Energy-Aware Computing Systems

Energiebewusste Rechensysteme

VII. Cluster Systems

Timo Hönig

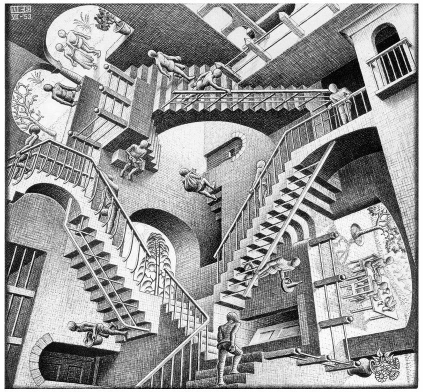
2019-06-13

EASY

Agenda

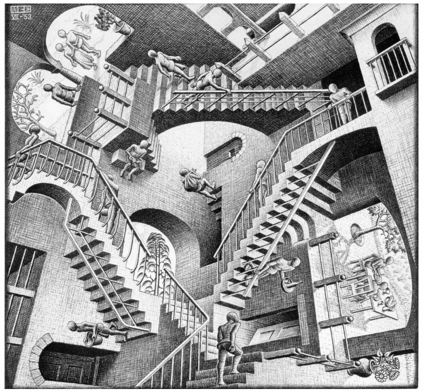
Preface: Changing the Perspective

- **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



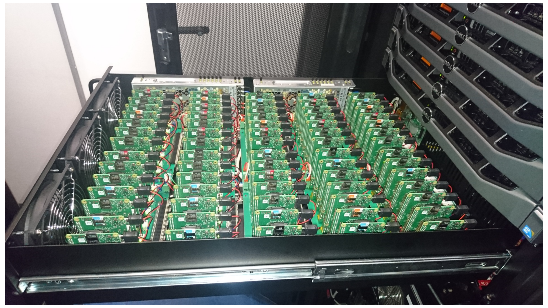
Abstract Concept: Cluster Systems

- **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/

Compound Structure

- **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

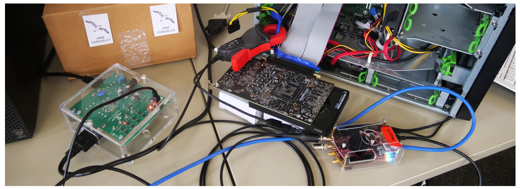


Compound Structure

- **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

Compound Structure

- **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 $\rightarrow$ power and energy proportionality

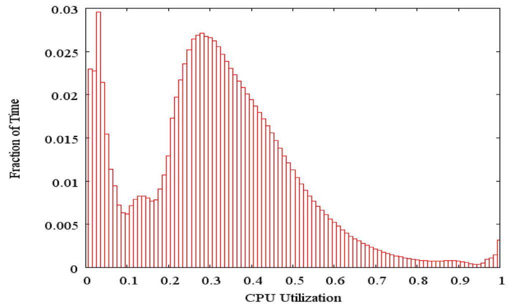


Provisioning and Load Control

- 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

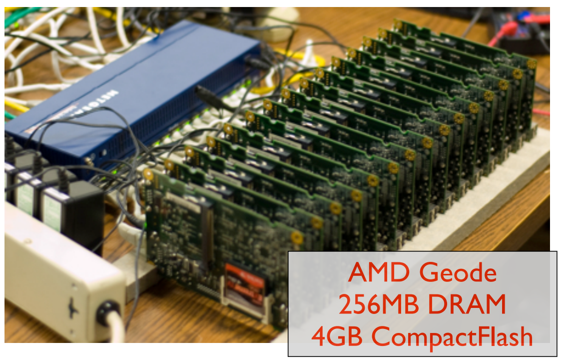


[3]
U. Hölzle and LA. Barroso:
The Case for
Energy-Proportional Computing (2007)

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



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

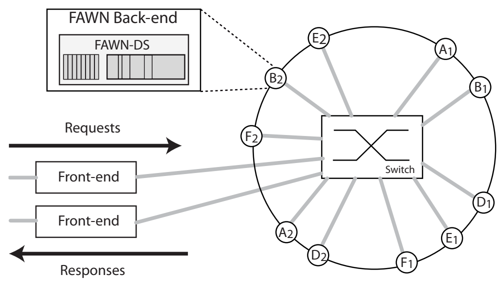
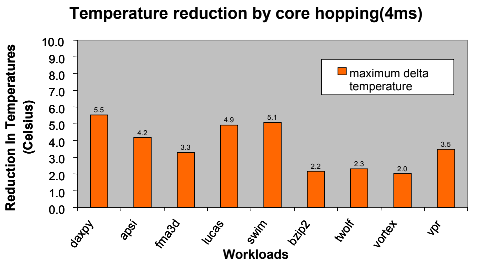


Figure 1: FAWN-KV Architecture.

Thermal Awareness and Control

- 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

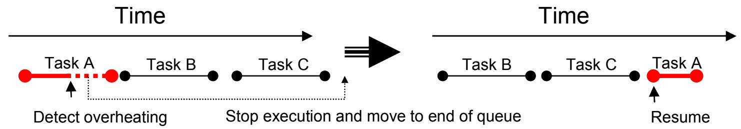


%slow down	0.1	-1.1	-0.5	0.4	1.0	1.1	1.6	0.9	2.5

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

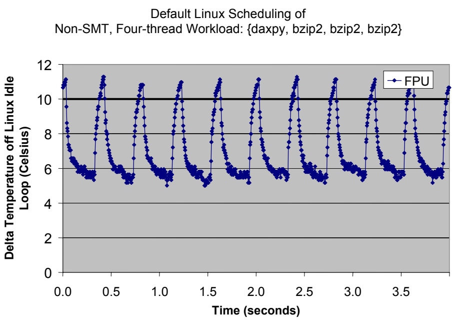
Thermal Awareness and Control

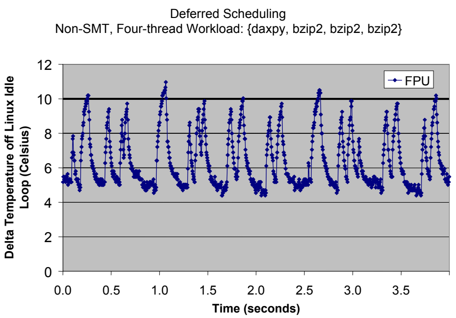
- Jeonghwan Choi et al.: thermal-aware task scheduling [2]
  - task deferral
    - reschedule hot-running tasks to be last in run queue
    - cool down ahead of (resumed) execution



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





Considerations and Caveats

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

Paper Discussion

- 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.

Subject Matter

- **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] Srikantaiah, 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

Reference List II

[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