Holistically Evaluating the Environmental Impacts in Modern Computing Systems

Donald Kline, Jr.*, Nikolas Parshook*, Xiaoyu Ge†, Erik Brunvand‡, Rami Melhem†, Panos K. Chrysanthis†, Alex K. Jones*

*Dept. of ECE, University of Pittsburgh
Pittsburgh, PA 15261
Email: {dek61,nbp3,akjones}@pitt.edu

†Dept. of CS, University of Pittsburgh
Pittsburgh, PA 15260
{xig34,melhem,panos}@pitt.edu

‡Dept. of CS, University of Utah
Salt Lake, UT 84112
elb@cs.utah.edu

***Abstract*—There is mounting evidence that manufacturing energy and environmental costs are a growing factor in the overall energy footprint of computing systems. The quantification of these impacts requires the evaluation of both the manufacturing and use phase energy/environmental costs of major integrated circuit (IC) components, including processing units, memory, and storage. In particular, expansions of memory and cache can potentially increase manufacturing costs beyond what can be recovered through use phase advantages for reasonable usage patterns. With this holistic view of sustainability in mind, we provide evaluations of the environmental impacts of memory and cache options for Parsec and SPEC multi-program workloads. Using indifference point analysis, we determine which architectural decisions are the most sustainable in the context of these workloads for various usage scenarios. Through a form of break even analysis, we show the impact of upgrading to a new technology node. Our analysis of current processor trends indicates that upgrading may require upwards of 10 years of service time to break even, and that designing systems with smaller cache and main memory sizes may provide an overall positive environmental trend without dramatically reducing performance.**

I. Introduction

The concept of green or sustainable computing in the computer science domain has become highly associated with reducing energy consumption of computational devices and their supporting electronics during their *use phase*. However, to be truly sustainable, all phases of the system life-cycle must be considered. In contrast to the significant effort that has been applied to address use phase energy consumption, especially in battery powered embedded systems and data center servers, there is limited attention to the impacts from manufacturing computing systems and, in particular, their integrated circuits (ICs). This is an important problem, as semiconductor manufacturing is a significant and rising contribution to the environmental impacts of computing systems [1].

The vast majority of the integrated circuits used in modern computing systems, from tablets to servers, are devoted to the processing elements and the memory hierarchy. Moreover, even as core counts increase, growing last-level on-chip caches (LLCs) dominate the area of modern processors. Furthermore, many ICs are dedicated to main memory (e.g., DRAM) and storage (e.g., Flash). New, smaller geometries, made possible by advances in semiconductor fabrication, have been successful in increasing storage density of these ICs throughout the memory hierarchy. Emerging strategies, such as 3D designs and new memory technologies, also provide ways to further increase density throughout the system and in many cases can even reduce the use phase energy costs of memory and storage (i.e., non-volatile memories). Unfortunately, these technology trends of decreasing feature sizes, 3D CMOS, and hybrid fabrication techniques tend to dramatically increase the negative environmental impacts of fabrication [2].

In this paper we present *GreenChip*, the first predictive manufacturing and use phase environmental impact estimation flow based on targeted technology node and computer architecture design choices such as number of processor cores, cache and main memory sizes and architectures, and solid state disks. GreenChip can provide detailed analysis of these choices with an end goal of supplying consumers with more holistic environmental data akin to fuel efficiency reports for cars. We demonstrate this idea by using GreenChip to compare the impact of different computer architecture choices. We classify the workloads based on memory access requirements as one example of how the data can be aggregated. Our comparisons are made with indifference point [3] and break even analyses.

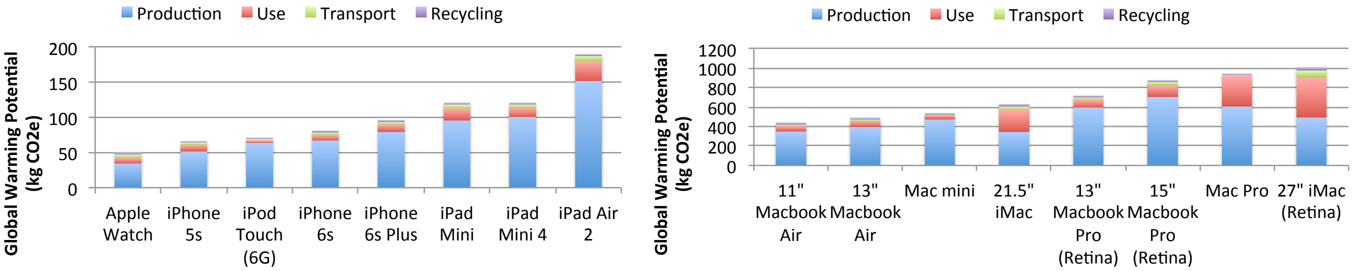
Indifference point analysis is a common economic metric to determine the point at which there is no difference in cost between two alternatives. For environmental analysis, we define the indifference point as the time when the energy to manufacture and operate two competing system architectures is equivalent. The indifference point can be compared to typical or projected product lifetimes to determine whether a change in manufacturing cost, either across technology generations or due to changes in system architecture, is justified. The break even time indicates the point when a new system will reach the same energy consumption of the system it will replace. This comparison assumes the manufacturing cost has already been invested for the original system. Thus, it identifies the *upgrade* time, when the energy for the new system will be less than leaving the original system in service.

In this work, we make the following contributions:

- We present GreenChip, the first predictive holistic sustainability evaluation flow of computing systems that considers architecture choices such as core count, cache size, main memory, and solid state storage architectures for different technology nodes. (Section III)

*This work was partially supported by NSF EFRI-1038139 and MCSI Seed awards. The authors also thank the support of A. Labrinidis and A. Zheng.

978-1-5090-5117-5/16/$31.00 ©2016 IEEE



(a) Tablets

(b) Computers

Fig. 1. Impact of manufacturing/production from ICs in "use phase energy" optimized systems.

- We apply indifference point and break even analysis to sustainable computing to evaluate when architectural decisions are appropriate for various system sleep and active time scenarios using workloads from the Parsec [4] and SPEC CPU [5] benchmark suites (Sections IV and V).
- We demonstrate the extensiblility of GreenChip to examine the interaction of main memory and solid state disks in the context of supercomputing applications using a SPEC OpenMP benchmark [6] (Section VI).

II. Background

The considerable attention and focus to use phase energy consumption in modern computing systems is a natural extension of research that aims to address thermal concerns caused by increases in power density associated with semiconductor technology scaling. These use phase energy reduction measures can help maximize battery life for mobile electronics and minimize operational energy costs of data centers.

To achieve holistic sustainability requires considering the entire computing life-cycle, for which a science called Life Cycle Assessment (LCA) is commonly used. LCA allows an engineer to quantitatively evaluate how processes and products use materials, water, and energy resources and the resulting environmental impacts throughout their lifetimes. Established guidelines for performing detailed LCAs are documented by the Environmental Protection Agency, Society for Environmental Toxicologists and Chemists, the International Organization of Standardization (ISO), and the American National Standards Institute [7, 8]. As defined by the ISO 14040 series, LCA is an iterative four-stage process including: 1) Scoping – defines the extent of analysis and the system boundaries; 2) Inventory Analysis – documents material and energy flows that occur within the system boundaries (life cycle inventory or LCI); 3) Impact Assessment – characterizes and assesses the environmental impacts using data obtained from the LCI (life cycle impact assessment or LCIA); and 4) Interpretation and Improvement – identifies opportunities to reduce the environmental burden throughout the product's life cycle.

There are three main LCA strategies: Process LCA, Economic Input/Output (EIO) LCA, and Hybrid LCA. Process LCA evaluates all steps involved in each stage of the LCA and directly evaluates their impacts as well as impacts from upstream elements such as the fundamental components used in the process. EIO LCA works from the principle that environmental impacts typically correlate with the financial cost of the process, and therefore uses the cost of a product to estimate its environmental impacts. Hybrid LCA uses a mixture of both Process and EIO LCA. Several life-cycle tools and databases have been developed such as the NREL LCI database [9] and the LCIA Tool for the Reduction and Analysis of Chemical and other environmental impacts (TRACI) for distinguishing carcinogenic impacts [10]In the next section we present some relevant research on LCA as it applies to computing systems.

A. LCA of Computing Systems

In Fig. 1, we present the carbon emissions from the life-cycles of various Apple products [11], demonstrating that the dominant phases are production (manufacturing) and use. Contrary to expectation, use phase impacts do not dominate; for tablets [Fig. 1(a)], manufacturing can reach as much as 90% of the overall carbon footprint. Additionally, while the use phase impact decreases across product generations, the manufacturing impact has continued to rise. For instance, comparing the iPhone 5s and 6s, the use phase impacts remain constant, but the manufacturing impact increases by more than 25%. A similar situation can be observed with the iPad Mini and Mini 4, which have nearly identical carbon impacts, but the use phase savings are entirely offset by the manufacturing increase. Looking at computing systems from laptops to workstations [Fig. 1(b)], we see a similar trend where manufacturing impacts are at least half, and in many cases far more than half, of the total carbon footprint.

Recent life-cycle studies [1, 2, 12, 13] have pinpointed ICs[1] and displays as having the dominant manufacturing environmental impacts of computing systems. As the use phase energy and resulting environmental impacts continue to decrease, there is mounting evidence that the environmental trends for IC manufacturing are becoming increasingly environmentally unfriendly. Considering the two desktop machines from Fig. 1(b) without an integrated display, the Mac Pro and Mac Mini gain 67% and 90% of their impacts from manufacturing of non-display components, respectively. In these cases, the IC components become the dominating contributors due to SSDs and large amounts of memory in addition to the processor and supporting circuitry. We explore IC manufacturing trends further in the next section.

[1] ICs are grouped with printed circuit board manufacturing [12] but shown to be negligible compared to ICs [2].

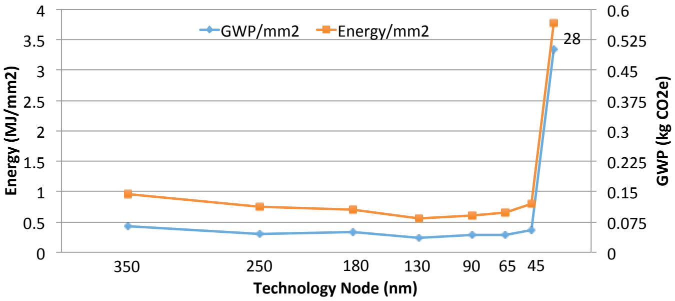


Fig. 2. IC fabrication energy and global warming potential (GWP) trends.

B. Impacts from IC Fabrication

A hybrid LCA of IC manufacturing over a 15 year period ranging from 350nm to 45nm [14] reveals problematic environmental trends. Environmental impacts from fabrication per area (Fig. 2) reached a minimum point at 130nm. Unfortunately, as technology descended below 90nm, the "environmental impacts per die area" increased with feature size. While these trends could be mitigated if IC die area per system decreased with the descent in feature size and resulting increase in device/transistor density, the opposite trend has been observed. Newer systems tend to have more IC area (for example, area increased from 750 to 1200 mm^2 between 2001 and 2010) due to trends to include more processor cores, embedded memory, accelerators such as graphics processing units, and solid state storage [12]. Moreover, trends such as "dark silicon," where many infrequently used hardware blocks and accelerators are included for use in improving energy efficiency and performance of relatively infrequent niche functions, also work against manufacturing sustainability.

Using the parameterized fabrication estimation method for ICs from Murphy et al. [15] combined with the Apple LCA data, it was possible to extend the chart in Fig. 2 from 45nm to 28nm. We examined the manufacturing cost of several apple tablet products over different generations implemented with processors fabricated at 45nm and 28nm. Using the reported breakdown of IC contributions [13] and the overall manufacturing effort at each node, the trend indicates a dramatic increase in manufacturing effort, supported by a transition in CMOS manufacturing from planar bulk CMOS to silicon-on-insulator at 36nm [16] (between 45nm and 28nm). This resulted in a significant savings in use phase energy [17] but seems to dramatically increase manufacturing effort. There is additional indication that lithography effort, currently the dominant component of manufacturing costs [18], may have seen a sharp increase.

Standard immersion lithography (193nm ArF source with immersion) provides a pitch size limited to approximately 60nm. 2X CMOS nodes (28nm and lower[2]) require some form of double patterning [18], which, depending on technique, can increase the number of lithography steps and resulting environmental impacts dramatically. The resulting data indicates a 5X increase in energy and GWP [11], which is consistent with this trend. Moreover, this trend appears to be poised to accelerate aggressively as nodes at 10nm and lower appear to require multiple patterning lithography. This is due to extreme ultraviolet lithography's earliest availability being predicted for the 7nm node [18] and is consistent with economic cost improvements of Dennard scaling breaking down at these nodes [19]. With current technology utilizing power-optimized hardware, production often exceeds 75% and reaches 90% of the total life-cycle cost for a 4-year service time (see Fig. 1) [11]. This fact, along with the aforementioned trends, points to a need to examine the holistic environmental cost.

The LCA results from many studies have identified use phase and manufacturing phase impacts as the dominant contributors to energy and carbon emissions for computing systems [2, 14]. This is demonstrated in Fig. 1, where phases like transportation and recycling are very low compared to the manufacturing and use phases of the system life cycles [11]. Our tool, GreenChip, focuses on these two dominant phases to provide a representative view of the system as shown in the next section.

III. The GreenChip Sustainable Computing Prediction and Evaluation Tool

To evaluate and compare the manufacturing and use phases of new computer architectures and systems we have created the GreenChip flow shown in Fig. 3. GreenChip can be used for the processor in isolation [Fig. 3(a)], combined with the main memory system [Fig. 3(b)], and even extended to consider secondary storage in the form of SSDs [Fig. 3(c)]. GreenChip first simulates the behavior of a mixture of workloads on a proposed architecture to generate performance statistics. The simulator output, architecture specification, and technology node are then fed into a use phase power estimation flow. To accomplish this, existing tools are available and may be leveraged. GreenChip is currently built from the Sniper full system simulator [20] and the McPAT use phase power estimation flow [17]. A more detailed, cycle-level simulator such as gem5 [21] can easily be used with GreenChip.

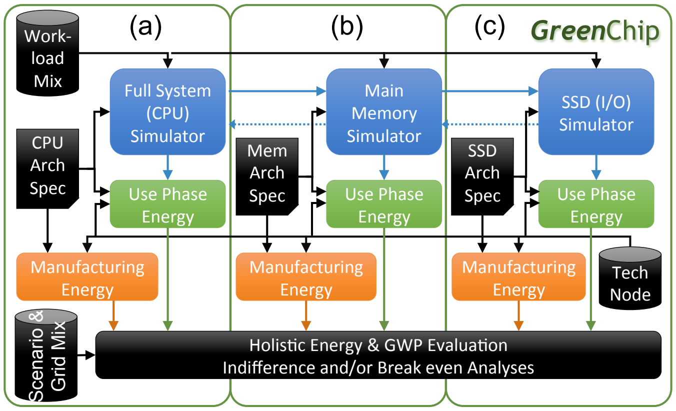


Fig. 3. GreenChip evaluation flow.

[2] 22nm is confirmed to require double patterning while 32nm only requires single patterning [18]. 28nm is assumed to require double patterning based on the increase in IC manufacturing impacts reported by Apple [11] which is a feasible changeover point for a 193+i lithography pitch limit of 60nm.

However, this level of detail may not be required for this type of analysis. Furthermore, the simulation time required for cycle level processor simulation makes it difficult to use large workloads that stress the main memory and disk in a reasonable amount of simulation time.

GreenChip extends this flow with a manufacturing environmental cost estimator[3] that uses a combination of the technology node impacts per area and the predicted area of the IC or ICs. For example, in the processor evaluation [Fig. 3(a)], area estimates for the processor for a particular technology node can be obtained from McPAT [17] and CACTI [22] and combined with the manufacturing cost per area of CMOS logic (Fig. 2) to determine overall manufacturing cost.

DRAM manufacturing cost is computed in a similar way to the processor cost using the DRAM technology trend [14]. DRAM tends to mirror the CMOS trends with an increasing cost per area starting after 70nm. Total cost is determined based on the die area, itself a function of the capacity per DRAM IC, combined with the cost per area. DRAM tends to trail CMOS by one technology node, so generally systems comparisons would consider the year a system was built rather than identical feature sizes (e.g., a 45nm processor would have a 57nm main memory system). Manufacturing cost can also be determined for Flash ICs used in SSDs, although the data is typically reported per capacity rather than die area [14], and is incorporated into GreenChip [Fig 3(c)] in the same fashion.

To determine the overall energy cost of a system during its lifetime, a usage scenario must be considered consisting of the time the system is awake versus asleep (*sleep ratio*), when awake, how much it is active versus idle (*active ratio*), and the time period it will be in service. We determine this number from the average power of the usage scenario shown in Eq. 1

$$P = (1 - r_S)(r_A(P_D + P_S) + (1 - r_A)P_S) + P_L \qquad (1)$$

where P_D, P_S, and P_L are dynamic, static, and sleep power, respectively, during workload execution, r_S is sleep ratio, and r_A is the active ratio of time spent executing the specified workload. Thus, the time t_W for the processor to be doing useful work is $t_W = t(1 - r_S)r_A$ where t is the service time. The overall energy cost is computed by $E = P \cdot t + M$ where M is the manufacturing cost described previously.

In addition to reporting raw environmental impact outputs for a particular system design, GreenChip also provides direct comparisons of two design choices. Using indifference analysis, the choice of system can be evaluated based on the expected service time. The indifference formula t_I of comparing two architectures, $System_0$ vs. $System_1$, is shown in Eq. 2. t_I is the time at which the increase in manufacturing cost will be outweighed by the savings in use phase cost. If the proposed service time $t < t_I$ the architecture with the lower manufacturing cost minimizes environmental impact and for a proposed service time $t > t_I$ the architecture with the lower use phase cost minimizes impact. If one system is lower in both costs, t_I is either < 0 or $= \infty$, making it invalid and pointing toward the selection of the lower cost system regardless of service time.

$$t_I = \frac{M_1 - M_0}{P_0 - P_1} \qquad\qquad t_B = \frac{M_1}{P_0 - P_1} \qquad (2)$$

The break even time t_B is also defined in Eq. 2. t_B represents, given an existing system ($system_0$), what service time for a new system ($system_1$) would be required to offset the upfront manufacturing cost to save overall energy. This is relevant to answer the "upgrade" question. For both of these comparisons, GreenChip automatically adjusts the selected usage scenario to account for the change in performance due to a different architecture configuration. Using the workload, the IPC of both proposed systems is determined. With $system_0$ as a baseline, the activity ratio of $system_1$ is adjusted by replacing r_A with $r'_A = r_A(\frac{IPC_0}{IPC_1})$ in Eq. 1[4].

Additionally, GreenChip is able to report various gas and byproduct emissions including carbon emissions and carcinogens from manufacturing and use phase energies for the U.S., China, and for a worldwide average using data from the literature [14] and electricity generation mix data [23]. For the remainder of this paper we focus on case study comparisons of energy from manufacturing and use phases of common architecture configurations at different technology nodes and for different workloads to highlight interesting trends and the importance of considering manufacturing impacts in developing next generation sustainable computers.

IV. Case Study I: Environmental Impacts of Recent Processor Trends

As processors have descended below the 90nm node, clock frequencies have become relatively fixed to manage thermal concerns. Performance improvements have instead been achieved by using the additional density per die to increase the number of processor cores and on-chip cache sizes. In this section we use GreenChip to demonstrate how these trends impact sustainability.

A. Experimental Setup

We consider pseudo ISO-area configurations across several different technology generations that mirror the processor products available in the corresponding years. In particular, a 90nm processor was configured with one core and 1MB of LLC, a 65nm processor with two cores and 2MB LLC, a 45nm processor with 4 cores and 4MB LLC, and a 28nm processor with eight cores and 8MB LLC. Each system employs a 4-way issue, out of order core model operating at 2.6GHz[5] with a bus-based interconnect to access the LLC and main memory. Cache and main memory latency and power consumption were taken from CACTI [22] and DRAMSim2 [24], respectively. Power consumption of the processor configurations (i.e., P_D, P_S, and P_L from Eq. 1) was determined using McPAT [17].

[3] When discussing manufacturing cost, effort or impact, we are referring to environmental impact/cost, not economic cost, unless specified.

[4] In our comparisons, the baseline system was typically set to the slower of the two systems. If $IPC_0 > IPC_1$, the activity ratio of $system_0$ would instead be adjusted by $r'_A = r_A(\frac{IPC_1}{IPC_0})$, with the activity ratio of $system_1$ unaltered.

[5] Clock speed is assumed invariant across technology nodes as is commonly the case due to power/thermal concerns.

TABLE I
MULTI-PROGRAM WORKLOADS AND MEMORY FOOTPRINTS FOR THE PARSEC AND SPEC BENCHMARKS. LOW (L), MEDIUM (M), AND HIGH (H) REPRESENTS THOSE RESPECTIVE MEMORY FOOTPRINTS.

Multi-Program Workload	*Abbr.*	*Memory Footprints*
Parsec Four Program Workloads		
blackscholes-vips-streamcluster-swaptions	BVSS	L-L-L-L
canneal-x264-blackscholes-vips	CXBV	H-H-L-L
canneal-x264-freqmine-dedup	CXFD	H-H-H-H
raytrace-fluidanimate-freqmine-bodytrack	RFFB	L-L-H-H
SPEC-CPU2006 Four Program Workloads		
bzip2-zeusmp-cactusADM-bwaves	BZCB	H-H-H-H
bzip2-gobmk-hmmer-libquantum	BGHL	H-L-L-L
GemsFDTD-lbm-milc-namd	GLMN	H-M-M-L
lbm-perlbench-leslie3d-astar	LPLA	M-M-M-M
mcf-sjeng-cactusADM-calculix	MSCC	H-L-H-M
povray-h264ref-calculix-soplex	PHCS	L-L-M-M
SPEC-CPU2006 Eight Program Workloads		
bzip2-gcc-zeusmp-cactusADM-mcf-GemsFDTD-milc-soplex	HIGH8	H-H-H-H-H-H-M-M
gobmk-hmmer-h264ref-gromacs-namd-povray-tonto-libquantum	LOW8	L-L-L-L-L-L-L-L
gobmk-namd-lbm-perlbench-calculix-soplex-bzip2-gcc	MIX8	L-L-M-M-M-M-H-H

We used GreenChip to analyze the indifference points, IPC, energy, and MPKI of a mix of the Parsec [4] and SPEC-CPU2006 [5] multi-program workloads. The memory impact and specific benchmarks to construct the workloads is shown in Table I. The Parsec workloads are both multi-threaded and multi-program, while the SPEC workloads are single threaded and multi-program. Unfortunately, due to limitations in the simulation environment, the Parsec multi-threaded workloads could only be run on the four and eight core configurations, limiting their experiments to the 45nm and 28nm processors. The estimations of the individual benchmark memory impacts were taken from the literature [25, 26] and are listed in the order of the benchmarks. Multi-program workloads were selected to represent systems with several concurrent processes.

In our sensitivity analysis, we evaluate different usage scenarios with four activity and sleep ratios (see Section III) shown in Table II representing the load experienced by a cloud server (Server) that is typically online but often underloaded, a high-performance machine (HPC) that is typically constantly online and heavily loaded, a desktop machine (Desktop) that is used often, but lightly during the working day, and a mobile device (Mobile) that is mostly asleep, but when it wakes up is heavily loaded [14, 27, 28].

B. Results

The manufacturing costs of the different architecture choices are shown in Table III. Manufacturing cost is reported for 90-45nm by Boyd [14] and 28nm is determined from Apple Environmental Reports [11] and normalized to 45nm from Boyd. The manufacturing cost per technology node tends to increase for each generation due to the increase in manufacturing cost per area (see Fig. 2) and even though the area decreases significantly for the 28nm node, the increase in manufacturing cost per area still results in a dramatic jump in manufacturing energy.

TABLE II
ACTIVITY AND SLEEP SCENARIOS

Name	*Activity Ratio r_A*	*Sleep Ratio r_S*
Server	30%	5%
HPC	95%	5%
Desktop	17%	77%
Mobile	90%	92%

TABLE III
MANUFACTURING COSTS FOR CHIPS AT DIFFERENT PROCESS NODES FOLLOWING PRODUCT TRENDS (PSEUDO ISO-AREA) [11, 14].

Process Node (nm)	90	65	45	28
Core Count	1	2	4	8
LLC size	1MB	2MB	4MB	8MB
Area (mm^2)	207	227	207	158
Manufacturing Energy (MJ)	124	148	164	598

In contrast, the use phase energy, shown in Fig. 4, shows how increasing the core count and cache size can dramatically reduce use phase energy by a combination of increasing performance and savings in use phase power. However, the the appropriate environmental design choice requires a combination of both manufacturing and use phase energy trends.

The indifference analysis of a selected benchmark (GLMN) is shown in Fig. 5 to illustrate the design space, with the four scenarios, Server, HPC, Desktop, and Mobile, represented by circles in the top left, bottom left, top right, and bottom right regions of the figures, respectively. For both the Desktop and Server scenarios, the higher manufacturing cost of the smaller process node in the 90nm vs 65nm and 65nm vs 45nm comparisons is not recovered through use phase gains. In contrast, the higher performance of the smaller node in the HPC and Mobile scenarios results in the larger manufacturing energy being offset by use phase gains in less than 2 years, suggesting the more environmentally sound approach is to choose the smaller technology node. In the 45nm vs 28nm comparison, HPC and Server scenarios reach the indifference point in less than 2 years, while the Mobile and Desktop scenarios approach 10 years.

The break even comparison from Fig. 6 shows similar trends but with sharper gradients through the design space. Across the three node comparisons, the break even time for the Desktop scenario is larger than 7 years, and always larger than 5 years for the Mobile scenario. While the Server scenario never

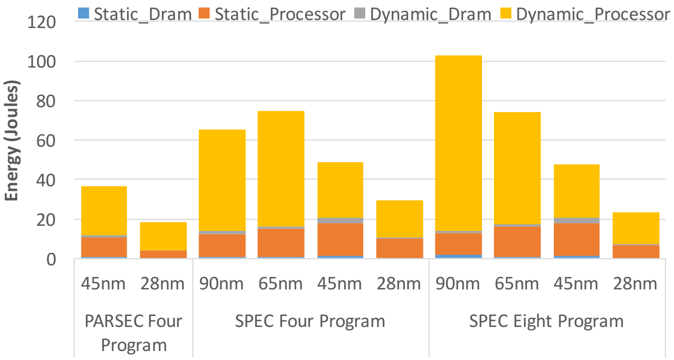
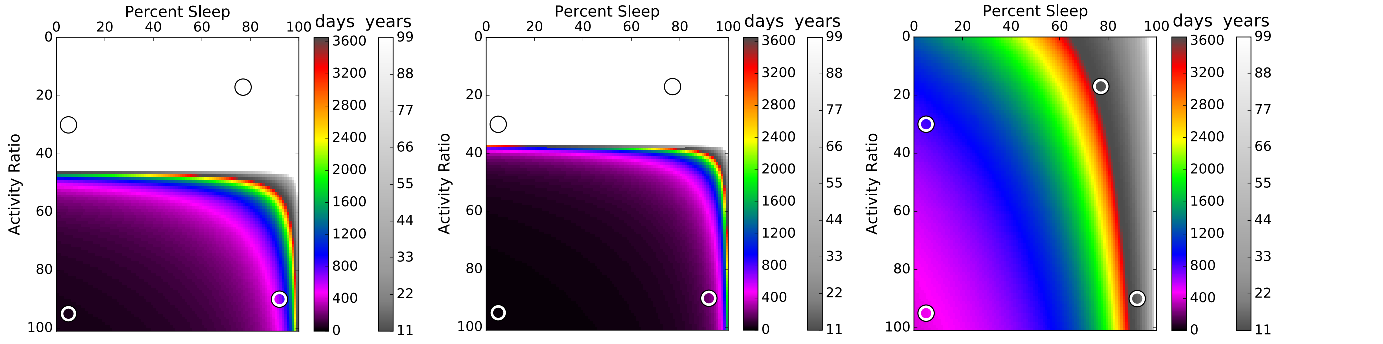


Fig. 4. Joules Per 10billion instructions for the Parsec and SPEC multiprogram workloads with different process node. All are run with the same chip area, as part of the iso-area comparison.

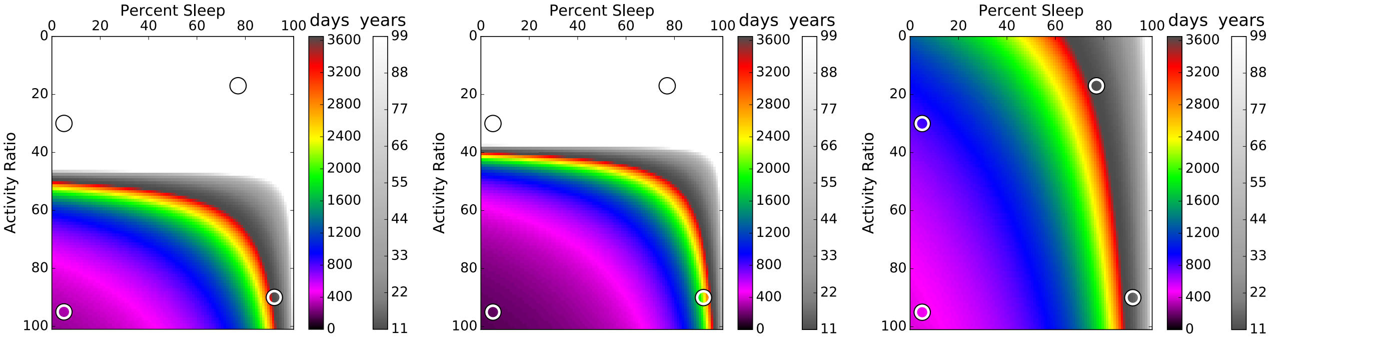


(a) 90nm vs 65nm

(b) 65nm vs 45nm

(c) 45nm vs 28nm

Fig. 5. Indifference points (t_I) for the pseudo ISO-area comparisons across adjacent technology nodes for GLMN highlighting four usage scenarios.



(a) 90nm vs 65nm

(b) 65nm vs 45nm

(c) 45nm vs 28nm

Fig. 6. Break even time (t_B) to move to the next technology node in a pseudo ISO-area comparison for GLMN highlighting four usage scenarios.

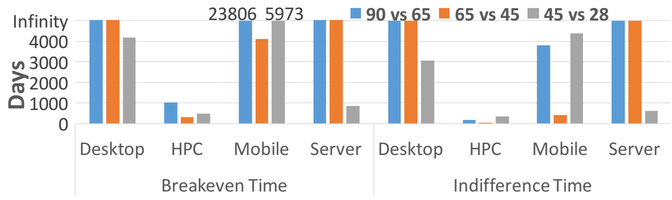


Fig. 7. Average break even times and indifference points across all benchmarks for pseudo iso-area comparison.

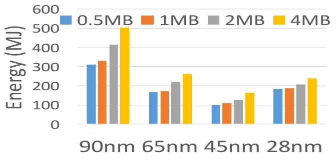


Fig. 8. Manufacturing energy for four-core systems with varying LLC capacities across technology nodes.

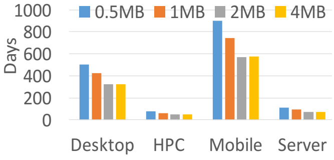


Fig. 9. Indifference time (t_I) between 45nm and 28nm for multiple LLC cache capacities.

breaks even for the 90→65 and 65→45 node comparisons, 45→28 breaks even after 3 years. Consistent with indifference analysis, the HPC scenario demonstrates that upgrading is the most sustainable decision, as long as the new device will be in use for at least a year.

To achieve a more global view, the average indifference points and break even times are shown in Fig. 7 for the four scenarios. The results follow similar trends as shown in GLMN example, with HPC systems typically pushing toward the new technology node quickly, Desktop, Server, and Mobile typically not pushing toward the new technology node with the exception of purchasing 45nm mobile and 28nm servers over 65nm and 45nm, respectively if within a 3-year usage time.

V. Case Study II: Sensitivity Analysis of the Impact of Cache Sizes on Sustainability

One common architecture configuration option is to change the size of the last level cache. In this case study we fix the processor into a four-core system and vary the LLC capacity from 0.5MB to 4MB and examine the impact on sustainability. The manufacturing cost of varying the LLC capacity is shown in Fig. 8. For all technology nodes, the increase of capacity is met with a significant increase in manufacturing cost, which attenuates as the feature size is reduced.

Considering the same scenarios and workloads described in Tables I and II, the indifference point analysis always selected the smaller of the two technology nodes regardless of scenario due to the reduction in manufacturing and use phase cost, with the exception of the 45nm to 28nm transition. This is due to the iso-architecture comparison, where the larger technology node areas are much higher, rather than the more realistic pseudo iso-area comparison from Section IV. Interestingly, the 45nm to 28nm indifference points (Fig. 9) vary widely by cache size, trending to become smaller as the LLC capacity increases within each usage scenario.

Moreover, this trend differs from the break even study in Fig. 10. The Desktop scenario never breaks even from 90nm→65nm at any cache size. For the Desktop 65nm→45nm comparisons, mobile 90nm→65nm and 45nm→28nm, and server 90nm→65nm comparisons, the break even times all

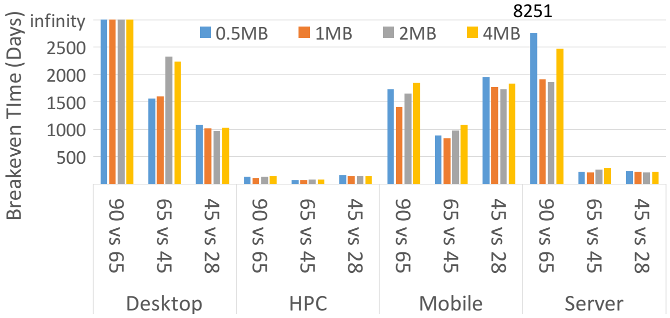


Fig. 10. Average break even times across all the benchmarks, iso-architecture comparison with 4 cores. Note: All 90nm vs. 65nm data points for desktop except one benchmark never broke even. Also, one benchmark for the server at 0.5MB never broke even, so the average is the remainder of the benchmarks

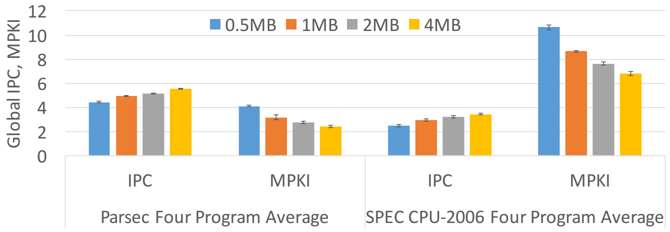


Fig. 11. Global IPC and MPKI averaged for Parsec and SPEC four process multi-program workloads across different technology nodes while varying LLC capacity. (min and max shown by error bars)

exceed four years. The Desktop 45nm→28nm and mobile 65nm→45nm comparisons are both around three years. Finally, the entire HPC scenario and Server 65nm→45nm and 45nm→28nm transitions all break even in less than one year.

To better understand these results we examined the performance in instructions per cycle (IPC). In these experiments, the IPC (Fig. 11) stays relatively constant across technology nodes but has varying effects for the different multi-program workloads; for example for RFFB the additional LLC capacity does not noticeably improve performance, while for the other workloads the IPC steadily improves as the capacity increases. Also, as expected, the misses per kilo-instructions (MPKI) decreases as the LLC capacity increases. The change from 0.5MB to 1MB has the largest MPKI decrease with larger LLC capacities having limited additional improvements.

The energy fluctuation (Fig. 12) for the different cache sizes within a workload and technology node depends on the trade off of additional performance from the larger LLCs against the increase in static power as the cache size increases. For example, CXFD at 28nm experienced sufficient performance benefits from increasing the LLC size to offset the static power increase resulting in a reduction in energy. In contrast, RFFB at 45nm experiences the opposite trend, as the static power increase offsets the nominal performance gains as the cache size increases. On average, there is a reduction in energy from a 0.5MB to 1MB LLC but for larger LLC sizes, the energy remains relatively consistent with the performance trends.

These trends point to 1MB caches providing the best trade off between performance, energy, and manufacturing cost. This is also supported by the breakeven results, with the 1MB LLC typically among the lowest break even times.

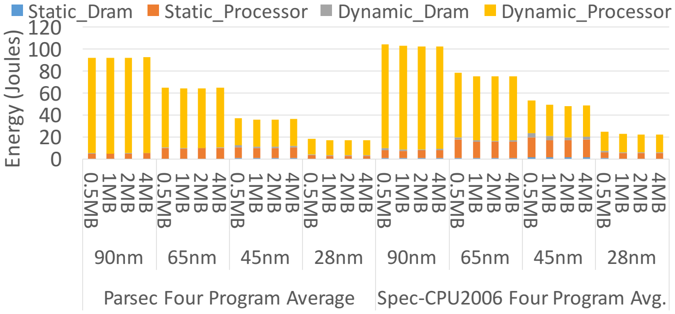


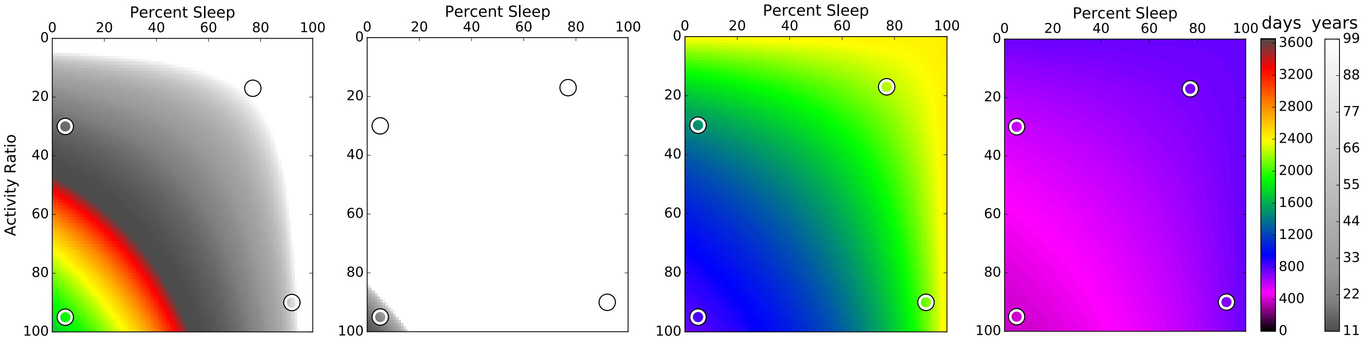
Fig. 12. Joules Per 10 billion instructions averaged for Parsec and SPEC four process multi-program workloads across different technology nodes while varying LLC capacity.

VI. Case Study III: Impact of Main Memory Density on Sustainability

As the number of cores and cache densities in modern processors increase, there is an increased trend toward larger workloads that place pressure on main memory. We show a simple analysis of the memory capacity impact for different technology nodes shown in Figs. 13(a)-13(d), including the power component of a fixed size Flash SSD that stores data, which spills from main memory in each scenario. The analysis was performed on the SPEC OpenMP benchmark (OMP2012) [6] “Swim,” which has about 6x the memory footprint of the largest SPEC CPU benchmark (around 6.2GB, in our own investigation). This is near the median of the OMP2012 benchmark footprints. 2GB and 4GB were selected as possible memory allotments for a virtual machine (VM) running the application on a server with multiple VMs and is likely provisioned with much larger amounts of total memory.

The first memory and disk result from our tool can be observed in Fig. 13(a) showing the 4GB and 2GB DRAM indifference points at the 65nm technology node. The indifference time for the server, desktop, and mobile scenarios all require more than 10 years of use for the 4GB memory to be more sustainable than the 2GB memory. Even for the HPC scenario, slightly more than 5 years is required before the additional memory overcomes its manufacturing and static power deficits. An almost identical trend is observed for 4GB vs. 2GB DRAM indifference comparison at the 55nm node but offset in the favor of 2GB. The HPC scenario now requires more than 10 years and the remaining scenarios require more than 100 years (i.e., essentially ∞) before the 4GB DRAM is more sustainable. In general, from a sustainability perspective, choosing 2GB of memory for workloads resembling Swim is better than choosing 4GB.

The indifference comparison between 65nm and 55nm for 2GB [Fig. 13(c)] indicates running times of 6.5, 5.5, 4.5, and 2 years are required before the 55nm is the sustainable choice for Desktop, Mobile, Server, and HPC scenarios, respectively. For the 4GB comparison, all scenarios reach the indifference point in less than 2.5 years, however, due to the intra-node indifference comparison of different memory sizes [Figs. 13(a) and 13(b)], 4GB was already not favorable compared to 2GB.



(a) 65nm Dram, 4GB vs 2GB (b) 55nm Dram, 4GB vs 2GB (c) 65nm vs 55nm Dram, 2GB (d) 65nm vs 55nm Dram, 4GB

Fig. 13. Memory indifference points for the SPEC OMP-2012 Benchmark "Swim."

VII. Conclusion

In this paper we present a holistic sustainability evaluation and prediction tool called GreenChip. GreenChip allows detailed manufacturing and use phase energy calculation and comparison of integrated circuits used for constructing computing systems from processors, main memory, and solid state storage. We presented several case studies that evaluate processor, cache, and main memory choices. In many cases, indifference and break even times can be compared with typical expected lifetimes. For example in Case Study I (Section IV), the break even points for upgrading desktop computers and mobile devices often exceeded five years and replacement cycles for such system is often less than two years. Also, it often did not make sense to upgrade servers even when the use phase gain was particularly helpful, recalling that the 45nm→28nm upgrade time still exceeded three years.

One interesting trend is that chasing higher core counts, caches, and memory/storage sizes may not always be the most sustainable solution, and there is potential with reaching fabrication technology limits for manufacturing cost to become an increasingly important factor in design choices. For example in Case Study II (Section V) the results pointed to a moderate LLC capacity (1MB) providing the best compromise of sustainability and performance. Additionally, considering main memory sizes, a moderate main memory can outperform a larger main memory from an environmental perspective as shown in Case Study III (Section VI).

GreenChip provides a flow to evaluate many future design choices for holistic sustainability such as server consolidation with larger core counts and memory capacity. Incorporating more holistic evaluations into standards such as Energy Star and presenting sustainability metrics for consumer electronics can empower consumers to make more informed choices and lead to new marketing strategies resulting in a more sustainable computing electronics industry.

References

[1] A. Jones, L. Liao, W. Collinge, H. Xu, L. Schaefer, A. Landis, and M. Bilec, "Green computing: A life cycle perspective," in *IGCC*, 2013.

[2] A. Jones, Y. Chen, W. Collinge, H. Xu, L. Schaefer, A. Landis, and M. Bilec, "Considering fabrication in sustainable computing," in *ICCAD*, 2013.

[3] M. J. Scott and E. K. Antonsson, "Using indifference points in engineering decisions," in *Proc. of ASME Des. Eng. Tech. Confs.*, 2000.

[4] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The parsec benchmark suite: Characterization and architectural implications," in *PACT*, 2008.

[5] J. L. Henning, "SPEC CPU2006 benchmark descriptions," *SIGARCH Comput. Archit. News*, vol. 34, no. 4, 2006.

[6] M. S. Müller *et al.*, "SPEC OMP2012 — an application benchmark suite for parallel systems using OpenMP," in *IWOMP*, 2012.

[7] ISO, "Environmental management – life cycle assessment – requirements and guidelines," Tech. Rep. 14044, 2006.

[8] UNEP/SETAC, "Life cycle approaches: The road from analysis to practice," Tech. Rep., 2005.

[9] "U.S. life cycle inventory database," 2012. [Online]. Available: https://www.lcacommons.gov/nrel/search

[10] J. Bare, "Traci 2.0: the tool for the reduction and assessment of chemical and other environmental impacts 2.0," *Clean Technologies and Environmental Policy*, vol. 13, no. 5, 2011.

[11] Apple Inc., "Environmental report," [Available Online]: http://www.apple.com/environment/reports/, 2015.

[12] M. A. Yao, T. G. Higgs, M. J. Cullen, S. Stewart, and T. A. Brady, "Comparative assessment of life cycle assessment methods used for personal computers." *Env. Sci. & Tech.*, vol. 44, no. 19, 2010.

[13] P. Teehan and M. Kandlikar, "Comparing embodied greenhouse gas emissions of modern computing and electronics products," *Environmental Science & Technology*, vol. 47, no. 9, 2013.

[14] S. B. Boyd, *Life-Cycle Assessment of Semiconductors*. Springer, 2012.

[15] C. F. Murphy, G. A. Kenig, D. T. Allen, J.-P. Laurent, and D. E. Dyer, "Development of parametric material, energy, and emission inventories for wafer fabrication in the semiconductor industry," *Env. Sci. & Tech.*, vol. 37, no. 23, 2003.

[16] SIA, "Model for assessment of cmos technologies and roadmaps," 2007. [Online]. Available: http://www.itrs.net/models.html

[17] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *MICRO*, 2009.

[18] M. Neisser and S. Wurm, "ITRS lithography roadmap: 2015 challenges," *Advanced Optical Technologies*, vol. 4, no. 4, 2015.

[19] R. I. Bahar, A. K. Jones, S. Katkoori, P. H. Madden, D. Marculescu, and I. L. Markov, "Workshops on extreme scale design automation," CCC, Tech. Rep., 2014. [Online]. Available: http://www.cra.org/ccc

[20] T. E. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *SC*, 2011.

[21] N. Binkert, B. Beckmann *et al.*, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, 2011.

[22] N. P. Jouppi and S. J. Wilton, "An enhanced access and cycle time model for on-chip caches," Compaq, Tech. Rep. TR-93-5, 1994.

[23] U.S. Energy Information Administration (EIA), "International energy statistics," [Accessed May 2016].

[24] P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "Dramsim2: A cycle accurate memory system simulator," *IEEE Comp. Arch. Let.*, vol. 10, no. 1, 2011.

[25] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The parsec benchmark suite: Characterization and architectural implications," Princeton University Technical Report TR-811-08, January 2008, Tech. Rep.

[26] J. L. Henning, "Spec cpu2006 memory footprint," *SIGARCH Comput. Archit. News*, vol. 35, no. 1, 2007.

[27] S. M. Pieper, J. M. Paul, and M. J. Schulte, "A new era of performance evaluation," *IEEE Computer*, vol. 40, no. 9, pp. 23–30, 2007.

[28] J. Suckling and J. Lee, "Redefining scope: the true environmental impact of smartphones?" *Intl. Jour. of Life Cycle Assess.*, vol. 20, no. 8, 2015.