How Much of a Role Software Plays in Energy Efficiency?

Panel at E2SC ’15
by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda
Panel Questions and Answers

• 1) How is system software expected to contribute to energy efficiency optimization?
  – Significantly

• 2) What are the expectations for introspective systems? Will they become a reality within the next decade?
  – Quite high, Black box approach may not be the best
  – Yes

• 3) What role will application software have? What are reasonable interfaces between the application and system software?
  – Significant role
  – Varied levels of interfaces are possible - MPI-T is one option
Reducing Power Consumption

• Power consumption is dominated by
  – CPU usage
    • Computation
  – Data movement
    • Memory access
    • Communication protocol
    • Network-level data transfer, etc.

• Two major approaches considered in the past to reduce power consumption in data movement
  – Determining slacks and use DVFS schemes
  – Scheduler-based solutions to have good power-performance trade-offs at the systems-level

• Runtime-level and Network-level power-optimizations have not been addressed
How to Reduce Power Consumption within a Runtime in an Intelligent Manner?

• MPI is a common runtime

• Primarily polling-driven for performance
  – Leads to increased power consumption

• Many times data has not reached the receivers and receivers are busy polling leading to un-necessary power consumption

• Can we have intelligent designs within runtime to avoid such un-necessary power consumption?

• Can we provide flexibility to have good power-performance trade-off?
MVAPICH2-EA: Application Oblivious Energy-Aware MPI (EAM)

- An energy efficient runtime that provides energy savings without application knowledge
- A white-box approach
- Automatically and transparently use the best energy lever
- Provides guarantees on maximum degradation with 5-41% savings at <= 5% degradation
- Pessimistic MPI applies energy reduction lever to each MPI call

---

Energy-Aware MVAPICH2 Library and OSU Energy Management Tool (OEMT)

- MVAPICH2-EA (Energy-Aware) MPI Library
  - Production-ready Energy-Aware MPI Library
  - New Energy-Efficient communication protocols for pt-pt and collective operations
  - Intelligently apply the appropriate energy saving techniques
  - Application oblivious energy saving
  - Released 08/28/15

- OEMT
  - A library utility to measure energy consumption for MPI applications
  - Works with all MPI runtimes
  - PRELOAD option for precompiled applications
  - Does not require ROOT permission:
    - A safe kernel module to read only a subset of MSRs

- Available from: [http://mvapich.cse.ohio-state.edu](http://mvapich.cse.ohio-state.edu)
Reducing Power with Intelligent Protocol Selection?

- Transport protocol and energy-aware designs for blocking All-to-all collectives for IB networks
- Identify the correct set of transport protocols and algorithms that lead to best energy savings for different All-to-all communication patterns (640 processes)

![Performance Chart]

**Performance**
- RC Protocol
  - Best performance at low to medium network load
  - Performance degrades as network load increases
  - Choose for applications / communication patterns with low to medium network load
- DC Protocol
  - Inherent serialization in DC causes
  - Performance overhead at low to medium network load
  - Good network behavior at high network load
  - Choose for applications / communication patterns with high network load
- RDMA-Power-Aware (R-P-Aware) + DC-E-UD
  - Significant energy savings
  - Able to save 1.7x (44%) energy
  - Improves communication performance
  - 8% improvement in latency
  - Significant reduction in network congestion
  - 8.8 times reduction in congestion


E2SC-Panel (SC’15)
Reducing Power Consumption at the Network-Level?

• In current HPC systems, networks are always `on’ irrespective of data transmission is taking place or not
  – Switches, Adapters and Links consume power all the time

• Can we turn off unused links, adapters and switch ports?

• Can low-overhead schemes (like DVFS for processors) be designed for turning off/turning on network components?

• Such schemes can be used by the upper-level runtime and middleware to reduce network-level power consumption further ...