

Energy-Aware Computing Systems

Energiebewusste Rechensysteme

VII. Cluster Systems

Timo Hönig

2018-11-28



Preface

Terminology

Composition and Strategies

- Compound Structure

- Provisioning and Load Control

Cluster Systems

- Energy Proportionality

- Energy-efficient Cluster Architecture

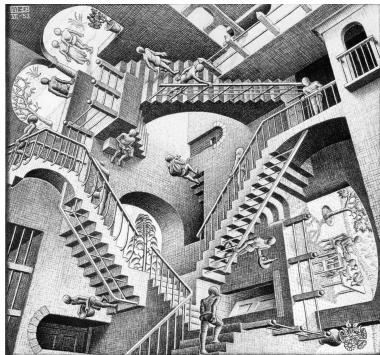
- Thermal Awareness and Control

Summary



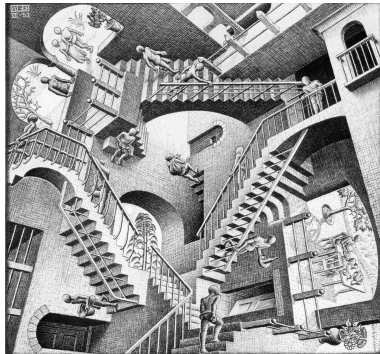
Preface: Changing the Perspective

- **small individual problems** that are processed to jointly provide a overall solution
 - deeply embedded systems, wireless sensor nodes in cyber-physical systems
 - **bottom up** approach: build (nested) control loops with self-contained solo systems
 - **heterogeneous tasks** across concerned systems



Preface: Changing the Perspective

- **large problems that** are split down to small problems, that contribute to a overall solution
 - clustered networked systems in a compound structure with manageable dynamicity
 - **top down** approach: divide and conquer; consider local and global energy demand
 - **homogeneous (sub-)tasks** across concerned systems



Abstract Concept: Cluster Systems

■ cluster systems

- a number of things of the same kind, growing or held together
- a bunch
- **swarm**
 - old English *swearm*
 - multitude, cluster
- cluster **composition**
 - **heterogeneous** nodes
 - **homogeneous** nodes
- cluster **linkage**
 - **wired** links
 - **wireless** links

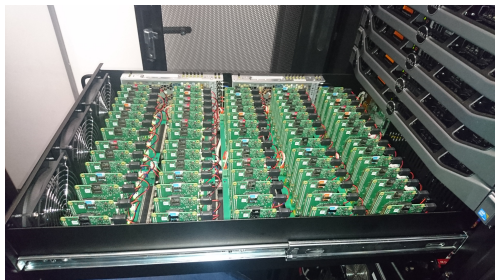


nationalgeographic.org/



■ cluster systems

- energy-efficient cluster architecture with homogeneous **low-power nodes**
- cheap hardware...
...but sensitive to errors
- RPi cluster
 - 1 350 systems
 - 5 400 cores
 - < 4 kW (idle)
 - > 13 kW (active)
 - small area requirements



■ cluster systems

- energy-efficient cluster architecture with homogeneous **high-performance nodes**
- powerful hardware...
...with complex wiring and administration
- mining cluster
 - energy-efficient special purpose hardware (e.g., GPUs)
 - yet, large clusters have an energy demand that exceeds the one of entire cities



■ cluster systems

- energy-efficient cluster architecture with heterogeneous **low-power and high-performance nodes**
- heterogeneous hardware components...
...enable an appropriate mapping of software requirements to hardware offerings
- mixed cluster
 - address heterogeneity of software requirements
 - highly dynamic → power and energy proportionality

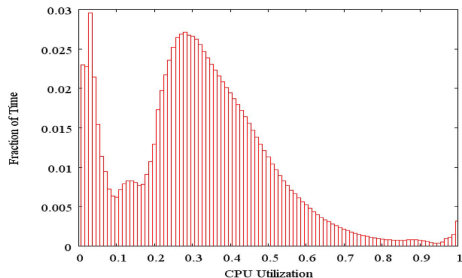


- provisioning and load control at level of the **system software**
- **workload distribution** [4]
 - software characterization → (available) hardware components
 - node assignment strategies → avoid under- and overload
- **scheduling**
 - thermal-awareness [2] → cluster locality and deferred execution
 - exploit parallelism where possible
- **distributed run-time power management**
 - cluster power cap [5]
 - steer progress speed of distributed tasks



Energy Proportionality

- considerations on **warehouse-scale computers**
 - the datacenter as a computer
 - provisioning of hardware components → impact on cost efficiency
 - operation of hardware components → impact on cost efficiency, too
- **utilization/workload vs. power demand**
 - depending on the workload of systems, the power demand must scale
 - best case: no power when idle → reasoning between blocking and non-blocking energy management control methods

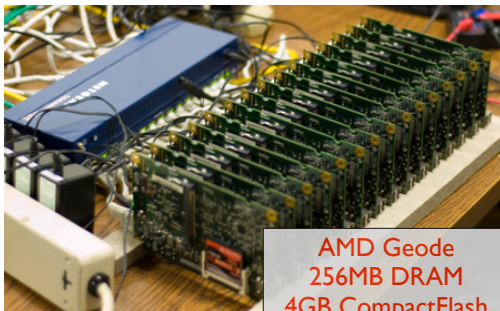


[3]
U. Hölzle and LA. Barroso:
The Case for
Energy-Proportional Computing (2007)

Energy-efficient Cluster Architecture

- David G. Andersen et al.: **fast array of wimpy nodes** (FAWN) [1]
 - cluster architecture that is composed of homogeneous low-power nodes („wimpy nodes”)
 - FAWN nodes and cluster have drastically different characteristics compared to server systems that employ so-called „beefy nodes”

FAWN



AMD Geode
256MB DRAM
4GB CompactFlash



Energy-efficient Cluster Architecture

- David G. Andersen et al.: **fast array of wimpy nodes (FAWN)** [1]
 - goal: efficient execution of I/O bound, computationally light workloads
 - multi-layered architecture: frontend node passes requests to responsible backend nodes → identified by hashes
 - joint **hardware/software architecture**
 - custom key-value store
 - low memory nodes
 - partitioning

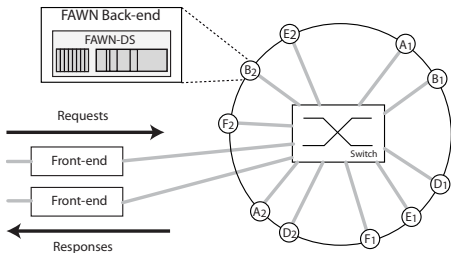


Figure 1: FAWN-KV Architecture.

- Jeonghwan Choi et al.: thermal-aware task scheduling [2]
 - goal: **hot spot mitigation** to reduce thermal stress
 - avoid performance loss as to overheating
 - reduce cooling efforts
 - **core hopping** vs. task deferral
 - spatial hot spot mitigation
 - temporal mitigation of overheating

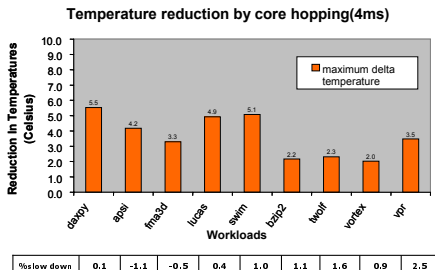
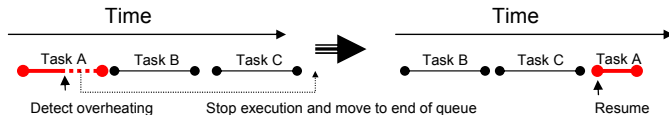


Figure 1: Core hopping reduces on-chip temperatures with small performance impact

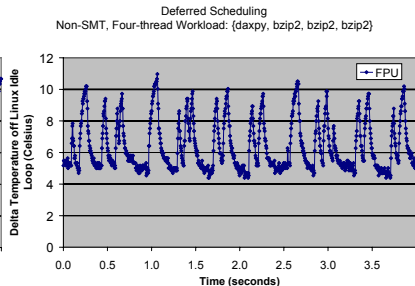
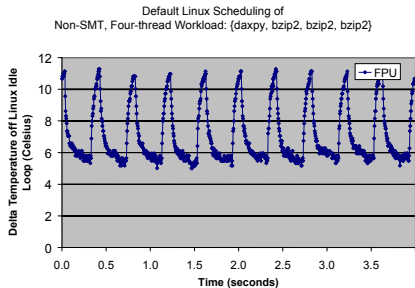
Thermal Awareness and Control

- Jeonghwan Choi et al.: thermal-aware task scheduling [2]
 - task deferral
 - reschedule hot-running tasks to be last in run queue
 - cool down ahead of (resumed) execution



Thermal Awareness and Control

- Jeonghwan Choi et al.: thermal-aware task scheduling [2]
 - task deferral
 - reschedule hot-running tasks to be last in run queue
 - cool down ahead of (resumed) execution



■ **cluster systems**

- compound systems consisting large number of nodes
- suitable mapping of **software requirements** to **hardware offerings**

■ **energy demand** depends on **system software**

- workload distribution and node assignment
- scheduling
- run-time controls (i.e. distributed power management)

■ **power and energy proportionality**

- as to varying workloads, power demand must scale
- consider blocking and non-blocking energy management methods



- hands-on:
power proportionality of a system with an Intel Core i5-8259U CPU
 - quad-core processor with hyper-threading
 - CPU clock up to 3.80 GHz
 - 2.3 GHz (base)
 - 3.6 GHz (turbo, 4 cores)
 - 3.8 GHz (turbo, ≤ 2 cores)
 - 6 MByte Cache, 16 GByte Memory
- paper discussion
 - ▶ Andrew Krioukov et al.
NapSAC: Design and Implementation of a Power-Proportional Web Cluster
Proceedings of the Workshop on Green Networking (GreenNet'10), 2010.



- **cluster systems** consist of **homogeneous** or **heterogeneous** nodes that cooperatively work on a solution for a large problem (e.g., scientific computing, number crunching)
- consider **overall** energy demand at cluster and **local** energy demand at node level to improve **energy proportionality**
- reading list for Lecture 8:
 - ▶ Rolf Neugebauer and Derek McAuley
Energy is just another resource: Energy accounting and energy pricing in the Nemesis OS
Proceedings of the 8th Workshop on Hot Topics in Operating Systems (HotOS'01), 2001.



Reference List I

- [1] ANDERSEN, D. G. ; FRANKLIN, J. ; KAMINSKY, M. ; PHANISHAYEE, A. ; TAN, L. ; VASUDEVAN, V. :
FAWN: A Fast Array of Wimpy Nodes.
In: *Proceedings of the 22nd ACM SIGOPS Symposium on Operating Systems Principles*, 2009, S. 1–14
- [2] CHOI, J. ; CHER, C.-Y. ; FRANKE, H. ; HAMANN, H. ; WEGER, A. ; BOSE, P. :
Thermal-aware Task Scheduling at the System Software Level.
In: *Proceedings of the 2007 International Symposium on Low Power Electronics and Design (ISLPED'07)*, 2007, S. 213–218
- [3] HÖLZLE, U. ; BARROSO, L. A.:
The Case for Energy-Proportional Computing.
In: *Computer* 40 (2007), 12, S. 33–37
- [4] SRIKANTIAH, S. ; KANSAL, A. ; ZHAO, F. :
Energy Aware Consolidation for Cloud Computing.
In: *Proceedings of the 2008 Workshop on Power Aware Computing and Systems (HotPower'08)*, 2008



- [5] ZHANG, H. ; HOFFMANN, H. :
Performance & Energy Tradeoffs for Dependent Distributed Applications Under System-wide Power Caps.
In: *Proceedings of the 47th International Conference on Parallel Processing (ICPP'18)*, 2018

