Energy-Driven Computing

Sivert T. Sliper, Oktay Cetinkaya, Alex S. Weddell, Bashir Al-Hashimi and Geoff V. Merrett

Centre for IoT and Pervasive Systems, School of Electronics and Computer Science, University of Southampton, United Kingdom

For decades, the design of untethered devices has been focused on delivering a fixed quality of service with minimum power consumption, to enable battery-powered devices with reasonably long deployment lifetime. However, to realise the promised tens of billions of connected devices in the Internet of Things, computers must operate autonomously and harvest ambient energy to avoid the cost and maintenance requirements imposed by mains- or battery-powered operation. But harvested power typically fluctuates, often unpredictably, and with large temporal and spatial variability. Energy-driven computers are designed to treat energy-availability as a first-class citizen, in order to gracefully adapt to the dynamics of energy harvesting. They may sleep through periods of no energy, endure periods of scarce energy, and capitalise on periods of ample energy. In this paper, we describe the promise and limitations of energy-driven computing, with an emphasis on intermittent operation.
1. Introduction

The Internet of Things (IoT) is driving the proliferation of low-power computing devices, bridging the cyber- and physical worlds with advanced monitoring, control, and management capabilities [1]. These devices are extensively employed in a variety of domains, such as smart buildings and cities, health monitoring, remote surveillance, industrial automation, and wearable computing [2]. However, despite gaining momentum, the services provided in the IoT context have some inevitable drawbacks.

The number of connected devices is projected to be 22 billion by 2024\(^1\). At this scale, maintenance operations, such as battery replacement and/or replenishment, are likely to become practically and economically infeasible. Due to the profusion of devices, IoT services will seek very low cost and compact solutions, i.e. sub-dollar and sub-cm\(^3\), capable of operating for decades without requiring maintenance, while also minimising environmental impact.

The architectural objectives for a computing device that can enable an IoT comprised of tens of billions of connected devices can be listed as: long maintenance-free deployment lifetime, low cost, small size, and minimal environmental impact. Figure 1 illustrates the potential of power-provisioning strategies in the context of the aforementioned architectural objectives. The following text introduces the three power-provisioning strategies, and qualitatively evaluates them against the axes of Fig. 1.

Non-rechargeable batteries can provide for computing devices with high quality of service (QoS)\(^2\), but impose substantial increases in cost and volume, and limit deployment lifetime; furthermore, their eventual replacement and disposal imposes maintenance cost and environmental impact (“battery-powered” in Fig. 1). A promising augmentation to battery-power is energy harvesting (EH), which is the conversion of ambient energy into usable electrical power. Ambient energy sources are practically inexhaustible, so using EH in conjunction with rechargeable batteries improves the longevity of devices by supplying energy during deployment, and thus mitigates the need for periodic maintenance. However, as shown in Table 1, EH typically only offers low power density, and the output is not only highly variable, but also uncontrollable, and in many cases unpredictable [3].

The dynamics of EH motivate a shift towards energy-driven computing [4], where energy availability, and the energy environment, is a fundamental component of the system and application design process. In other words, energy-driven computers must not only consider application requirements but also the dynamics of their energy harvester and environment. For example, an energy-driven computer could dynamically adapt in order to provide

\(^1\)Number of devices connected to the Internet, excluding PCs, laptops, tablets and mobile phones. Source: https://www.ericsson.com/en/mobility-report/mobility-visualizer

\(^2\)In this paper we regard QoS qualitatively as delivering accurate results on time and/or frequently. Lower QoS could entail reduced accuracy, reduced throughput and/or increased latency.
Table 1: Representative power density range of tiny energy harvesters.

<table>
<thead>
<tr>
<th>EH Category</th>
<th>Energy Source</th>
<th>Power Density</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal</td>
<td>Heat</td>
<td>10μW/cm² - 1mW/cm² (Thermoelectric)</td>
<td>[5] [6,7]</td>
</tr>
<tr>
<td>Fluid Flow</td>
<td>Wind</td>
<td>100mW/cm² @ 2 ~ 9m/s, airflow</td>
<td>[5]</td>
</tr>
<tr>
<td></td>
<td>Hydro</td>
<td>30.67μW/cm² (0.35ml @3.43m/s)</td>
<td></td>
</tr>
<tr>
<td>Radiated</td>
<td>RF</td>
<td>0.2nW/cm²-1μW/cm²</td>
<td>[6,8]</td>
</tr>
<tr>
<td></td>
<td>Light</td>
<td>100mW/cm² (Outdoor, solar)</td>
<td>[9,10]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100μW/cm² (Office, diffused light)</td>
<td></td>
</tr>
<tr>
<td>Kinetic</td>
<td>Vibration</td>
<td>200μW/cm³ (Piezoelectrics)</td>
<td>[6,9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~800μW/cm³ (Machines - kHz)</td>
<td></td>
</tr>
<tr>
<td>Biochemical</td>
<td>Glucose</td>
<td>180μW/cm²</td>
<td>[11,12]</td>
</tr>
<tr>
<td></td>
<td>Lactate</td>
<td>44μW/cm²</td>
<td></td>
</tr>
<tr>
<td>Human Body</td>
<td>Blood pressure</td>
<td>6μJ/cm³</td>
<td></td>
</tr>
<tr>
<td>Biomechanical</td>
<td>Heartbeat</td>
<td>1.2μW/cm²</td>
<td>[9,13]</td>
</tr>
<tr>
<td></td>
<td>Breathing</td>
<td>N/A</td>
<td>[11,12]</td>
</tr>
<tr>
<td></td>
<td>Locomotion</td>
<td>~4μW/cm³</td>
<td></td>
</tr>
</tbody>
</table>

maximum achievable QoS for the energy available during deployment, as opposed to a traditional application-driven computer, which aims to minimise overall power consumption for a fixed QoS requirement.

Current energy-driven computers have different mechanisms for managing the dynamics of EH, which can be categorised into three groups: energy-neutral, power-neutral, and intermittent operation.

Energy-neutral operation [3] copes with the dynamics of EH by using energy storage, i.e. a rechargeable battery or a supercapacitor, that act as a buffer between energy supply and demand. Compared to their battery-powered counterparts, energy-neutral computers can maintain high QoS, while reducing the necessary storage capacity (and therefore cost/size), but are still limited by their use of rechargeable batteries or supercapacitors which impose cost and volume, and eventually expire due to limited charge cycles and/or ageing.

In contrast, power-neutral operation [4] seeks to cope with the dynamics of EH by adapting power consumption to instantaneously match harvested power, reducing, and potentially eliminating, the need for energy storage. However, without storage, volatile application state is lost when the harvested power drops too low; when power returns, the application has to be restarted from the beginning. Consequently, power-neutral operation often cannot sustain long-running applications.

Intermittent operation is an orthogonal approach that does not attempt to hide the dynamics of EH, but rather accepts that power failures can be frequent and unpredictable [14]. By retaining computational progress through power cycles, intermittent computing systems (ICSs) allow long-running applications to progress incrementally whenever energy is available. The key merit of intermittent operation is that it can enable maintenance-free operation and theoretically unbounded lifetime at very low cost and volume, and with little environmental impact. But QoS must often be sacrificed because ICSs are not continuously operational (by definition). Hence,
intermittent operation is only applicable to certain applications. However, in the ideal case, where the harvested power directly coincides with the event of interest and is sufficiently powerful such that the system is guaranteed power during the events, the QoS may match that of a battery-powered system (see Section 3(a)).

Energy-driven computing has, in recent years, seen increasing research interest spanning areas such as programming languages and compiler techniques [15–17], embedded systems [18–22], hardware-software codesign [23], nonvolatile processors [24] and asynchronous circuits [25]. This paper outlines the state-of-the-art of energy-driven computing, with an emphasis on intermittent operation, and points out open research challenges. The contributions of this paper are:

- Our conditions for intermittent operation provides high-level answers to the question “Which applications are amenable to intermittent operation?”.
- A systematic taxonomy of intermittent computing (IC) methods that divides prior works into three classes, each evaluated against our set of correctness requirements, performance objectives and scalability objectives.
- An overview of unsolved challenges in energy-driven computing, with a particular focus on issues pertaining to communication under intermittent operation.

The remainder of this paper is organized as follows. First, energy- and power-neutral computing is discussed in Section 2. Then, in Section 3, conditions, correctness criteria, performance objectives and scalability objectives for intermittent operation are presented, followed by a taxonomy of prior works. We then extend our study to discuss unsolved challenges in energy-driven computing, in Section 4.

2. Energy- and power-neutral computing

This section describes two possible ways of operating a computer under an unstable power supply. The first, energy-neutral operation [3], seeks to hide the dynamics of EH by buffering energy, and potentially adapting application behaviour, to balance energy supply and demand over a long period. The second alternative, power-neutral operation, modulates power consumption to instantaneously match production, ideally obviating the need for energy storage.

Both methods may exploit trade-offs between QoS and power consumption by modulating their operating frequency, activating and deactivating peripherals, selectively sampling sensors, and/or using techniques such as approximate computing [26–29].

(a) Energy-neutral computing

Energy-neutral systems cope with an unstable power supply by storing energy in a battery, or a supercapacitor, when harvesting conditions are good [3]. During periods of scarce energy, the stored energy sustains operation until conditions improve.

Fundamentally, energy-neutral operation strives to balance energy consumption and harvesting over a long period $T$, related to the periodicity of the ambient source [4]:

$$\int_{(n-1)T}^{nT} P_h(t) dt = \int_{(n-1)T}^{nT} P_c(t) dt, \quad (2.1)$$

where $P_h(t)$ and $P_c(t)$ is the power harvested and consumed, respectively, at time $t$.

As an example, an outdoor solar-powered device may have a period of $T = 24$ hours, and thus can satisfy (2.1) by ensuring that its energy consumption matches the harvested energy over the past 24 hours.

Modulating the average power consumption over a long period is typically done through duty cycling or other means of exploiting power-QoS trade-offs. For example, a node can wake up from sleep after a period of duration $p$, quickly take sensor measurements, perform its computational
tasks, and go back to sleep again. The period $p$ can then be adjusted according to harvesting conditions.

A representative and recent example of a wireless sensor node that can be operated in energy-neutral mode is Permamote [20], shown in Fig. 2. An energy harvesting front-end that includes maximum power point tracking (MPPT), charging and battery monitoring circuitry, selects between a rechargeable energy store and a non-rechargeable battery (used when the rechargeable energy store is depleted) to power the system. A switching regulator then boosts the battery voltage for the system load (microcontroller, radio and sensors). The backup battery used in Permamote is superfluous when successfully operating in energy-neutral mode, but since EH is so unpredictable (both in long-term and short-term temporally, in addition to spatially), it is added as a backup to provide a minimum lifetime where the harvesting conditions are worse than those expected at design time.

Although energy-neutral systems such as Permamote can provide high reliability and QoS with potentially decades of deployment lifetime, the extra conversion and storage circuitry adds cost, mass, volume and complexity; all of which can be prohibitive for deployment at scale. Additionally, the energy loss through the conversion circuitry degrades energy efficiency: the simulator for Permamote used 80% in each of the two conversion steps (Fig. 2), leading to an overall energy efficiency of 64%, before accounting for leakage currents and energy spent in sleep-modes [20,30]. Greater end-to-end efficiency is possible for specialised circuits, e.g. 82% for a reciprocal converter with selective direct operation that tailors to the specific characteristics of subthreshold CMOS processors and tiny energy harvesters [31].

(b) Power-neutral computing

Instead of relying on energy storage and conversion to smooth out an unstable source of power, power-neutral operation aims to instantaneously match power consumption to harvested power so that the system load can be connected directly to the energy harvester [4]. Theoretically, an ideal power-neutral device would scale power consumption such that

$$P_c(t) = P_h(t),$$

and thus, no energy storage would be necessary. Power-neutral operation thus promises to both reduce complexity and improve energy efficiency by removing the storage element and the associated charging, monitoring and conversion circuitry. Modulating the power consumption is typically done through dynamic performance scaling (e.g. dynamic voltage and/or frequency scaling) [4,27–29], although it could potentially also employ other techniques such as approximate computing.
However, real computational devices have fixed operating points in terms of power and performance, and switching between them takes time, so real power-neutral devices still have to obey (2.1), albeit over a much smaller period $T$ (typically on the order of milliseconds). Therefore, some capacitance is needed to give the system sufficient time to react, and switch to another operating point, when the power level changes. The necessary capacitance is dependent on the granularity of operating points, and the latency to switch between them. Typically, in microcontroller based systems, the capacitance that is already present for other purposes, such as decoupling, suffices; for more high-performance systems, a moderate amount of capacitance becomes necessary. In Momentum [28], power-neutral operation was demonstrated on a microcontroller based system with no extra capacitance, and on a multiprocessor single-board computer with an additional $47 \, \text{mF}$ capacitor.

A limitation of power-neutral operation becomes apparent if temporal application requirements vary independently of the harvested power. In this case, the application may have high performance demands in a period where harvested power is low, and so QoS can be severely degraded. Similarly, if the harvested power is ample in a period with low application demands, a large fraction of the harvested power becomes superfluous. This stands in contrast to energy-neutral systems, where harvested power and application demands are temporally decoupled.

3. Intermittent computing systems

Both energy- and power-neutral systems can survive through moderately long periods (relative to the size of their energy buffer) of scarce power-availability by utilizing low-power sleep modes. But, if the harvested power remains lower than the current consumption in their lowest-power sleep mode for an extended period, they will eventually brown out. Consequently, the volatile state is lost, and the system has to recharge, reboot, and then restart the application from the beginning; potentially losing important data, or worse: leaving a corrupt state in nonvolatile memory.

Intermittent computing (IC) solves this problem by retaining computational progress through power cycles, thus allowing long-running applications to progress incrementally whenever energy is available. IC is typically used stand-alone or in conjunction with power-neutral operation [28,29].

This section begins by defining the conditions under which an application is amenable to intermittent operation, then presents consistency requirements, performance objectives, and scalability objectives. Then, a taxonomy divides published works into three classes of IC and evaluates them against the aforementioned requirements and objectives.

(a) Conditions for intermittent operation

Because intermittent operation does not, by definition, sustain continuous operation, an application that is to be executed intermittently must:

1. not depend on absolute time, unless it can be gathered from an external source (e.g. via wireless communication);
2. be guaranteed sufficient power during events to be detected, or permit some lost events;
3. be delay-insensitive or amenable to low-precision computing during periods of scarce energy, or accept lost data points due to a lack of processing power.

A typical application of an ICS is to sense the environment, detect events, and report them to a central sink/server [32,33]. Without a consistent source of power, an ICS cannot maintain an absolute sense of time; it can only keep track of time within an on-period. Note that all three conditions typically entail unreliable and/or reduced QoS when compared to a battery-powered or energy-neutral system.
Recent works have proposed methods that can maintain an approximate sense of time through short off-periods by measuring the decay of a charged capacitor or the gradual information loss in volatile memory [34], but they are far too inaccurate for synchronizing wireless communication (3.7-8.7% error over a period of 35 seconds). Wireless communication between ICSs is, therefore, a major challenge, and still an unsolved problem, so today’s ICSs generally communicate with an always-on server. Thus, the central server only needs to be notified when an event happens, and can itself annotate absolute timing. An ideal example is the wireless bicycle trip counter by Bing et al. [22], which counts and reports the revolutions of a bicycle wheel. The device itself is powered by the current induced when a magnet attached to the bicycle wheel passes a coil mounted on the bicycle’s fork, hence the event to be detected directly coincides with the power source. By use of intermittent operation, their sensor node is made with a minimum of components, leading to small size and cost without introducing maintenance requirements or sacrificing lifetime. Other potentially suitable applications for ICSs include predictive maintenance on industrial equipment via vibrational and acoustical monitoring [35]; energy for the system could be supplied by a piezoelectric vibration harvester.

(b) Correctness requirements, performance objectives and scalability objectives

For systematic evaluation of approaches to intermittent operation, this paper will use the following correctness requirements (C1-C4), performance objectives (P1-P3), and scalability objectives (S1-S3). Correctness requirements must be met for an ICS to function as expected. Performance and scalability objectives are ideals that should be sought after, but are not critical —and will inevitably lead to compromises. The requirements and objectives are inspired by the ones introduced in Alpaca [16], but have been generalised and expanded in order to include recent progress and insights in the field [32,33,36–39].

For correctness, an ICS must:
- C1 retain computational progress through power loss;
- C2 ensure consistency between volatile and nonvolatile memory despite power loss;
- C3 allow blocks of code to be executed atomically within a single on-period;
- C4 avoid sisyphean tasks.

For performance, i.e. to maximise QoS, an ICS should:
- P1 maximise forward progress through applications;
- P2 be reactive to external events;
- P3 minimise extra memory footprint imposed by ICS framework.

For scalability and to accommodate widespread adoption, an ICS should:
- S1 require minimal ICS-specific expertise from the programmer;
- S2 be compatible with existing software;
- S3 be portable across hardware platforms.

C1 is the base premise of intermittent operation. In certain situations, arbitrary sections of code can be re-executed because the system has to restore to an earlier state after a power failure. If the re-executed code is not idempotent, re-execution can lead to a violation of C2 [14]; i.e. an inconsistency between volatile memory, which loses its state when power fails, and nonvolatile memory, which does not. C3 enables the programmer to specify sections of code which must be executed atomically in the same on-period. This is necessary for operations that must be restarted entirely if interrupted by a power failure (of unknown duration). Examples include sampling

4In this context, a section of code is idempotent if it can be interrupted at any point, restarted from the beginning, and still produce the same result.

5In this context, a function which is executed atomically will be executed in its entirety without power interruptions. If it is interrupted, it must be restarted from the beginning.
temporally correlated sensors, using on-chip peripherals, serial communication with peripherals, and transmitting radio packets [33,36,38].

A *Sisyphean task* is a task that requires more energy than the device can muster, hence it will be restarted indefinitely, rendering the ICS non-functional [14]. The requirement to avoid sisyphean tasks (C4) follows from C3: if certain tasks are to be executed atomically, it is imperative that they can be completed within a power cycle.

P1 implies that an ICS should minimise the time spent on state retention, boot, and rollback. To remain reactive to external events (P2), an ICS should be able to handle asynchronous interrupts, and should generally wake up frequently so as to not miss an event while recharging for the next on-period. Note that P1 and P2 often conflict with each other. To optimise for P1 means that reboots are rare, so that boot and rollback overheads are amortised over a long period of application-execution. On the contrary, optimising for P2 means frequent reboots to ensure that no events are missed. In addition to time and energy overheads, IC methods generally impose extra memory footprint to, for example, store snapshots of system state, or to instrument application code with checkpoints; this extra footprint should be minimised, because the ICSs typically have strict memory constraints (P3).

(c) Taxonomy

In order to differentiate and evaluate existing works, this section presents a taxonomy based on the following three classes of intermittent computing methods [40]:

**Static IC:** where snapshots are saved at predetermined (design-time or compile-time) checkpoints in the program.

**Task-based IC:** where the application is divided into small tasks that are executed atomically by a runtime.

**Reactive IC:** where snapshots are saved when a power failure is detected.

The fundamental divergence is between static and reactive IC; task-based IC can be regarded as a recent development of static IC, where checkpoints are replaced by task boundaries. Static and reactive IC typically retain progress through power cycles by periodically (static) or reactively (reactive) saving the volatile state as a snapshot in nonvolatile memory. Task-based IC needs, in principle, only to save the result from a task’s execution and an indicator to the next task to be completed.

State diagrams for the three classes are shown in Fig. 3; these diagrams show the fundamental states, although most methods diverge slightly —typically by having additional states and edges to improve performance. A key insight is that both static and task-based methods have power-off edges (red) from all states, meaning that the system could potentially instantaneously shut down at any point in the program/runtime. Reactive IC (Fig. 3c), on the other hand, ensures that the system only powers off after reaching a safe sleep state.

Figure 4 illustrates the behaviour of each class in relation to the power supply voltage. Static IC restores (R) as soon as the supply voltage, $v_{cc}$, exceeds the on-threshold of the processor, $V_{on}$, and starts useful computation (C) while checkpointing (CP) state according to heuristics, for example every few milliseconds. Task-based IC also restores (boot runtime and restart the latest task) as soon as $v_{cc} > V_{on}$. The *restore* operation is typically quick because only the runtime and a single task need to be booted, not the entire program. When complete, the results of the current tasks are saved, and the next task is booted and started; this is termed a task-transition (TT). Reactive IC sleeps until an interrupt triggers *restore*, when $v_{cc}$ exceeds the restore threshold $V_R$. Reactive IC does not take checkpoints along the way, but rather continues useful computation until $v_{cc}$ drops below $V_S$, the suspend threshold, which triggers *suspend* (S), saving a snapshot just before power is lost. In the literature, this is also referred to as Just-In-Time checkpointing [33]. The circles show the latest snapshot before power is lost; any computation after the latest snapshot is a waste of resources. Note that some task-based or static approaches may mitigate wasted computation after a checkpoint by measuring stored energy and comparing it to predictions [21] or observations [41] of the energy required to reach the next task-boundary or checkpoint.
Because static and task-based methods place checkpoints or task-boundaries a priori, they risk encountering a *sisyphean task* [42]. This is where the section between checkpoints or task-boundaries is too large to be completed within a single power cycle, and thus it gets restarted forever. Reactive IC can typically avoid sisyphean tasks by proper use of voltage thresholds and sufficient capacitance (see Section 3(c)iii).

The following subsections analyse the three classes of IC in the context of the aforementioned correctness requirements, performance objectives and scalability objectives. Table 2 summarises our findings.
Table 2: Summarised analysis of the three classes of intermittent computing.

<table>
<thead>
<tr>
<th>Objective/Requirement</th>
<th>Static IC</th>
<th>Task-based IC</th>
<th>Reactive IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1: Comp. progress</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>C2: Consistent memory</td>
<td>Challenging</td>
<td>Challenging</td>
<td>✓</td>
</tr>
<tr>
<td>C3: Atomic sections</td>
<td>✓</td>
<td>✓</td>
<td>Challenging</td>
</tr>
<tr>
<td>C4: Avoid sisyph. tasks</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>P1: Forward progress</td>
<td>Slower¹</td>
<td>Slower¹</td>
<td>Faster¹</td>
</tr>
<tr>
<td>P2: Reactive to events</td>
<td>Challenging</td>
<td>Challenging</td>
<td>✓</td>
</tr>
<tr>
<td>P3: Memory overhead</td>
<td>Instr.², checkpoint</td>
<td>Instr.², runtime, logs, tiny checkpoint</td>
<td>Checkpoint</td>
</tr>
<tr>
<td>S1: Required ICS expertise</td>
<td>Moderate</td>
<td>Major</td>
<td>Minor</td>
</tr>
<tr>
<td>S2: Software compatibility</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>S3: Hardware portable</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>

¹ Depends on specific implementation. ² Code instrumentation. ³ With full recompilation.

(i) Static IC

Static IC approaches are based on instrumenting application code with checkpoints by the user during application development, or automatically at compile time [21, 42–44]. Inserting checkpoints at compile time can alleviate the burden of porting existing software to run intermittently, if source code is available (S2). Because the checkpoints are inserted a priori, offline analysis of program execution and control flow graphs can be applied to guarantee memory consistency (C2) and improve performance (P1–P3) [21, 43, 45, 46]. However, static analysis is greatly complicated if asynchronous interrupts are involved (P2).

The main challenges to static IC are in the placement of checkpoints: placing them too far apart minimises superfluous checkpoints, but makes it unlikely or even impossible to reach the next checkpoint (C4); placing them too densely wastes time and energy on superfluous checkpoints (P1). The optimal checkpoint placement depends both on hardware and on harvesting conditions, making solutions unlikely to be portable across hardware platforms (S3), and requiring a moderate degree of ICS expertise (S1).

Static IC also suffers from frequent code re-execution because the state rolls back to the latest snapshot when power returns after a power failure (the computation after the circle in Fig. 4 is re-executed in the next power cycle). Code re-execution is a waste of energy (P1), and can corrupt memory if the re-executed section of code is not idempotent (C2) [14].

(ii) Task-based IC

Task-based IC is a recent development of static IC where the application is divided into a set of small tasks that are executed atomically by a runtime [15–17, 41]. This development is motivated by protecting static IC against idempotency violations, by carefully controlling or recording each task’s accesses to nonvolatile memory, thus making code re-execution safe (C2) [15]. By definition, task-based systems handle atomic regions correctly and efficiently (C3). Additionally, systems like Alpaca [16] offer fast reboot and state-saving because only the runtime and the current task is booted, and only persistent data from the task needs to be saved (P1). However, in addition to re-execution overheads, handling idempotency is expensive [33]; Alpaca, an efficient task-based method reports 1.3 – 3.6× overhead when continuously powered (P1) [16].

In addition to technical challenges, such as

- expensive and complicated roll-back schemes necessary for protection against idempotency violations (C2),
- large memory footprint (P3),
- expensive task-transitions (P1), and
- handling interrupts to react to asynchronous events (P2),
task-based systems struggle to meet any of the three scalability objectives. The imposed programming model is not compatible with existing software (S2), as it requires redesigning applications into tasks. Finding optimal task boundaries (analogous to checkpoint placement in static IC) is difficult, and depends on both the specific underlying hardware and the energy harvesting conditions; great care must be taken to avoid sisyphean tasks (C4). Thus designing a task-based application generally requires a high degree of IC-specific expertise (S1) [39]. Furthermore, since optimal task-boundaries are hardware-specific, applications are not portable across hardware platforms (S3).

CleanCut [39] is a recent method that checks for sisyphean tasks (termed "non-terminating path bugs" in the paper) and performs automatic task decomposition. The method is a significant step towards making task-based methods scalable (S1-S3). However, limitations of the method, such as the requirement that the programmer must specify an iteration bound for each unbounded loop in the program, restrict its utility for applications that depend on existing libraries (S2).

Classical task-based approaches, such as Alpaca [16], are incompatible with asynchronous interrupts because they would break the premise of an application divided into atomically executed tasks. Ink [36] aims to resolve this by use of task-threads and scheduling. However, tasks in Ink cannot be pre-empted, making reaction time slow and unpredictable. A more recent alternative is Coati, which "employs a split-phase model that handles time-critical I/O immediately in a brief interrupt handler, but defers processing the interrupt’s results until after the interrupted task or transaction completes" [47] thus accomplishes real-time response to interrupts.

(iii) Reactive IC

Contrary to static and task-based approaches, reactive IC is mostly implemented in an application agnostic manner [18,19,48], greatly improving S1 and S2 when compared to typical task-based or static methods. Reactive IC consists of the following two operations:

**Suspend:** Save a snapshot of volatile state to nonvolatile memory, and enter low power mode or shut down.

**Restore:** Restore volatile state from a snapshot.

The set of volatile state is architecture and platform specific. However, reactive IC remains portable across hardware (S3), because only the suspend and restore functions need be ported for each platform, not the entire application.

Power supply monitoring is set up to generate interrupts that trigger suspend and restore operations when the supply voltage, \( v_{cc} \), crosses a threshold. When the restore threshold is exceeded, an interrupt triggers restore, which restores state from a previous snapshot. Similarly, when \( v_{cc} \) drops below the suspend threshold, \( V_S \), a snapshot is saved and the system enters low-power mode (where volatile state is retained) or shuts down. If power returns while the system is still in low-power mode, restore is unnecessary because the volatile state was retained, so the application can continue execution directly, thus improving P1 [19]. When powered by a low-current source, suspend can often be avoided altogether by pre-emptively entering sleep mode before \( v_{cc} \) drops below \( V_S \) [49]; this can, in some cases, improve P1.

The suspend threshold ensures that snapshots are only saved when power failure is imminent, thus nearly eliminating superfluous save operations (P1) [18,33,50]. The restore threshold is used to guarantee the success of the next checkpoint; it eliminates code re-execution, and thus also idempotency violations (C2). Determining such a \( V_R \) is intractable for static and task-based IC because all possible execution paths must be exhaustively analysed. In reactive IC, \( V_R \) can be determined by online measurement [19], or calculated based on the amount of state to suspend and restore [50]. Note that to guarantee the success of the next checkpoint, a minimum amount of energy buffering is required. The energy buffer must suffice to power suspend, but additional capacity can improve performance.
With proper, pre-emptive, interrupt prioritization, reactive IC is natively reactive to asynchronous external events (P2), and furthermore can guarantee forward progress in application code (assuming that any atomic functions used are sufficiently small).

The main drawback of reactive IC is that saving and restoring the entire state is expensive compared to task-based systems where only one task is loaded/saved at a time. This is especially wasteful if the power cycle is short, because only a small part of the program is expected to execute. This problem can be partially alleviated by techniques such as: comparing the current state to the latest snapshot, and only saving the difference [51,52]; or IC-specific memory management [50]. Another drawback is that reactive IC does not generally handle high-level atomicity constraints (C3) as gracefully as task-based or static methods. A solution to this is proposed in [53], where a formal division between application code (which accesses only application memory) and driver code (which accesses peripherals and nonvolatile memory) assists a kernel to ensure atomic execution of driver-functions without introducing idempotency violations (C2). But adding atomicity to a reactive system can introduce sisyphean tasks (C4). Samoyed [33] handles sisyphean tasks by profiling all atomic tasks in hardware, and, for suitable peripherals/operations, dynamically breaks down the task into parts that are small enough to execute in a single on-period. Not all peripheral operations can be divided into smaller parts, but as shown in the paper, this method is very effective for memory mapped accelerators.

4. Unsolved challenges in energy-driven computing

Energy-driven computing brings with it unique challenges that slow down adoption in the real world. While there have been a plethora of applications for battery-powered devices in the IoT literature, energy-driven devices are still novel. Demonstrations of real world applications are starting to emerge for energy-neutral systems, e.g. a self-powered wireless sensor node for predictive maintenance [54]. However, few realistic applications have been demonstrated using intermittent or power-neutral operation. Developing applications for power-neutral and/or intermittent computers is a challenge, requiring new thinking throughout the stack: from the design of circuits for power management, selection and integration of energy harvesters, software, and the entire communications stack. On the topic of Energy-Driven Computers, there are many research challenges yet to be solved. We highlight what we regard as the main remaining challenges that are currently hindering adoption: development tools for energy-driven computers, and wireless communication under intermittent operation.

(a) Simulation, emulation and debugging

Simulation and debugging of energy-driven computers is inherently more complex than for traditional application-driven ones. This stems from the fact that energy-driven computers adapt their behaviour according to the amount of available energy. Traditionally, simulations are performed under the assumption of a steady power supply, so emulating the ISA and platform is usually sufficient for finding bugs, evaluating performance and developing applications. However, to correctly simulate the behaviour of an energy-driven computer, execution, power consumption and power production (harvested energy) must be simulated simultaneously to correctly model their influence on each other. While previous works have explored simulation of energy-driven computers, their solutions target only specific processors, and they replay voltage traces rather than model the complex dynamics of a real energy harvester [55].

Another option is to record current-voltage traces for real energy harvesters, then replay them with a power-synthesizer to produce a repeatable power source [56,57]. While useful for performance evaluation, this method does not provide the introspection capabilities of simulation, so is only moderately useful for validation/debugging.

Interactive debuggers like the ubiquitous GNU debugger (GDB) typically control the execution and state of a target processor via a JTAG-like interface, providing invaluable introspection into program and device behaviour. Key to their utility is the ability to halt and single-step a program.
However, in an energy-driven system, halting the processor is insufficient; ideally one should be able to halt the energy dynamics too. EDB (Energy-interference-free Debugging) [58] provides the ability to both monitor and control the processor’s execution simultaneously with a platform’s energy buffer. This enables new debugging primitives such as energy-breakpoints. However, this does not “pause” the power generation (energy harvester). Combining the works in recording and replaying EH dynamics [56,57] with EDB [58] could be a promising direction for a holistic debugging environment for ICS.

(b) Wireless communication under intermittent operation

Due to the nature of sources that ICSs rely on, which can be highly unpredictable and fluctuating (both temporally and spatially), coordinating a number of intermittently-powered, distributed devices/systems is an ongoing challenge. Thus, the state-of-the-art intermittent networks mostly embody mains- or battery-powered always-on sinks for data collection and coordination. This is regarded as the most convenient/ reliable way for uplink communications through these networks. Here, the always-on sink collects data from the ICSs in a single-hop manner and establishes a gateway to the Internet, whereby the ICS data become accessible from anywhere at any time. This centralized structure, i.e. star topology, helps to alleviate the uncertainty occurred due to non-deterministic on/off periods, enabling ICSs with active radios having enough energy to freely transmit their data by knowing that the sink is constantly listening to the channel for data reception. These systems are often not acknowledged, which means that a collision could go unnoticed. Yet, depending on the parameters of interest, confirmation of reception is highly required, especially for mission- or information-critical applications, to guarantee reliable detection of events and/or reporting of vital parameters. However, this is hard to accomplish in intermittent operation, after uplink data transfer more specifically, where traditional medium access control (MAC) layer solutions have to be revisited.

Downlink communications, i.e. from sink to ICSs, on the other hand, is more problematic as neither the ICSs know when they should listen nor the sink anticipates when an ICS becomes active and/or is ready to receive. In practice, the ICSs can perform listening for a very short duration only, mostly in a random fashion, due to their limited energy buffer. Thus, precise timing, i.e. synchronisation, between them and the sink is a challenging issue. To avoid loss of important data and so retransmissions, the ICSs go to sleep after computation and wait for an interrupt from the sink while recharging their buffer. For this particular reason, intermittent operation needs competent and efficient wake-up mechanisms.

The very basic way of waking up an ICS is with wireless power transfer (WPT) from a sink [59]. An ICS with a rectifying antenna (rectenna) can trigger an interrupt by using the intercepted RF signals radiated from the sink to wake up and get ready to receive. By following the power signal, the sink can transmit data to the ICS that is now active and listening. This operation is costly for the sink as it has to transmit twice, once for wake-up and once again for communication, which also decreases bandwidth utilization efficiency. After reception of data, the ICS acknowledges the sink. Alternatively, instead of sending the control frame (acknowledgement) back to the sink immediately, it can be hooked onto the next transmission (which is known as piggybacking); however, this does incur a delay [60]. Despite simplicity, rectennas are prone to false wake-ups since any signal with a matching frequency, not originating from the sink of interest, may unintentionally wake the ICS up.

For more energy-constrained applications, passive radios, with significantly lower power consumption than the active radios, can be utilized. In the very well-known method of passive radio communications called backscattering [61], the ICSs with passive radios, i.e. tags, use the same signal directed from a centralized carrier generator to reflect their data back by changing their antenna impedance. Although backscattering enables small form factors, low-cost and low-power solutions, backscattering-backed communications suffer from frequent power interrupts, collisions, low data rates, and short communication ranges. The majority of existing research
efforts on this field is on uplink communications; however, there are some recent studies aiming at speeding up the downlink data transfer [62,63].

Apart from environment- and physical layer-related matters, the fundamental factor limiting the widespread adoption of ICSs is synchronization, which severely restricts peer-to-peer communications. This can be only achieved when both ends are circumstantially active for a long enough period, at the same time. Thus, a power-failure-resilient timekeeping solution is evident [64]. Once the challenge of establishing peer-to-peer communication in ICSs is overcome, and hence ICSs can be considered as distributed systems, a wealth of further challenges emerge – for example how to coordinate, schedule and manage energy across a distributed and federated system. These will be followed by debugging, simulation, and optimization of the networked ICSs. Due to the very same reason, i.e. lack of synchronization/timekeeping, the ICSs currently does not support mesh topology and multi-hop communications, i.e. routing protocols, which need further investigation. Moreover, the existing protocol stack, including cross-layer communication solutions, has to be redesigned to allow intermittent operation to reach its full potential.

5. Discussion

Energy-driven computers are gaining traction as a solution for the demands of IoT end devices. While battery-powered devices are suitable for many applications, their deployment lifetimes are limited by their fixed capacity. Energy-harvesters can be used to extend the deployment lifetime of devices, but bring unique challenges due to the dynamics of their power output. This motivates a shift into energy-driven computing, where energy-availability is considered from the beginning of the design process. In this paper, we described three prominent examples of energy-driven computing: energy-neutral, power-neutral and intermittent operation.

Energy-neutral operation augments a rechargeable battery with an energy harvester that supplies energy during deployment; thus lifetime is extended without increasing battery size. However, the charging circuitry imposes electrical losses and the rechargeable batteries are still a limiting factor in terms of size, cost, environmental impact, and their limited charge-cycles limit deployment lifetime. Power-neutral computers seek to operate without rechargeable batteries, using only a tiny amount of capacitance for buffering, to simplify the circuit and thus substantially reduce cost, size and environmental impact while simultaneously improving lifetime. However, power-neutral computers cannot sustain operation if harvested power drops too low, so long-running applications become infeasible. Intermittent operation solves this problem by retaining computational progress through power cycles; they do not attempt to hide the dynamics of energy harvesting, but rather accept that the system will shut down frequently and unpredictably. However, this imposes new challenges, including losing track of time, and ultimately reducing achievable QoS for most applications.

Development of energy-driven computers is complicated because the developer has to reason about application requirements, energy demands, and energy availability simultaneously. Current development tools lack the ability to sufficiently control both the energy-availability and execution simultaneously, hindering the adoption of energy-driven computers.

Wireless communication is tremendously important for achieving a scalable IoT and enabling a multitude of sensing applications. However, this necessitates a reliable energy supply and a precise notion of time; energy-driven computers are severely constrained on energy, and intermittent computing systems lack a sense of time. These require novel physical layer solutions to cope with the non-deterministic factors in both applications achieving sufficient QoS in uplink/downlink communications. In addition, lightweight synchronization and power-failure-resilient timekeeping mechanisms are of utmost importance. Directions for future research into this area include establishing peer-to-peer communications, efficient routing protocols and application-specific network topologies.

Data Accessibility. No data beyond those explicitly stated in this article were generated for this study.
Competing Interests. The authors declare that they have no competing interests.

Funding. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under an iCASE award and grant EP/P010164/1.

References


26. Han J, Orshansky M. 2013 Approximate computing: An emerging paradigm for energy-efficient design. In 2013 18th IEEE European Test Symposium (ETS), pp. 1–6. doi:10.1109/ETS.2013.6569370


computing systems.
ACM Transactions on Embedded Computing Systems (TECS) 17, 93.
doi: 10.1145/3281300
doi: 10.1109/TCAD.2016.2547919
30. Jackson N. 2019, Permanote git repository.
IEEE Transactions on Circuits and Systems I: Regular Papers 64, 2370–2379.
doi: 10.1109/TCSI.2017.2707304
doi: 10.1145/3173162.3173210
doi: 10.1145/3314221.3314613
doi: 10.1109/2903140
IEEE Access 6, 665–676.
doi: 10.1109/ACCESS.2017.2774261
doi: 10.1145/3274283.3274837
doi: 10.1145/2809695.2809707
doi: 10.1145/3131672.3131673
IEEE Micro pp. 1–1.
doi: 10.1109/MM.2017.265085515
doi: 10.1145/3279753.3279759
doi: 10.1145/1950365.1950386


44. Woude JVD, Hicks M. 2016 Intermittent computation without hardware support or programmer intervention. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI) 16, pp. 17–32.


doi: 10.1145/2744769.2744809

doi: 10.1145/3314221.3314583


doi: 10.1145/3142992.3142998

doi: 10.1145/3316781.3317812


doi: 10.3390/s19122747

doi: 10.1109/RSP.2018.8631991

doi: 10.1145/2668332.2668336

doi: 10.1145/2820645.2820653

doi: 10.1145/3178372.3179525

*IEEE Wireless Communications* **23**, 10–18. 
doi: 10.1109/MWC.2016.7462480

60. Tanenbaum AS, Wetherall DJ. 2010 *Computer Networks*. 

*arXiv:1901.10274 [cs]*

doi: 10.1109/INFOCOM.2016.7524574

doi: 10.1109/INFOCOM.2017.8056987

64. Donne CD, Yildirim KS, Majid AY, Hester J, Pawelczak P. 2018 Backing out of Backscatter for Intermittent Wireless Networks. 
doi: 10.1145/3279755.3279758