System Software Research for Extreme-Scale Computing

Approved for Public Release

CScADS Workshop
July 19-22, 2010

James H. Laros III
Sandia National Labs
Team Members

- James Laros
- Sue Kelly
- Ron Oldfield
- Ron Brightwell
- Kevin Pedretti
- Kurt Ferreira
- Rolf Riesen
- Todd Kordenbrock
Outline of Our Plans for the INCITE Allocation

INCITE provides platforms necessary to continue research in system software

- Research Activities Briefly
  - Lightweight Kernel OS and Virtualization
    - Kitten (Sandia) and Palacios VMM (from North Western University)
      - Less than 5% performance impact on applications
  - Resilience
    - Redundant MPI
      - At very large scale reduces runtime and total resource usages
  - Scalable I/O
    - Leverage available compute/service node resources for I/O caching and data processing
  - Power Efficiency and Utilization
    - Goal: Reduce power use while maintaining performance
  - Debugging
    - Fast debugging capability for light-weight kernels
Motivation

- Power is one of or the most important considerations in fielding current and next generation HPC systems.
- HPC application power use and factors impacting this use are not well studied.
- Power saving techniques used in commodity operating systems will greatly impact HPC application performance.

Modifications to RAS and Catamount to support power savings

- RAS
  - Added instrumentation and collection capabilities to RAS
- Catamount
  - Power savings during OS idle, per core
  - OS-level frequency scaling capability
  - User space library interface to frequency scaling
  - MPI profiling layer instrumentation
Power Frequency and Analysis
Phase 1

Based on previous power analysis studies

- Laros et.al. "Topics on Measuring Real Power Usage on High Performance Computing Platforms"

Analyze performance vs. power efficiency (at scale)

- **STATIC** frequency modification during application run-time.
- Procedure
  - Execute application suite using a range of Pstates defining both frequency and input voltage of CPU.
  - Collect power usage during runs and analyze total energy use vs. application run-time
  - Our early results show a very favorable trade-off!!
Power Frequency Analysis: LAAMPs
Small scale results of multiple LAAMPs runs
Analyze performance vs. power efficiency (at scale)

- **DYNAMIC** frequency modification during application run-time
- **DYNAMIC** frequency modification defined as deterministic frequency change driven by application characteristics. Pstate change during MPI barrier for example

Phase 3 testing, if necessary, will be based on Phase 1 and 2 analysis

Additionally, power data will be collected during a range of other systems software testing accomplished as part of this overall project
Additional Information

For information about the other research topics mentioned see: