On Energy Proportionality and Time-energy Performance of Heterogeneous Clusters

Lavanya Ramapantulu, Dumitrel Loghin and Yong Meng Teo
Department of Computer Science
National University of Singapore
13 Computing Drive, Singapore, 117417
Email: {lavanya,dumitrel,teoym}@comp.nus.edu.sg

***Abstract*—Energy efficiency is an area of growing importance in datacenters. While the energy proportionality wall poses a challenge to further improve the dynamic power range of servers, heterogeneous systems offer a new opportunity to achieve higher energy efficiency by improving the match between application workload demands on the large heterogeneous system configuration space. This paper proposes a model-driven energy proportionality analysis of heterogeneous clusters consisting of server nodes with different performance-to-power ratio (PPR). Our analysis shows that inter-node heterogeneity has a positive effect of scaling the energy proportionality wall by exposing configurations with sub-linear energy proportionality. Secondly, analysis of these sub-linear configurations on the 95th percentile response time shows that heterogeneity is beneficial in workloads where the PPR of wimpy nodes is higher than brawny nodes.**

I. Introduction

Energy proportionality was proposed as an important server design metric to address the mismatch between the amount of useful work performed and the energy consumed [10]. Energy proportionality has stalled at 80% for individual servers. This stall is referred to as the energy proportionality wall, and is attributed to the lack of improvements in the dynamic power range of servers [44].

While the energy proportionality wall leads to the introduction of different low-power modes, these modes such as sleep or shutdown are not viable options due to (i) longer response time during traffic spikes and (ii) the necessity to execute many background tasks in typical datacenters [26], [10]. Therefore, research directions in the area of active low-power modes have been explored, where the server continues to perform some amount of useful work in a low-power state. While Somniloquy [4] and barely alive servers [5] propose servers performing only I/O operations in a low-power state, KnighShift uses a low-power processor (knight) at lower utilization levels. However, all of these works explore energy efficiency at an individual server level. Complementing these techniques, we analyze heterogeneous mixes of nodes with diverse performance-to-power ratios to understand the impact of *heterogeneity on cluster-wide energy proportionality*.

Energy efficiency techniques in the parallel computing landscape mainly focus on clusters with traditional high-performance multi-core systems such as x86/64 Intel/AMD processor nodes, also known as brawny nodes [30], [39], [40]. With energy efficiency increasingly becoming a concern, the use of low-power (wimpy) processors as an alternative in designing clusters has been explored [38], [6], [18]. While these works focus on using either high-performance or low-power nodes, this paper analyzes the energy proportionality of clusters having a mix of both wimpy and brawny nodes.

Heterogeneous clusters have many challenges including selecting a good system configuration with respect to the number of nodes of each type and dynamic adaptation of the workload demands to available heterogeneous resources. This paper addresses the first challenge and determines a static mapping of the application to a system configuration. Dynamic adaptation of workload during the execution of a program complements our approach and can be used in conjunction with the proposed approach.

For a given application with a time deadline and energy budget, it is non-trivial to determine an energy-proportional configuration among the large system configuration space due to heterogeneity [31]. To analyze the energy proportionality of such a mix of high-performance and low-power nodes, we use an energy model to determine the energy proportionality of heterogeneous clusters. Based on this model, we first analyze single-node energy proportionality of both wimpy and brawny nodes. We compare these nodes across a range of metrics and show that while wimpy nodes are more power-efficient and have a better performance-to-power ratio (PPR), they are less energy-proportional than brawny nodes. Using this analysis, we show that energy proportionality need not necessarily imply energy efficiency, specifically when comparing nodes with diverse peak power usage.

Secondly, we analyze the cluster-wide energy proportionality of low-power and high-performance homogeneous clusters and compare with heterogeneous clusters having a mix of both wimpy and brawny nodes. This comparison among the clusters again exposes the point that while traditionally energy proportionality implies energy efficiency, this implication might not hold when comparing systems with a diverse power ratio among them. While energy proportionality advocates the usage of brawny nodes, PPR metric using the throughput per watt shows wimpy nodes to be more energy efficient.

Our previous work [31] showed that heterogeneity introduces a "sweet region" consisting of a set of system configurations that meet a given execution time deadline with minimum energy, thus forming an energy-deadline Pareto frontier. It is

2168-9253/16 $31.00 © 2016 IEEE
DOI 10.1109/CLUSTER.2016.44

imperative to understand the energy proportionality implications of these heterogeneous configurations on the sweet spot region. To this end, we show how heterogeneity has a positive effect of scaling the energy proportionality wall by exposing configurations with *sub-linear proportionality*. Furthermore, we analyze the impact of a given execution time deadline on these sub-linearly proportional configurations and show that they have minimal implications on the 95th percentile response time. Such an analysis provides useful insights to lower energy usage while still meeting a given execution time deadline target, by using energy efficient system configurations, to reduce overall system inefficiencies.

Our approach is applied to analyze the energy proportionality for a range of typical datacenter applications such as *memcached* key-value store used by web applications , H.264 encoding used in multimedia streaming, the financial analytics program *blackscholes* from PARSEC, the real-time speech recognition engine *Julius*, and the openssl implementation of the *RSA-2048 key verification* step of the TLS/SSL encryption mechanism. The time and energy models used in our approach have been validated across all these workloads and across different heterogeneous system configurations.

Traditionally, energy proportionality has been perceived as synonymous to energy efficiency [10], [33], [43]. With the emergence of several metrics to quantify energy proportionality, we review these metrics and answer the question: Can energy-proportionality using these metrics alone translate into cluster-wide energy efficiency, specifically for clusters with a heterogeneous mix of wimpy and brawny nodes. Furthermore, this paper tackles implications of heterogeneity on the ideal energy proportionality curve by making the following key contributions:

1) we define a measurement-driven model to determine the energy proportionality of clusters with brawny and wimpy nodes
2) we show that inter-node heterogeneity has a positive effect of scaling the energy proportionality wall by enabling sub-linear configurations
3) we show that these sub-linear configurations have minimal impact on the 95th percentile response time

The rest of the paper is organized as follows. Section II discusses our methodology and we present the energy proportionality analysis in Section III. We discuss the related work in Section IV and summarize in Section V.

II. Methodology

A. Overview

With heterogeneity becoming ubiquitous in datacenters [29], [32], [12], it offers a new opportunity to obtain a better resource match between application demands and the heterogeneous platforms. However, the large configuration space due to heterogeneity poses a challenge to both datacenter architects and users of such heterogeneous clusters to choose the right system configuration for executing a workload. Our previous work [31] proposed a methodology to address this challenge by providing a technique to determine the set of Pareto-optimal system configurations to execute a program. Such a configuration is defined using a set of tuples consisting of the types of nodes, number of nodes for each type, the active cores per node and the operating core clock frequency. While

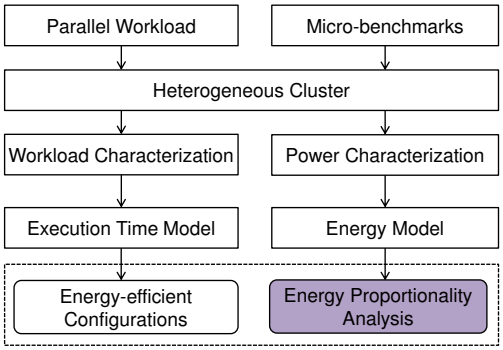


Figure 1: Methodology overview

our previous work [31] models the execution time and energy using a measurement-driven modeling approach, those models did not support energy proportionality. In our previous work we derived the time energy models shown in Figure 1. To understand the work presented in this paper, we summarize the definitions of the model parameters and the time energy model in Table 1 and Table 2 respectively.

Table 1: Model parameters

Symbol	Description
	Workload Parameters
$\mathcal{P}$	program
$\mathcal{P}_s$	program $\mathcal{P}$ with smaller input size
$\lambda_{I/O}$	I/O requests inter-arrival rate
	System Parameters
d_{max}	maximum degree of inter-node heterogeneity of the system
n_{max}	maximum number of nodes of type i
c_{max}	maximum number of cores for nodes of type i
f_{max}	maximum core clock frequency for nodes of type i
U	average system utilization during time period T
	Time Model
n	number of nodes
c	number of active cores per node
f	operating core clock frequency
T_{CPU}	total CPU response time for $\mathcal{P}$
T_{core}	total core response time for $\mathcal{P}$
T_{mem}	total memory response time for $\mathcal{P}$
$T_{I/O}$	total I/O response time for $\mathcal{P}$
T_{I/O_T}	total I/O transfer time for $\mathcal{P}$
$T_{\mathcal{P}}$	total execution time of program $\mathcal{P}$
	Power Parameters[W]
$P_{CPU,act}$	CPU power when executing work cycles
$P_{CPU,stall}$	CPU power when memory-related stalls
P_{mem}	power consumed by memory operations
P_{net}	power consumed by network card
$P_{sys,idle}$	power consumed by idle system
	Energy Model[J]
$E_{CPU,act}$	total energy consumed when CPU is active
$E_{CPU,stall}$	total energy consumed when CPU is stalling
E_{mem}	total energy consumed by memory sub-system
E_{net}	total energy consumed by network sub-system
E_{idle}	total energy consumed by idle system
$E_{\mathcal{P}}$	total energy consumed by a program $\mathcal{P}$
E	total energy consumed during time period T

Table 2: Summary of time-energy model

Time Performance	
$T_{\mathcal{P}}$	$\max_{i=1}^{d_{max}}(T_i)$
T_i	$max(T_{i,CPU}, T_{i,I/O})$
$T_{i,CPU}$	$max(T_{i,core}, T_{i,mem})$
$T_{i,core}$	$\frac{cycles_{i,core}}{f_i}$
$T_{i,mem}$	$\frac{cycles_{i,mem}}{f_i}$
$T_{i,I/O}$	$\frac{\max(T_{i,I/O_T}, \frac{1}{\lambda_{I/O}})}{n_i}$
Energy Performance	
$E_{\mathcal{P}}$	$\sum_{i=1}^{n} E_i$
E_i	$(E_{i,CPU} + E_{i,mem} + E_{i,I/O} + E_{i,idle}) \cdot n_i$
$E_{i,CPU}$	$(P_{i,CPU,act} \cdot T_{i,act}) + (P_{i,CPU,stall} \cdot T_{i,stall})$
$E_{i,mem}$	$P_{i,mem} \cdot T_{i,mem}$
$E_{i,I/O}$	$T_{i,I/O} \cdot P_{i,I/O}$
$E_{i,idle}$	$T_i \cdot P_{i,idle}$

B. Energy Proportionality Extensions

As outlined in Figure 1, we extend the time and energy models to determine the energy proportionality metrics: Dynamic Power Range (DPR), Idle-to-peak Power Ratio (IPR), Energy Proportionality Metric (EPM) and Linear Deviation Ratio (LDR). Using these modeled values, we analyze the energy proportionality of both individual servers and clusters and determine whether inter-node heterogeneity aids in scaling the energy proportionality wall.

During execution, a processor consumes varying amount of power depending on the number of active components. CPU active power, $P_{CPU,act}$, is measured across cores and frequencies for each type of node, using a micro-benchmark that maximizes the CPU utilization. Power incurred by CPU stall cycles, $P_{CPU,stall}$, is measured using a micro-benchmark that generates a stream of cache misses to maximize the number of stall cycles. Power used by active memory, P_{mem} is derived from specifications [1], [23]. Networking I/O power, $P_{I/O}$, is obtained through direct measurement when the NIC is used and the idle system power, P_{idle}, is measured without any workload. It suffices to do the measurements on a single node of each type because all the nodes of the same type exhibit similar power characteristics.

We model the arrivals and departures of jobs to a datacenter using an M/D/1 queueing model. Jobs are assumed to arrive with inter-arrival time exponentially distributed with parameter λ_{job}, and are queued in a dispatcher node until all the previous jobs have been serviced. The service time for a job is considered fixed and according to the M/D/1 queueing model, the utilization of the cluster is $U = T_{\mathcal{P}}\lambda_{job}$, where $T_{\mathcal{P}}$ is the service time of a job. We simulate the impact of utilization on the server or cluster by varying the arrival rate such that the utilization varies between 0 and 1 for a given time period T. When the utilization of the server or cluster is zero, the system is idle for the entire observation duration T.

We derive the peak power consumed by a node by modeling the energy consumed during a time period T and the utilization is 1.

$$P_{peak,\mathcal{P}} = \frac{E_{U=1}}{T}$$

The idle power is determined when the system is not executing any job and utilization is 0.

$$P_{idle,\mathcal{P}} = \frac{E_{U=0}}{T}$$

Table 3: Summary of energy proportionality extensions

Metrics	
DPR	$100 - P_{idle}(\%)$
IPR	$\frac{P_{idle}}{P_{peak}}$
EPM	$1 - \frac{\int_0^{100} P_{server} \cdot du - \int_0^{100} P_{ideal} \cdot du}{\int_0^{100} P_{ideal} \cdot du}$
LDR	$\max_u^{\lvert\cdot\rvert} \frac{P(u) - ((P_{peak} - P_{idle})u + P_{idle})}{(P_{peak} - P_{idle})u + P_{idle}}$
$PG(u)$	$\frac{P(u)_{server} - P(u)_{ideal}}{P(u)_{ideal}}$

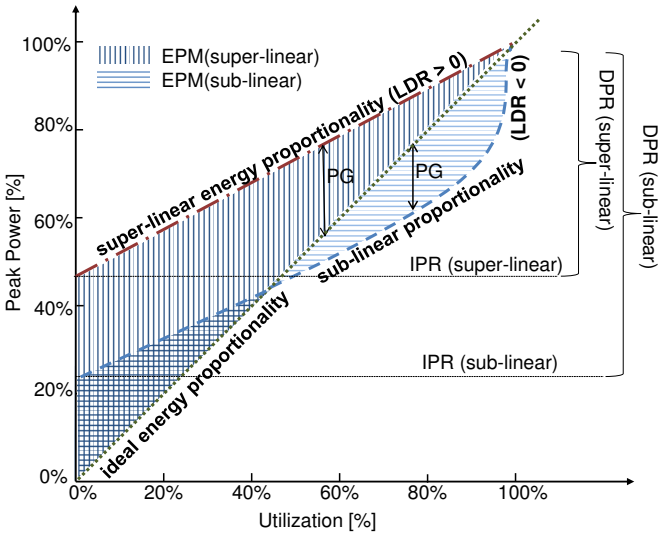


Figure 2: Energy proportionality metric relationships

An ideal energy-proportional system consumes no power when idle and its power consumption grows linearly with the amount of utilization of its resources. For example, the ideal energy-proportional system consumes 10% of its peak power at 10% utilization as shown in Figure 2. However, the actual proportionality of servers or clusters can be super-linear or sub-linear. Multiple metrics have been proposed by researchers to quantify the energy proportionality of individual server nodes. While both the DPR and IPR metrics capture server power at zero utilization, server's power consumption is known to increase non-linearly with utilization [40], which is not accounted for by these metrics. Since the majority of datacenters over-provision their servers to achieve reasonable QoS at peak utilization, most servers operate at 30% utilization on an average [9]. Hence, a more meaningful metric would be to capture the ratio between power consumption at 30% and 100% utilization. However, datacenter operators will find it convenient to shift workloads accordingly if they have prior knowledge of server proportionality at different utilization levels, which is determined by our approach.

As shown in Figure 2, the EPM metric measures server's power consumption at different utilization levels using the area between server's power consumption and the ideal power consumption curve. Varsamopoulos et al. [42] propose a metric called the Linear Deviation Ratio (LDR) to account for the linearity of server's energy proportionality curve across utilization . While both EPM and LDR account for proportionality across utilization levels, the results are expressed as a single value. For the EPM metric, a value of one indicates that server consumes power proportional to its load, while a value of zero indicates that the server consumes a constant amount of power irrespective of its load [33]. For LDR, lower values indicate a close-to-linear system, negative values represent *sub-linear energy proportionality* and positive values represent *super-linear proportionality*. The aggregation of these metrics into a single value limits the analysis of energy-efficient configurations across cluster utilization levels. In contrast, the Proportionality Gap (PG) metric [44], is defined at each utilization and a lower value of PG is indicative of a more energy-proportional server.

Figure 2 summarizes the relationships among recent energy proportionality metrics. In the sections that follow, we use these metrics to analyze the energy proportionality of both homogeneous and heterogeneous clusters with wimpy and brawny nodes. While the energy proportionality metrics factor in only the power consumption with respect to the peak power and utilization of the server (or cluster), another metric that factors throughput along with power is the performance-to-power ratio (PPR). It is defined as the throughput of the workload per unit power across utilization levels,

$$PPR(u) = \frac{Throughput[operations/s]}{Power[W]}$$

where throughput denotes the number of useful operations performed by the system per unit time. This metric is also used in SPEC benchmark [35].

C. Workloads and System Setup

As we are targeting datacenter systems, we select six typical workloads with different deadline requirements and exposing different performance bottlenecks. *EP*, from NPB benchmark [8], is an embarrassingly parallel distributed-memory program that generates random numbers for Monte-Carlo numerical simulation. *Memcached* is widely used by Facebook, Amazon, Twitter, among others, as an in-memory key-value distributed storage. When a key request arrives, a front-end node dispatches the request to a set of nodes that are responsible for storing the key-values belonging to an application. All nodes in the pool perform a key look-up computation, but typically few nodes return the value. However, this operation exerts complex service demands on core, memory and I/O devices [21], [41]. We use memslap running on another system to trigger requests to the memcached server over a 1 Gbps network connection. This memslap program generates requests with fixed key-value size and uniform popularity.

From the PARSEC benchmark suite [11], *x264* represents the widely used encoding algorithm for streaming video, and *blackscholes* represents a quantitative model for determining option pricing. The open source speech recognition engine *Julius* [3] represents the increasing adoption of real-time speech processing workloads originating from smart devices. To analyze the energy efficiency of web security, we use the openssl *RSA-2048* speed benchmark because major web players are increasingly concerned with the in-transit data security and are hardening the https encryption [2]. These representative workloads along with the cluster validation results are summarized in Table 4 and each type of workload constitutes a single job. The errors denote the percentage difference between the model and the measured values.

Table 4: Cluster validation

Domain	Program	Execution time error[%]	Energy error[%]
HPC	EP	3	10
Web Server	memcached	10	8
Streaming video	x264	11	10
Financial	blackscholes	4	7
Speech recognition	Julius	13	1
Web security	RSA-2048	2	8

Datacenters typically receive multiple jobs concurrently from many users. To represent the arrival of multiple jobs, we vary the number of jobs per batch. In our analysis, this variation in the number of jobs per batch and number of batches in an observation interval varies the utilization of the individual servers and the cluster system.

D. Inter-node Heterogeneous Cluster System

In this paper, we consider scale-out workloads [16], [25], which are highly parallelizable with negligible inter-node communication. Such programs have repeating parallel phases of execution and different service demands for the cores, memory and network I/O resources depending on application domain and problem size. Many datacenter workloads must obey strict service time deadlines. To service requests within a deadline, processing is distributed over hundreds of server nodes. Jobs arrive at front-end nodes and are forwarded to a cluster of back-end leaf nodes that service job requests. Both response time and the energy incurred by a job are dominated by leaf nodes [25]. Thus, we focus on the energy proportionality of heterogeneous leaf nodes with diverse PPRs as shown in Figure 3.

All nodes are multicore systems, and all cores inside a node operate at a core clock frequency $f \in [f_{min}, f_{max}]$, where f_{min} and f_{max} are specific to each type of node. Because we target typical server systems, we consider that the cores inside a node are super-scalar and support out-of-order execution where at least one integer instruction, one floating point instruction and one memory request instruction can be issued within each CPU cycle. Because of out-of-order architecture, the execution of instructions for which the data is available can be overlapped with the time required to retrieve the data for subsequent instructions [14].

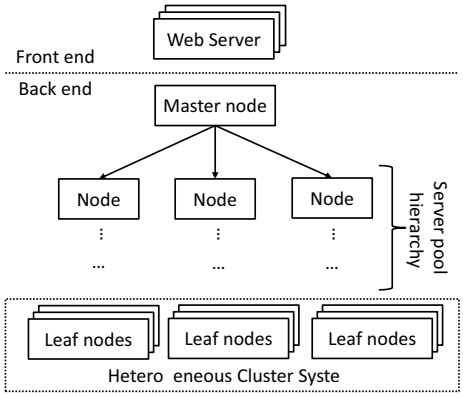


Figure 3: Heterogeneous cluster

For analysing cluster-wide energy proportionality, we use the same metrics as for the single node case and the power consumed by the whole cluster under consideration. For cluster utilization, we consider the same workload being executed by all the nodes in the cluster. While all nodes of the same type execute similar proportion of the workload, the amount of workload executed by nodes of different types is determined by matching the execution rates among the different types of nodes, such that all nodes finish executing at the same time as shown in Table 2. Thus, in our approach, the idling period of all nodes in a system configuration is approximately the same and depends only on the cluster utilization level.

For simplicity, each node has a single memory controller (i.e. Uniform Memory Architecture) that is equally shared among all the cores of the system. The I/O devices in modern server systems are memory-mapped and can transfer data to and from the main memory with minimal intervention from the CPU, because the transfers are controlled by a specialized processor called the DMA controller. Thus, the activities of the network I/O devices can be completely overlapped with the CPU activities. Most modern multicore systems are covered by this model of execution, including high-performance Intel Xeon or AMD Opteron systems, and low-power ARM Cortex-A8, Cortex-A9, Cortex-A15 and Cortex-A57 systems. In this paper, we consider systems with one I/O network device with workloads having negligible storage I/O requirements.

While our proposed approach can analyze a generic mix of heterogeneous nodes, for validation, we consider a mix of **A9** (ARM), and **K10** (AMD), as shown in Table 5. These representative nodes cover the broad spectrum of performance and power offered by computing platforms today. At one end of the spectrum is the low-power A9 node which consumes a peak power of only 5W. At the other end of the power spectrum, we select the K10 node that consumes a peak power of about 60W and offers a peak performance around 50 GFLOPS. We use *perf* to access hardware event counters and to measure execution time, and a Yokogawa WT210 power monitor to measure the power and energy, as shown in Fig. 4.

Table 5: Types of heterogeneous nodes

Node	ARM Cortex A9	AMD Opteron K10
ISA	ARMv7-A	x86_64
Clock Freq	0.2–1.4 GHz	0.8–2.1 GHz
Cores/node	4	6
L1 data cache	32KB / core	64KB / core
L2 cache	1MB / node	512KB / core
L3 cache	NA	6MB / node
Memory	1GB LP-DDR2	8GB DDR3
I/O bandwidth	100Mbps	1Gbps

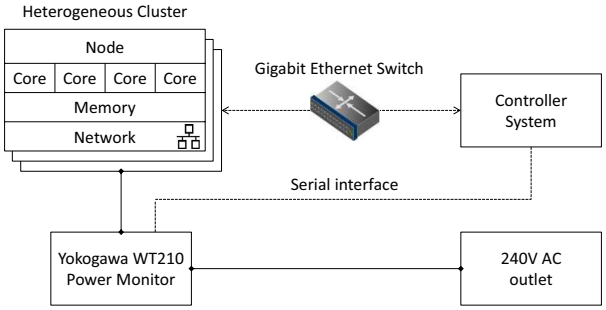


Figure 4: Validation setup

III. Analysis

A. Performance-to-Power Ratios

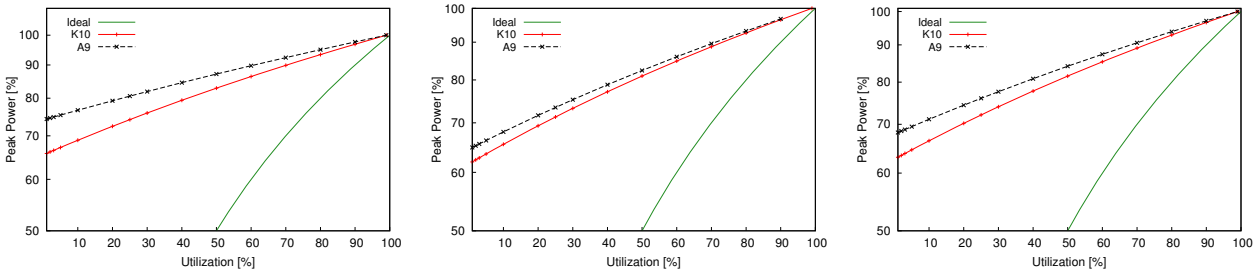
PPR is defined as the work done per unit of time, normalized by the average power consumption. This is equivalent to the work done per unit of energy. The unit of work for different workloads depends on the program and are shown in Table 6. The PPRs computed for the most energy-efficient configuration per type of node are shown in Table 6. As observed from the table, A9 has a better PPR than K10, but with two notable exceptions. For web-security applications such as RSA-2048, K10 has better PPR due to its special instructions that accelerate cryptography processing. x264 encoding algorithm is memory-bound [11], and performs much better on the K10 node which has a higher memory bandwidth. For the other applications, A9 has a better PPR but lower overall performance. This makes a case for mixing these two nodes with diverse PPR and analyze if such a cluster scales the energy proportionality wall.

Table 6: Performance-to-power ratio

Program	Performance per Watt (PPR)	A9 node	K10 node
EP	(random no./s)/W	6,048,057	1,414,922
memcached	(bytes/s)/W	5,224,004	2,68,067
x264	(frames/s)/W	0.7	1
blackscholes	(options/s)/W	11,413	2,902
Julius	(samples/s)/W	69,654	21,390
RSA-2048	(verify/s)/W	968	1091

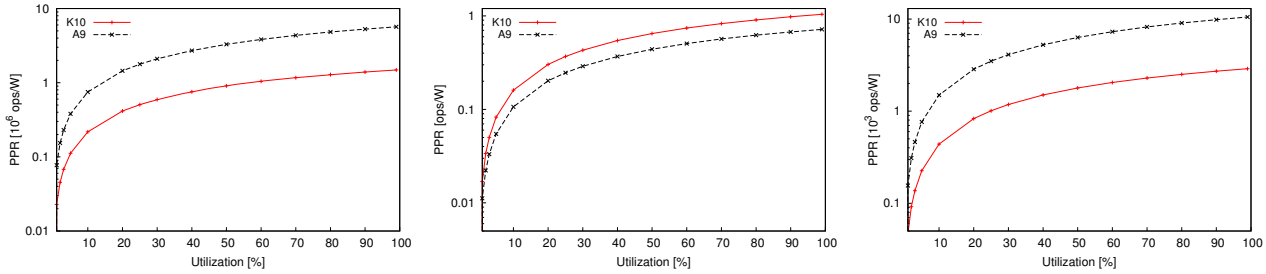
B. Energy Proportionality Analysis of Brawny and Wimpy Node

Table 7 shows the different energy proportionality metrics described in Section II-B for the A9 and K10 nodes across



(a) EP (b) x264 (c) blackscholes

Figure 5: Energy proportionality of brawny and wimpy nodes[1]



(a) EP (b) x264 (c) blackscholes

Figure 6: PPR of brawny and wimpy nodes[2]

all the workloads considered in this paper. From the results in the table, brawny K10 node has better energy proportionality compared to the wimpy low-power A9 node. Secondly, an interesting observation from the table is all of these metrics indicate the same value for a given program. While the EPM and LDR values are equal to $1 - IPR$, the DPR value is $(1 - IPR) \times 100$. All of these metrics use P_{peak} and P_{idle} to quantify proportionality over a given period. As the metrics in the table are cumulative and do not indicate the proportionality of nodes at individual utilization levels, we plot the percentage of peak power consumed by these nodes to further analyze the variation of the proportionality gap with utilization.

Figures 5a, 5b and 5c plot the energy proportionality of a single AMD Opteron K10 node and ARM Cortex-A9 node for the EP, x264 and blackscholes programs respectively. The utilization percentage for all the plots is determined by varying the number of jobs in a given observation interval, T, based on the time per job $T_{\mathcal{P}}$ as described in Section II-B. From the plots, we conclude that usage of K10 nodes in system clusters is more energy-proportional than using the A9 node, for compute and memory intensive workloads. However, comparison between the absolute values of the idle power consumed by the two nodes shows that the idle power of A9 (≈1.8W) is at least 25 times lower than that of K10 (≈45W). This counter-intuitive result is because the energy proportionality metrics described in Section II-B do not portray a complete picture as they only provide the percentage of the power consumption with respect to the peak at different utilization levels. The metrics neither consider the absolute power values nor do they consider performance characteristics.

Table 7: Single-node energy proportionality

Program	DPR		IPR		EPM		LDR	
	A9	K10	A9	K10	A9	K10	A9	K10
EP	25.97	34.57	0.74	0.65	0.26	0.34	0.26	0.35
memcached	16.78	11.05	0.83	0.89	0.17	0.11	0.17	0.11
x264	35.54	38.41	0.64	0.62	0.36	0.38	0.36	0.39
blackscholes	32.11	37.30	0.68	0.63	0.32	0.37	0.32	0.37
Julius	30.48	38.10	0.70	0.62	0.30	0.38	0.31	0.38
RSA-2048	35.62	41.19	0.64	0.59	0.36	0.41	0.36	0.41

Figures 6a, 6b and 6c plot the PPR across utilization levels for a single node of K10 and A9 executing EP, x264 and blackscholes workloads respectively. For certain workloads like x264, both the PPR and energy proportionality metrics concur (Figures 5b and 6b), wherein K10 has both a better PPR and proportionality gap compared to A9. However, the comparison of the energy proportionality and the PPR values of the EP and blackscholes workloads show contradictory results in terms of determining the more efficient node.

While the PPR of the A9 wimpy node is better than that of brawny K10 for executing both EP and blackscholes workload, the proportionality gap of the A9 node is bigger than that of K10 for these workloads. This contradicting result stems from the fact that while energy proportionality determines how the server power consumption adapts to different utilization levels, it does not consider the throughput.

[1]These plots depict the percentage of peak power consumed by the system for a specific utilization value as defined in Section II. EPM is the area between the energy proportionality of the system and the ideal energy proportionality.

[2]For all the PPR plots presented in this paper, higher is better

Table 8: Cluster-wide energy proportionality

Program	DPR			IPR			EPM			LDR		
	128 A9 0 K10	64 A9 8 K10	0 A9 16 K10	128 A9 0 K10	64 A9 8 K10	0 A9 16 K10	128 A9 0 K10	64 A9 8 K10	0 A9 16 K10	128 A9 0 K10	64 A9 8 K10	0 A9 16 K10
EP	25.97	32.66	34.57	0.74	0.67	0.65	0.26	0.33	0.34	0.26	0.33	0.35
memcached	16.78	12.44	11.05	0.83	0.88	0.89	0.17	0.12	0.11	0.17	0.12	0.11
x264	35.54	37.73	38.41	0.64	0.62	0.62	0.36	0.38	0.38	0.36	0.38	0.38
blackscholes	32.11	36.10	37.30	0.68	0.64	0.63	0.32	0.36	0.37	0.32	0.36	0.37
Julius	30.48	36.39	38.09	0.70	0.64	0.62	0.30	0.36	0.38	0.30	0.37	0.38
RSA-2048	35.62	39.92	41.19	0.64	0.60	0.59	0.36	0.40	0.41	0.36	0.40	0.41

C. Cluster-wide Energy-proportionality

We assume a fixed system configuration across all utilization levels to ensure that the energy proportionality analysis is an unbiased comparison among different cluster mixes. Furthermore, to ensure a fair comparison among cluster mixes, we constrain the peak power of the cluster using a fixed power budget. This is motivated by the fact that datacenters often have an upper bound on their peak power consumption. Based on peak power consumed by the A9 and K10 node, we analyze both homogeneous clusters and cluster mixes such that the total peak power is within the allocated budget.

For this analysis we consider a peak power budget of 1kW. The combination of the different heterogeneous cluster mixes within a 1kW power budget can be determined using a power substitution ratio of 8:1 between the A9 and K10 nodes[3].

The values of the energy proportionality metrics for the homogeneous clusters and a heterogeneous cluster are shown in Table 8. The results of the cluster-wide energy proportionality are similar to the single-node case. As seen from the values in the table, high-power homogeneous clusters consisting of K10 nodes have better energy proportionality compared to the homogeneous cluster with A9 nodes. However, the K10 cluster consumes an idle power of around 720W which is about three times higher compared to the A9 cluster. Thus, this contradiction shows that using only energy proportionality metrics do not always reveal the most energy-efficient system configuration. This conclusion is further augmented by comparing the cluster-wide energy proportionality and PPR for different workloads.

Figures 7 and 8 plot the energy proportionality and PPR for the different cluster configurations executing EP workload respectively. Comparing the proportionality gaps of different clusters and their respective PPRs, it is evident that the efficient *heterogeneous* system configurations determined by the two metrics are completely different. While the energy proportionality advocates the use of 32 A9 and 12 K10 node mix, the PPR advocates the mix with 96 A9 and 4 K10 nodes.

The cluster-wide energy proportionality plots for all the workloads indicate that the homogeneous configuration using K10 nodes has the least proportionality gap. On the contrary, the PPR plots of EP and blackscholes workloads illustrate that

[3]This ratio is derived based on the peak powers of the A9 and K10 nodes. Since one A9 node draws a peak power of 5W and one K10 node draws a peak power of 60W, one K10 node can be replaced by 12 A9 nodes. Factoring about 20W peak power drawn by the switch that connects the A9 nodes, gives us a power substitution ratio of 8:1.

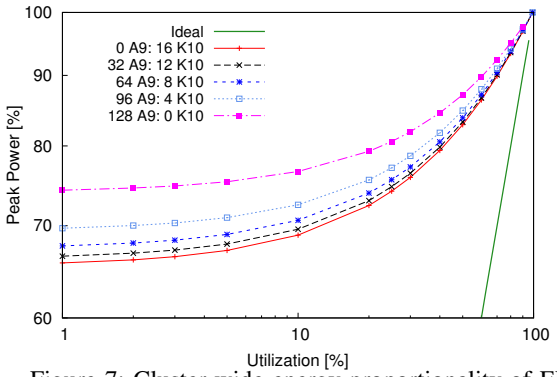


Figure 7: Cluster-wide energy proportionality of EP

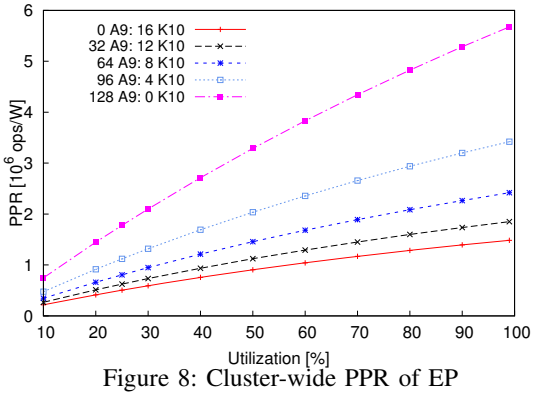


Figure 8: Cluster-wide PPR of EP

the homogeneous configuration consisting of 128 A9 nodes exhibits the best PPR. This contradiction between energy proportionality and PPR is also exposed when determining energy-efficient heterogeneous configurations. For the blackscholes workload, while the energy proportionality shows that a mix with 32 A9 and 12 K10 nodes has the least proportionality gap, while the PPR indicates that this mix has the worst PPR among the heterogeneous mixes. For determining efficient system execution configurations, PPR offers better insights as it additionally factors in the performance along with the power used. Thus, in comparison with the energy proportionality metrics, PPR provides better insights about system resource imbalances and inefficiencies.

D. Does Inter-node Heterogeneity Scale the Energy Proportionality Wall?

Inter-node heterogeneous systems allow for better matching between an application's resource demands and available system resources due to different combinations of system

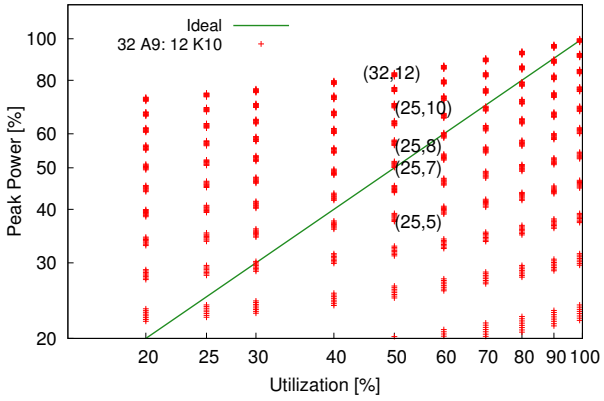


Figure 9: Energy proportionality of Pareto-optimal configurations for EP

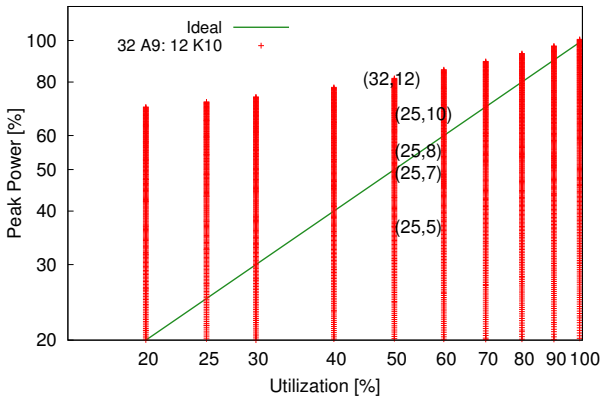


Figure 10: Energy proportionality of Pareto-optimal configurations for x264

parameters such as types of nodes, number of nodes of each type, number of active cores per node and the operating core clock frequency. For example, a system with ten AMD and ten ARM nodes results in a total of 36,380 possible heterogeneous configurations[4]. An approach to reduce the configuration space is beyond the scope of this paper.

While our previous work [31] shows that, among the large set of configurations there exists a Pareto-optimal set of heterogeneous configurations that form the energy-deadline Pareto frontier, the impact of these Pareto-optimal configurations on cluster-wide energy proportionality is non-obvious. Figures 9 and 10 plot the energy proportionality for the configurations on the Pareto-frontier using a maximum of 32 A9 and 12 K10 nodes, executing the EP and x264 workload respectively. Different configurations achieve the same utilization by varying the number of jobs executed in a given observation interval T.

As observed, several Pareto-optimal configurations have sub-linear energy proportionality, as they fall below the ideal

[4]a) Mix of ARM and AMD nodes = 10 (ARM nodes) $\times$ 5 (core frequencies per ARM node) $\times$ 4 (number of cores per ARM node) $\times$ 10 (AMD nodes) $\times$ 3 (core frequencies per AMD node) $\times$ 6 (cores per AMD node) = 36,000; b) Considering only ARM nodes, 10 $\times$ 5 $\times$ 4 = 200; c) Considering only AMD nodes, 10 $\times$ 3 $\times$ 6 = 180. Total $= 36,000 + 200 + 180 = 36,380$

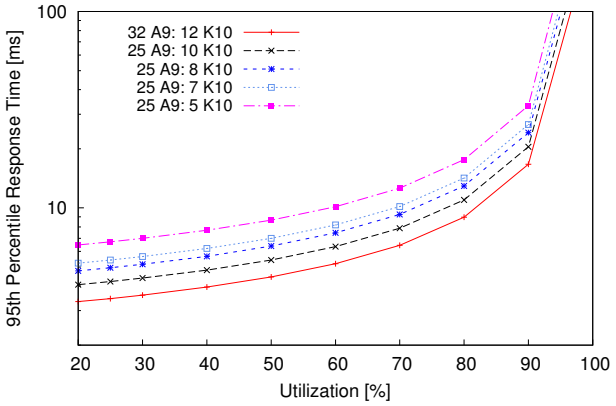


Figure 11: Response time of sub-linear heterogeneous mixes for EP

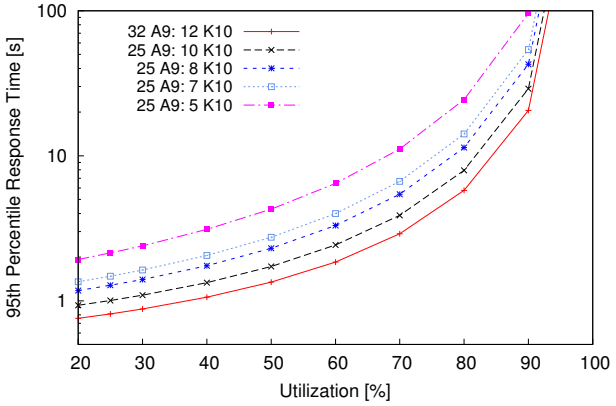


Figure 12: Response time of sub-linear heterogeneous mixes for x264

energy proportionality. These sub-linearly proportional configurations arise by reducing the number of brawny nodes. As seen from the plots, the configurations below the ideal proportionality have decreasing number of brwany nodes thus decreasing the idle power and becoming sub-linearly proportional. For example, given a maximum of 32 A9 and 12 K10 nodes executing the EP workload, a configuration with 25 A9 and 8 K10 nodes is above the ideal proportionality, but a configuration with 25 A9 and 7 K10 nodes exhibits sub-linear proportionality for cluster utilization of 50% as shown in Figure 9. While these configurations with smaller number of K10 nodes consume less energy than ideal, they trade-off execution time to save energy. This trade-off with respect to the response time deadline is not clear from the energy proportionality plots alone. Hence, the response time analysis for these configurations gives additional insights in choosing an energy- and time-efficient configuration.

E. Impact of Sub-linear Configurations on Response Time

Figure 11 plots the 95th percentile response times for the EP workload executed on different heterogeneous mixes that have sub-linear energy proportionality. This plot shows that the response time differences among the configurations is in

the sub-millisecond range. *Thus, we show that heterogeneity introduces sub-linearly proportional configurations that do not impact response times.* However, this observation holds only when the PPR of wimpy nodes is better than the PPR of brawny nodes because these sub-linear configurations exist by removing brawny nodes and reducing the power consumed. Therefore, for workloads such as x264 where the brawny clusters outperform wimpy clusters, heterogeneity introduces sub-linear energy-proportional configurations, but the execution time is degraded to the order of seconds.

Figures 10 and 12 plot the energy proportionality and the 95th percentile response time for heterogeneous clusters with a maximum 32 A9 and 12 K10 nodes executing the x264 workload. While the number of sub-linear configurations for x264 is larger compared to the EP workload, these configurations suffer from response time degradation in the order of seconds as shown in Figure 12. Thus, while heterogeneity scales the proportionality wall, these sub-linear configurations are not time-efficient for workloads that have a lower PPR on wimpy nodes compared to brawny ones, such as x264. The existence of the energy- and time-efficient configurations among the sub-linearly proportional configurations is directly derived from the PPR of the programs on the two types of nodes. As high-performing nodes are removed, execution time increases, but if the PPR of the low-power nodes is better then these sub-linear configurations are both energy and time efficient.

IV. Related Work

There are many research works on managing energy inefficiencies of datacenter workloads and server systems. We classify the related work pertaining to this paper using two broad categories (i) energy proportionality and (ii) energy efficiency techniques for heterogeneous clusters, and compare them with the contributions of this paper.

A. Energy Proportionality

Energy proportionality studies of warehouse scale computers and strategies to improve non-peak power efficiency has been widely explored [15], [22], [36]. There are many research studies that employ dynamic strategies using software driven transitions to exploit multiple low-power server modes. Such strategies and techniques complement the energy proportionality study in this paper and can be applied to further reduce the inefficiencies of heterogeneous cluster mixes. While Hsu et al. [17] question the linearity of the energy proportionality curve and show that most modern servers follow a quadratic trend, we show that the energy proportionality metric alone does not suffice to study clusters consisting of nodes with diverse PPRs.

Barroso et al. [9] suggest that energy proportionality at the system level cannot be achieved through CPU optimizations alone, but instead requires improvements across all components, such as memory and network devices. Energy proportionality of server-level memories for datacenter workloads and datacenter network architectures have been proposed [23], [34], but these studies are tangential to our work, as we study the impact of inter-node heterogeneity on cluster-wide energy proportionality. With servers exhibiting lesser proportionality gaps at higher utilization levels, workload co-location strategies have been proposed to increase cluster utilization [20], [24], [46]. The implications of high energy-proportional servers such as KnightShift [44] on cluster wide energy proportionality was studied [45]. However, this paper addresses energy proportionality for inter-node heterogeneous clusters consisting of node mixes with diverse PPRs.

B. Energy efficiency techniques for heterogeneous systems

Many research studies explore dynamic algorithms for both power management [13], [19], [27], and energy efficiency of heterogeneous clusters [7], [47], [48]. These techniques complement our approach and can be applied to the energy proportional heterogeneous mixes analyzed in this paper. For a given power budget, a proposed cluster-in-a-box production system using *only* low-power CPUs improves the performance per watt-hour [37]. However, we analyze energy proportionality for a given power budget using a heterogeneous *mix* of nodes with diverse PPRs.

More recently, with the emergence of ARM big.LITTLE, a hierarchical power management approach to optimize the performance per watt within a thermal-design power budget is proposed by Muthukaruppan et al. [28]. While these works focus on *intra-chip* heterogeneity, our proposed approach is applicable for *inter-node* heterogeneity and addresses multi-ISA heterogeneity. Nathuji et al. [29] propose an intelligent workload allocation method to exploit across-platform heterogeneity for power efficiency. However, we analyze the energy proportionality of heterogeneous systems having nodes with diverse PPRs. Our previous work [31] modeled the execution time and energy of heterogeneous systems having nodes with diverse PPRs, but does not address energy proportionality.

V. Conclusion

With heterogeneity becoming ubiquitous in datacenters due to the paradigm shift from high-performance to low-power designs in server systems, new opportunities arise for an energy-efficient matching of workload service demands and resource capabilities. Motivated by this, our paper proposes an energy proportionality analysis to determine the impact of brawny and wimpy inter-node heterogeneity on energy proportionality. By extending our time-energy model to include a range of energy proportionality metrics, we analyze the energy proportionality of both homogeneous and heterogeneous clusters with high-performance (*brawny*) and low-power (*wimpy*) nodes. We perform this analysis for a wide range of datacenter workloads representing application domains such as web-hosting, multimedia streaming, financial analytics, real-time speech recognition, and web-security. Using two different types of nodes, AMD (brawny) and ARM (wimpy) processors, we cover the broad spectrum of performance and power offered by computing platforms today.

For a given power budget, we show that inter-node heterogeneous clusters exhibit the positive effect of scaling the

energy proportionality wall by enabling sub-linear energy-proportional configurations. Furthermore, we show that for workloads that have better PPR on wimpy systems, these configurations have minimal impact on the 95th percentile response time.

VI. Acknowledgements

This work is supported by the National University of Singapore under grant number R-252-000-601-112.

References

[1] DDR3 Specification, http://www.webcitation.org/6JN7G4r3x, 2010.
[2] Google finishes 2048-bit RSA migration, Yahoo to encrypt all data early next year, http://www.webcitation.org/6LaPMkEj0, 2013.
[3] Julius, http://julius.sourceforge.jp/, 2013.
[4] Y. Agarwal, S. Hodges, R. Chandra, J. Scott, P. Bahl, R. Gupta, Somniloquy: Augmenting Network Interfaces to Reduce PC Energy Usage, *Proc. of NSDI*, pages 365–380, 2009.
[5] V. Anagnostopoulou, S. Biswas, H. Saadeldeen, A. Savage, R. Bianchini, T. Yang, D. Franklin, F. T. Chong, Barely alive memory servers: Keeping data active in a low-power state, *J. Emerg. Technol. Comput. Syst.*, 8(4):31:1–31:20, Nov. 2012.
[6] D. G. Andersen, J. Franklin, M. Kaminsky, A. Phanishayee, L. Tan, V. Vasudevan, FAWN: A Fast Array of Wimpy Nodes, *Proc. of 22nd SOSP*, pages 1–14, 2009.
[7] N. Auluck, S. Betha, B. Mangipudi, Contention Aware Energy Efficient Scheduling on Heterogeneous Multiprocessors, *IEEE Transactions on Parallel and Distributed Systems*, 2014.
[8] D. Bailey, T. Harris, W. Saphir, R. Van Der Wijngaart, A. Woo, M. Yarrow, The NAS Parallel Benchmarks 2.0, Technical Report NAS-95-020, NASA Ames Research Center, 1995.
[9] L. A. Barroso, J. Clidaras, U. Hölzle, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second edition, *Synthesis Lectures on Computer Architecture*, 8(3):1–154, 2013.
[10] L. A. Barroso, U. Hölzle, The Case for Energy-Proportional Computing, *IEEE Computer*, 40, 2007.
[11] C. Bienia, S. Kumar, J. P. Singh, K. Li, The PARSEC Benchmark Suite: Characterization and Architectural Implications, *Proc. of PACT*, pages 72–81, 2008.
[12] J. Burge, P. Ranganathan, J. Wiener, Cost-aware Sheduling for Heterogeneous Enterprise Machines (CASH'EM), *Proc. of Cluster Computing*, pages 481–487, 2007.
[13] J.-J. Chen, K. Huang, L. Thiele, Power Management Schemes for Heterogeneous Clusters under Quality of Service Requirements, *Proc. of 26th ACM Symposium on Applied Computing*, pages 546–553, 2011.
[14] J. Corbet, A. Rubini, G. Kroah-Hartman, *Linux Device Drivers, 3rd Edition*, O'Reilly Media, Inc., 2005.
[15] X. Fan, W.-D. Weber, L. A. Barroso, Power Provisioning for a Warehouse-sized Computer, *Proc. of 34th ISCA*, pages 13–23, 2007.
[16] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, B. Falsafi, Quantifying the Mismatch between Emerging Scale-Out Applications and Modern Processors, *ACM Transactions on Computer Systems*, 30(4):15:1–15:24, 2012.
[17] C.-H. Hsu, S. Poole, Revisiting Server Energy Proportionality, *Proc. of 42nd ICPP*, pages 834–840, 2013.
[18] V. Janapa Reddi, B. C. Lee, T. Chilimbi, K. Vaid, Web Search Using Mobile Cores: Quantifying and Mitigating the Price of Efficiency, *Proc. of 37th ISCA*, pages 314–325, 2010.
[19] U. R. Karpuzcu, A. Sinkar, N. S. Kim, J. Torrellas, EnergySmart: Toward energy-efficient manycores for Near-Threshold Computing, *Proc. of 19th HPCA*, pages 542–553, 2013.
[20] J. Leverich, C. Kozyrakis, Reconciling High Server Utilization and Sub-millisecond Quality-of-service, *Proc. of 9th EuroSys*, pages 4:1–4:14, 2014.
[21] K. Lim, D. Meisner, A. G. Saidi, P. Ranganathan, T. F. Wenisch, Thin Servers with Smart Pipes: Designing SoC Accelerators for Memcached, *Proc. of 40th ISCA*, pages 36–47, 2013.
[22] D. Lo, L. Cheng, R. Govindaraju, L. A. Barroso, C. Kozyrakis, Towards Energy Proportionality for Large-scale Latency-critical Workloads, *Proc. of 41st ISCA*, pages 301–312, 2014.
[23] K. T. Malladi, B. C. Lee, F. A. Nothaft, C. Kozyrakis, K. Periyathambi, M. Horowitz, Towards Energy-proportional Datacenter Memory with Mobile DRAM, *Proc. of 39th ISCA*, pages 37–48, 2012.
[24] J. Mars, L. Tang, R. Hundt, K. Skadron, M. L. Soffa, Bubble-Up: Increasing Utilization in Modern Warehouse Scale Computers via Sensible Co-locations, *Proc. of 44th MICRO*, pages 248–259, 2011.
[25] D. Meisner, C. M. Sadler, L. A. Barroso, W.-D. Weber, T. F. Wenisch, Power Management of Online Data-intensive Services, *Proceedings of the 38th ISCA*, pages 319–330, 2011.
[26] D. Meisner, T. F. Wenisch, Dreamweaver: Architectural support for deep sleep, *Proc. of ASPLOS*, pages 313–324, 2012.
[27] K. Meng, R. Joseph, R. P. Dick, L. Shang, Multi-optimization Power Management for Chip Multiprocessors, *Proc. of 17th PACT*, pages 177–186, 2008.
[28] T. S. Muthukaruppan, M. Pricopi, V. Venkataramani, T. Mitra, S. Vishin, Hierarchical Power Management for Asymmetric Multi-core in Dark Silicon Era, *Proc. of 50th DAC*, pages 174:1–174:9, 2013.
[29] R. Nathuji, C. Isci, E. Gorbatov, Exploiting Platform Heterogeneity for Power Efficient Data Centers, *Proc. of 4th ICAC*, pages 5–5, 2007.
[30] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, X. Zhu, No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center, *Proc. of 13th ASPLOS*, pages 48–59, 2008.
[31] L. Ramapantulu, B. M. Tudor, D. Loghin, T. Vu, Y. M. Teo, Modeling the Energy Efficiency of Heterogeneous Clusters, *Proc. of 43rd ICPP*, pages 321–330, 2014.
[32] C. Rusu, A. Ferreira, C. Scordino, A. Watson, Energy-Efficient Real-Time Heterogeneous Server Clusters, *Proc. of 12th RTAS*, pages 418–428, 2006.
[33] F. Ryckbosch, S. Polfliet, L. Eeckhout, Trends in Server Energy Proportionality, *Computer*, 44(9):69–72, 2011.
[34] Y. Shang, D. Li, M. Xu, A Comparison Study of Energy Proportionality of Data Center Network Architectures, *Proc. of 32nd ICDCSW*, pages 1–7, 2012.
[35] SPEC, SPEC Power and Performance, Benchmark Methodology V2.1, *https://www.spec.org/power/docs/SPEC-Power_and_Performance_Methodology.pdf*, 2011.
[36] B. Subramaniam, W.-c. Feng, Towards Energy-proportional Computing for Enterprise-class Server Workloads, *Proc. of 4th ACM/SPEC International Conference on Performance Engineering*, pages 15–26, 2013.
[37] K. Sudan, S. Balakrishnan, S. Lie, M. Xu, D. Mallick, G. Lauterbach, R. Balasubramonian, A Novel System Architecture for Web Scale Applications using Lightweight CPUs and Virtualized I/O, *Proc. of 19th HPCA*, pages 167–178, 2013.
[38] A. S. Szalay, G. C. Bell, H. H. Huang, A. Terzis, A. White, Low-power Amdahl-balanced Blades for Data Intensive Computing, *SIGOPS Oper. Syst. Rev.*, 44(1):71–75, 2010.
[39] N. Tolia, Z. Wang, M. Marwah, C. Bash, P. Ranganathan, X. Zhu, Delivering Energy Proportionality with Non Energy-proportional Systems: Optimizing the Ensemble, *Proc. of HotPower*, pages 2–2, 2008.
[40] D. Tsirogiannis, S. Harizopoulos, M. A. Shah, Analyzing the Energy Efficiency of a Database Server, *Proc. of SIGMOD*, pages 231–242, 2010.
[41] B. M. Tudor, Y. M. Teo, On Understanding the Energy Consumption of ARM-based Multicore Servers, *Proc. of SIGMETRICS*, pages 267–278, 2013.
[42] G. Varsamopoulos, S. Gupta, Energy Proportionality and the Future: Metrics and Directions, *Proc. of 39th ICPPW*, pages 461–467, 2010.
[43] D. Wong, M. Annavaram, Knightshift: Scaling the Energy Proportionality Wall through Server-level Heterogeneity, *Proc. of 45th International Symposium on Microarchitecture*, pages 119–130, 2012.
[44] D. Wong, M. Annavaram, Scaling the Energy Proportionality Wall with KnightShift, *Micro, IEEE*, 33(3):28–37, 2013.
[45] D. Wong, M. Annavaram, Implications of High Energy Proportional Servers on Cluster-wide Energy Proportionality, *Proc. of 20th HPCA*, pages 142–153, 2014.
[46] H. Yang, A. Breslow, J. Mars, L. Tang, Bubble-flux: Precise Online QoS Management for Increased Utilization in Warehouse Scale Computers, *Proc. of 40th ISCA*, pages 607–618, 2013.
[47] X. Zhu, C. He, K. Li, X. Qin, Adaptive Energy-efficient Scheduling for Real-time Tasks on DVS-enabled Heterogeneous Clusters, *Journal of Parallel and Distributed Computing*, 72(6):751–763, 2012.
[48] Z. Zong, X. Qin, X. Ruan, K. Bellam, M. Nijim, M. Alghamdi, Energy-efficient Scheduling for Parallel Applications Running on Heterogeneous Clusters, *Proc. of ICPP*, 2007.