Quantifying the Dominance of Leakage Energy in Large-Scale System Caches

Aditya Deshpande, Jeffrey Draper {adeshpan. draper} @ ISI.EDU
Marina Group :: Information Sciences Institute :: Ming Hsieh Department of Electrical Engineering

Motivation

- Energy consumption is a big concern for VLSI designers at both extremes of computing —
  - ⇒ Embedded (Battery life limitation)
  - ⇒ Exa-scale (Recurring electricity costs)
- Prior research has extensively focused on reducing energy in the core, but little work done to address other components
- **On-chip caches** make an ideal candidate for energy studies from **S**ystem, **A**rchitecture and **T**echnology standpoint —

S
- Large number of processing cores
- Increased parallelization in application code (Sync. leads to increased aggregated runtime across all the cores)

A
- Trend of having large on-chip caches in a deep hierarchy to increase performance of multi-core systems

T
- Leakage in transistors increasing with every generation of transistor scaling
- Caches occupy upto 75% of chip area & upto 50% total chip power

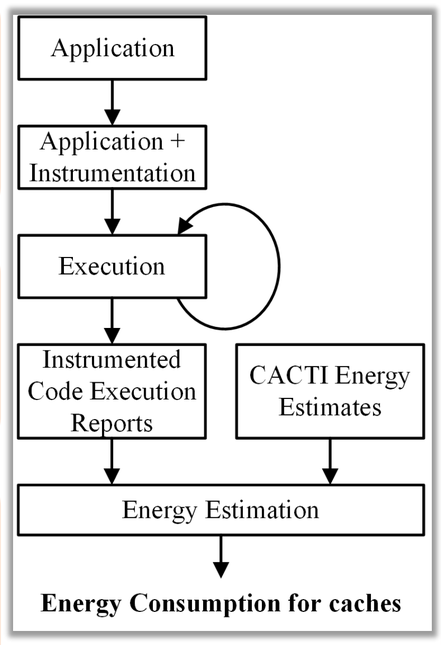
- We focus on the energy consumption in the memory sub-system during an application execution (in particular on the on-chip caches)

Our Modeling Scheme

Architecture
- Experiment conducted on Hopper, a Cray XE6 machine located at NERSC
- Compute node uses AMD processors at 45nm process technology

Benchmarks
- NAS Parallel Benchmarks (NPB)
- Mantevo Miniapps /Real World Apps
- SCALAPACK

Cache
- L1 D/I cache: A=2, B=64B, C=64KB
- L2 Unified 512KB, L3 shared 12MB
- L1 modeled as fast cache, 350K temperature, 45nm process in CACTI

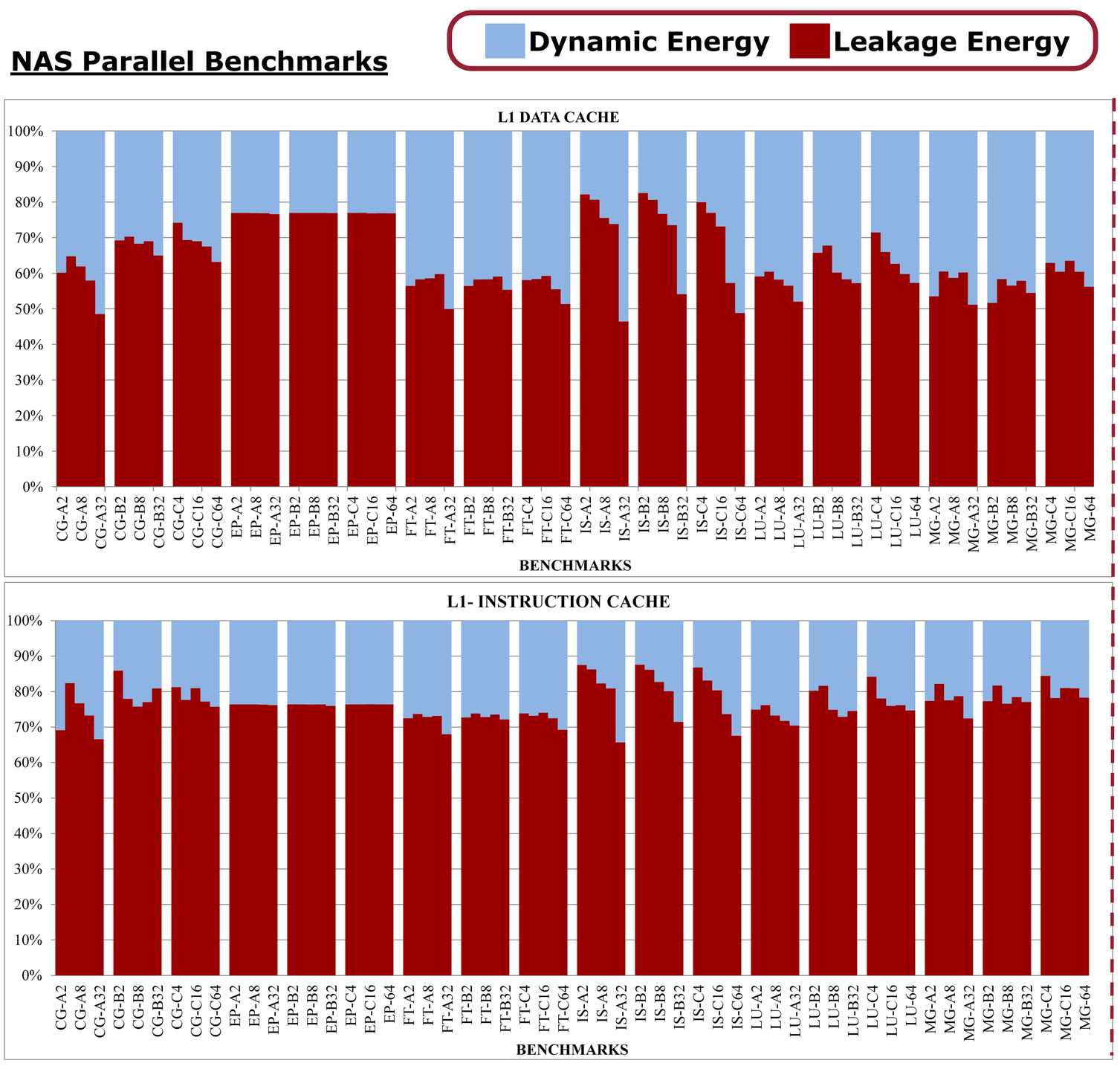


Our scheme combines cache activity information with a cache-energy modeling tool to generate fast and accurate energy consumption information for various levels of caches

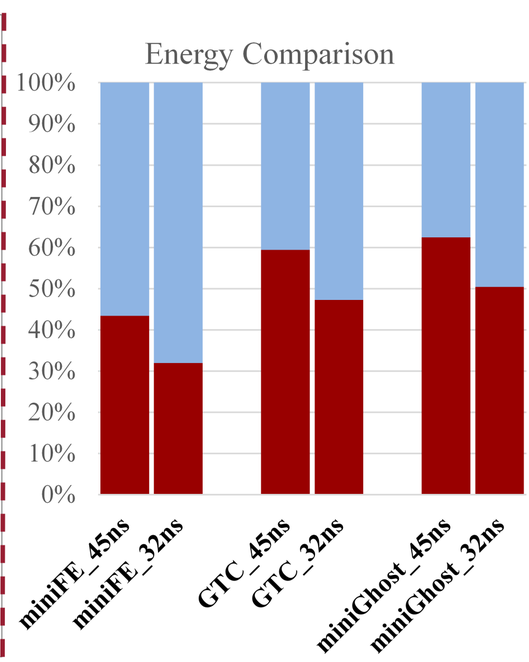
We compute dynamic and leakage energy dissipated in various levels of on-chip caches for various applications. Here, results for the L1 cache energy consumption across a wide range of HPC applications is presented

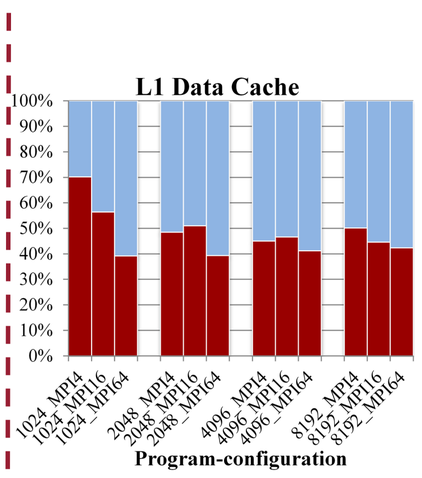
Results

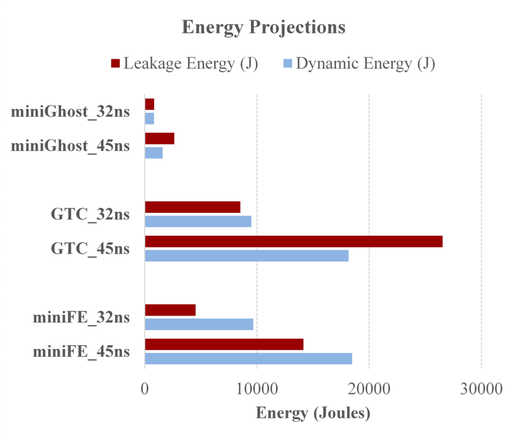
NAS Parallel Benchmarks

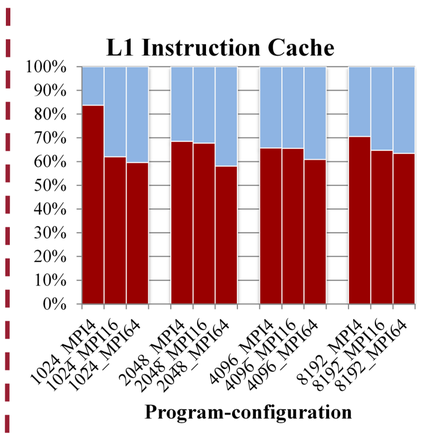


Miniapps/Real world Apps/ SCALAPACK









Observation

- Leakage energy percentage varies from 40-80% for D-cache and 60-80% for I-cache
- As MPI nodes increase, more cache accesses occur increasing total energy
- Even though leakage energy percentage reduces as more nodes are used, the total energy in caches increases significantly with increase in nodes
- With the use of high-K metal gate devices, leakage is still between 30-50% for real-world applications— a worrying sign!

Conclusions

- Cache utilization has strong impact on leakage energy component of total cache energy
- Our results show leakage is the dominant form of energy dissipation in smaller L1 caches and expected to be more dominant for L2/L3 caches

Trade-offs

- **Performance Improvement** — Increasing nodes for computation (requires more energy) OR optimal use of resources (requires more effort)
- **Cache Leakage Energy** — Magnitude of performance improvement AND reduction in execution time