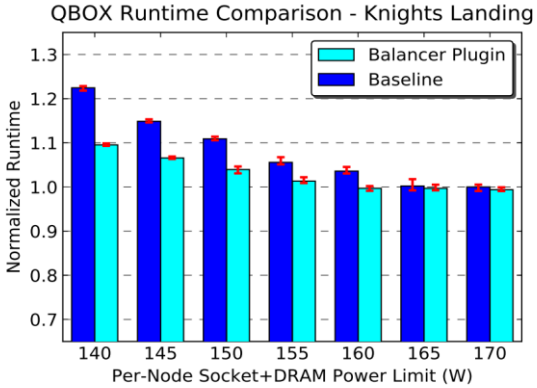


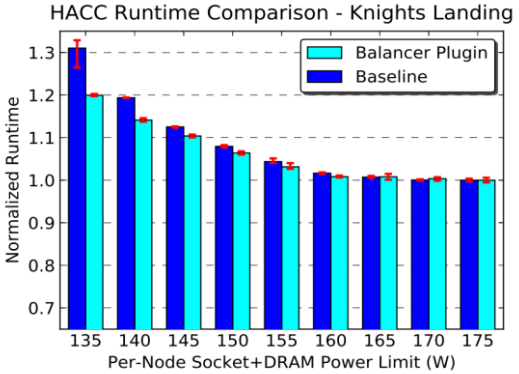


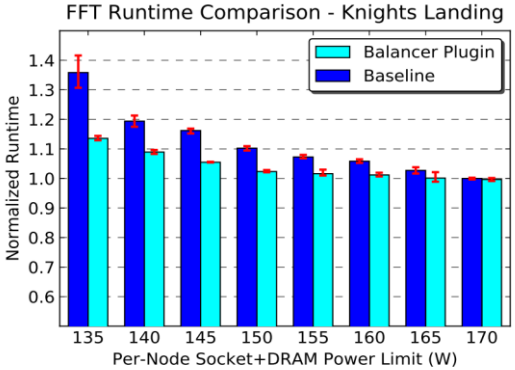


Results: Four Additional Workloads









Deployment Status

- GEOPM is expected to be a general product offering
  - Accepted for inclusion in OpenHPC
- Expecting deployment on CORAL systems at Argonne
  - Additional deployment discussions with LLNL, LANL, Sandia, LRZ
- Basis of Software Development Project within the USDOE Exascale Computing Project (ECP)
  - “A Runtime System for Application-Level Power Steering on Exascale Systems,” in collaboration with LLNL, U. of Arizona and TUM

GEOPM Open Source Community

Institution	Principal Investigator	Project Name	Project Scope	Contribution Type	Time Span	Quality Level	Funded?
Argonne	Ti Leggett Paul Rich Kalyan Kumaran	CORAL -> A21	1. GEOPM 1.0 product development 2. GEOPM >1.0 feature development 3. GEOPM enablement for system power capping + EAS in Cobalt	Sponsor	Q2'15 – Q4'21	Product	Yes
IBM STFC LLNL	Vadim Elisseev Tapasya Patki Aniruddha Marathe		1. GEOPM port to Power8 + NVLink 2. Integrate GEOPM with EAS	Contributor	Q4'16 – TBD	Near-Product	Yes
LLNL Argonne U. Arizona U. of Tokyo	Tapasya Patki Aniruddha Marathe Pete Beckman Dave Lowenthal	ECP PS ECP Argo-GRM	1. Exascale power stack leveraging GEOPM 2. Integrate GEOPM + Caliper framework 3. Integrate GEOPM w/ SLURM power capping and power-aware scheduling extensions 4. Port of GEOPM to non-x86 architectures	Contributor	Q1'17 – Q4'19 SLURM PoC in '18	Near-Product	Yes
LRZ	Herbert Huber Et al.	Super MUC-NG	1. Enhance GEOPM monitoring features 2. Energy optimization plugin for GEOPM 1.0	Contributor	Q3'17 – Q4'20	Product	Yes
Sandia	James Laros Ryan Grant	Power API	1. GEOPM and Power API xface compatibility 2. Power API community WG kickoff at Intel	User	Q4'14 - TBD	Industry Standard	Yes
UniBo CINECA	Andrea Bartolini Carlo Cavazzoni		1. Enhance GEOPM monitoring features 2. Energy optimization plugin for GEOPM 3. Integrate GEOPM + EXAMON 4. Integrate GEOPM w/ SLURM extensions	Contributor	Q2'18 – Q4'19	Near-Product	Yes

Timeline GEOPM Development Progress

Schedule:

A) GEOPM Beta release date on track. Good progress allowed for additional improvements in of monitoring and tracing.

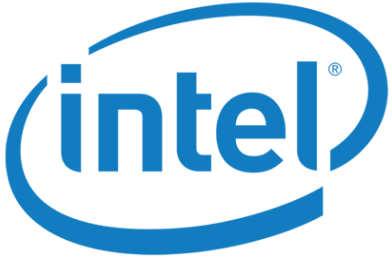
B) GEOPM accepted in OpenHPC, expected to be released ahead of ISC'18

C) GEOPM 1.0 release date on track for SC'18 release date

D) GEOPM tutorial accepted at ISC'18.
Also covering Intel processor controls/monitors

Global Extensible Open Power Manager

http://geopm.github.io/geopm



- Open source runtime for power management and framework for HPC community collaboration. (BSD-3 license)
- Scalable, extensible through plugins!
- Contribute, use & adapt for your HPC center / users / research groups
- Everything you need to get started: http://geopm.github.io/geopm

Matthias Maiterth
[matthias.maiterth@intel.com]