ENERGY-EFFICIENT COMPUTING PATTERNS FOR COST-AWARE INFORMATION-SYSTEM MODERNIZATION

Boltayev Alisher Ataboy o‘g‘li
Assistant at the Department of Information Systems and Technologies, TSUE.
Email: alisherboltaev66@gmail.com

Received: 14.10.2025
Revised: 15.10.2025
Accepted: 16.10.2025

Keywords:
energy-aware scheduling; cost-aware modernization; approximate computing; virtual-machine consolidation; thermal-aware placement; compute-in-memory; energy telemetry; sustainable IT; digital transformation; green cloud computing.

Annotation. This article investigates modern approaches to incorporating patterns of energy-efficient computing into the modernization of enterprise Information Systems (IS) in the digital economy. The rapid digitalization has increased the workload of computing, leading to increases in electricity usage, which has raised operating costs, necessitating the proposed action. This paper delivers a synthesis of recent studies discussing energy-aware scheduling, virtual-machine consolidation, approximate computing, compute-in-memory architectures, and thermal-aware resource placement and describes how these technologies fit systematically within a cost-aware approach to the modernization of IS. Drawing on Masdari et al. (2022); Al-Haija et al. (2023), this study develops a conceptual framework to connect hardware-level optimizations (e.g., DVFS, power gating), algorithmic efficiency (approximation, locality), and governance (telemetry, KPIs, and policy enforcement). The article concludes by reinforcing that measurable energy efficiency should be a first-class requirement of IS design that equals performance and reliability. A viable navigation path is recommended that starts baseline measuring of energy, implementation of low-risk optimization (e.g., Energy Efficient Ethernet, consolidation), algorithm adaptation, and continuous monitoring with RAPL counters and feedback control. The synthesis notes that multi-layer integration of these patterns can result in 30–50 % total energy consumption savings while maintaining quality of service. Finally, the paper concludes that integrating energy awareness in IS modernization reduces total cost of ownership and enhances corporate sustainability and readiness for low-power computing approaches in the future, such as compute-in-memory

Introduction. In the digital economy of the 21st century, businesses have a dual challenge: modernizing their information systems (IS) for agility, analytics, and automation, but doing so in a manner that does not allow electricity costs and carbon footprints to spiral out of control. Each part of the modernization process—moving workloads to the cloud, decomposing monoliths into microservices, embedding AI and analytics—drags increased computational intensity, which increases energy consumption. Data centers currently consume approximately 1.5% of the world's electricity demand and are forecasted by the International Energy Agency to double their consumption by 2030, partly due to increased demand from AI and big-data workloads. Without a conscious energy strategy, digital transformation may quietly escalate operating expenses and environmental impacts. Energy effectiveness needs to shift from a facilities issue to a first-class design principle; work in computer engineering and cloud engineering has already established mature methods to address energy consumption, including energy-aware scheduling and putting workloads on virtual machines (VMs) to reduce idle power consumption, approximate computing to reduce arithmetic overhead, thermal-aware placement to relieve cooling load, and compute-in-memory (CIM) architectures that minimize data movement. These patterns together can be embedded in IS modernization, allowing organizations to reduce total cost of ownership (TCO) while advancing organizational performance and sustainability goals.

Energy-aware scheduling, vm consolidation, and placement.

Scheduling is a central factor in compute efficiency due to the fact that it determines where and when every workload executes. Traditional schedulers, whether in operating systems or Kubernetes clusters, largely optimize for performance metrics like CPU load or response time. However, one scheduling decision can result in an energy use difference of many tens of percent based on voltages, the workload's resource utilization, and efficiency of the cooling.

Since that time, reinforcement-learning methods such as the EETS deep-Q scheduler have evolved this idea: an agent learns in real-time how to pack tasks onto nodes to amortize kilowatt-hours while remaining within their latency requirements[1]. In virtualized, or cloud, environments, the same idea applies just at a higher granularity. Masdari et al developed CEWOAVM, a cost-aware and energy-aware migration algorithm using whale optimization; their experimental results showed energy savings between 18% and 36% from the standard policies[2]. Also, Nguyen et al created the EPOBF heuristic which allocates VMs based on performance-per-watt, and out-performed load-balancing alone[3].

In a contemporary enterprise setting, this means that every orchestrator from Kubernetes to VMware should add power and/or temperature as a dimension for scheduling. Workloads can be tagged for energy criticality: customer-facing microservices stay on high-performance nodes, while, say, non-critical analytics, or background synchronization jobs can wait for off-peak hours or to run on low power hosts. Live migration then enables nightly consolidation, enabling low-energy hosts to sleep when they are unused. Even while live migration incurs a small energy cost,

literature such as the Energy-Aware Task Scheduling (EATS) framework demonstrates that if the policies are tuned correctly, the migration overhead will be more than offset.

The principle of energy proportionality—that servers consume energy proportional to their workload—is the basis of these methods. DVFS, parking CPU cores, and implementing energy caps on a per-VM basis help enterprises move toward this principle. In addition, when modernization consists of refactoring legacy systems into cloud-native components, embedding these types of controls in orchestration logic can deliver lasting reductions to OPEX without touching the business logic.

Approximate computing, data locality, and compute-in-memory.

In terms of results, we want to look at the fundamental algorithmic frontier. A significant portion of energy consumed in digital systems is not expended on computation, rather, it is used to move data, or transferring bytes, from memory or storage to cpu registers. Research from Han et al has found that for many applications that tolerate errors—multimedia, AI inference (and other sets of tech analysis), sensor fusion and integration of various modalities—relaxed precision results in a much larger reduction in energy use than a corresponding loss in accuracy. They even put together a paper called "Approximate Computing: An Emerging Paradigm for Energy-Efficient Design" which talks about loop perforation, reduced bit-width accuracies, or even skipping multiple iterations when non-essential, in studies that find over 50% reductions in power with little (or no) losses in accuracy. With the these studies in mind, we can think about systems as whole, De et al. Recommend dynamic quality scaling, wherein systems can be designed to switch between precise and approximate computations that are tied to the urgency of the workload or energy budget[4].

Hardware innovation supports this. The ADCIM architecture combines rapidly AES (Approximate Enabling Subnets) into compute-in-memory (CIM). For example, they used this for the self-attention layers needed for transformer models. The findings were that this improved the energy efficiency of these transformers, compared to a GPU, between 10x and 100x or more[5]. While CIM hardware is still considered "emerging," enterprise pipelines can/should be developed to treat back-end compute as plug-able architectures, in anticipation for such accelerators to become available. In the meantime, energy gains and performance improvements can be realized purely through the consideration of locality of similar tasks through: In-memory aggregation of variables, columnar storage, caching sub-systems, or co-locating computer with data devices.

Approximation and locality together create a continuum of trade-offs: exactness versus efficiency. In enterprise analytics, the pragmatic rule is "be precise where it counts and efficient everywhere else." A dashboard can sample the data, a recommendation engine can quantize weights, and log analysis can down-sample the events—these are all minor changes that can accumulate into something quite substantial across millions of transactions.

Thermal-aware consolidation and infrastructure synergy.

Energy waste happens at levels beyond just the chips, it can extend to the fans, chillers, and pumps that keep systems operating within thermal specs. When heat loads are concentrated to a small area, components suffer reduced lifetimes, and cooling systems have to work harder. Al-Haija et al noticed this coupling, and developed a TM-VMC (Thermal-aware Virtual-Machine Consolidation) (TM-VMC) algorithm, which incorporates heat dissipation at a rack level, and allocates VMs so that no hot spots occur[6]. Their results showed total facility energy use reduced up to 12 % when comparing TM-VMC to just energy based consolidation.

Additional complementary work EAFT (Energy-Aware and Fault-Tolerant Scheduling, 2022) shows as well that adding redundancy and improving reliability does not has to compromise efficiency, by optimizing the location of replicas in the system the authors found an available system while using less energy.

In practice, companies can co-optimize their IT and facility operations by allocating compute loads to thermal zones. Servers located in cooled aisles or with direct liquid cooling can be assigned compute loads related to high levels of AI processing; meanwhile, lighter loads can be assigned to warmed racks. Some companies manage workload scheduling related to the energy grid and tariff costs, a high energy process can run at night when ambient temperature is lower and cooling is cheaper. Others have gone as far as to relocate batch compute clusters to areas with renewable power and climates that are naturally cooler, potentially reducing scheduled compute and dynamic workloads[7].

The synergy between compute and cooling can be summarized as follows:

Synergy Level	Typical Technique	Expected Reduction	Reference Example
Chip / Board	DVFS + power gating	15 – 40 % CPU energy saving	Abrishami & Naghibzadeh (2013)
Node / Cluster	VM consolidation + migration	18 – 36 % system energy saving	Masdari et al. (2022)
Rack / Facility	Thermal-aware placement + cooling optimization	5 – 12 % total PUE improvement	Al-Haija et al. (2023)
Application / Algorithm	Approximate computing + data locality	10 – 50 % compute energy saving	Han et al. (2013); De et al. (2023)

The cumulative effect across these layers can exceed 50 % energy reduction, translating directly into cost savings and sustainability metrics.

Governance, telemetry, and roadmap for modernization.

Technical measures can only be effective if they are anchored in measurement and governance. Today's CPUs offer RAPL (Running Average Power Limit) counters, and ARM, AMD, and others expose similar telemetry interfaces. When these measurements and telemetry feed into observability stacks (Prometheus, Grafana), you can accurately monitor joules per transaction or watts per virtual machine. Once energy is visible and counted as a key performance indicator (KPI), improvement is often a natural byproduct.

A pragmatic modernization roadmap begins with measurement, because without measurement, energy cannot be optimized. The first step is to enable hardware counters and collect baseline measurements over a typical workload cycle. In step two, you will determine hotspots by evaluating output in applications or services that consume energy disproportionally versus delivered business value. Step three considers 'low-risk' optimizations, such enabling Energy-Efficient Ethernet (IEEE 802.3az), tuning power states, or consolidating idle servers. Step four considers intelligent scheduling that has proven results using methods like EETS or CEWOAVM. Step five refactors algorithms for approximation and locality, with step six being the feedback loop to gauge temperature, and to monitor and enforce policies.

Governance promotes persistence. Each service team should own an energy budget similar to their ownership of latency and availability objectives. Any proposals for architectural changes must come with an energy-impact statement indicating expected changes in consumption. CI/CD pipelines could include automated energy regression tests-once established-in their testing column, which could block any deployments that exceed energy thresholds. Quarterly dashboards that benchmark energy per transaction for departments provide transparency that leads to positive competition.

In summary, measurement creates accountability to change theoretical understanding into action. Governance creates transformations from individual optimizations to significant sustainable modernization.

Conclusion. Energy efficiency is quickly establishing itself as the central constraint of digital transformation. Evidence from the literature shows that efficiency and performance are not mutually exclusive: Berral et al. show that intelligent consolidation saves power with little overhead; Abrishami, and Naghibzadeh show that DVFS-based multi-objective scheduling produces both speed and efficiency; Masdari et al., and Nguyen et al. report double digit benefits from energy aware VM migration and placement; Han et al., and De et al. confirm that approximation reduces computational energy by half in tolerance applications; Al-Haija et al. includes thermal considerations to mitigate cooling demand; and green-IT surveys provide governance that locks in improvements. Together these studies provide a path to governance-aware IS modernization based on energy-efficient computing patterns.

Businesses that incorporate these procedure-embedded patterns—scheduling by watts as well as load, consolidating while being informed by thermal zones, using approximate algorithms, architecting for compute-in-memory, and exercising governance via telemetry—will likely benefit from tangible financial and

environmental returns. The return on modernizing is two-fold: a reduced cost of operating and strengthened corporate sustainability. At a time when global computing is confronted with stresses related to grid capacity and climate policy, these practices change energy from being a liability to allowing the use of energy as a competitive advantage.

REFERENCES

1. EETS Authors (2024). *"Energy-Efficient Task Scheduling Using Deep Reinforcement Learning in Cloud Computing." Alexandria Engineering Journal.*

2. Masdari M., Jafarian A., Ghahramani M. (2022). *"CEWOAVM: Cost-Effective and Energy-Aware VM Migration in Cloud Data Centers." Electronics 11(21): 3639.*

3. Nguyen Q.-H., Thoai N., Nguyen T.-S. *"EPOBF: Energy-Efficient Allocation of Virtual Machines in HPC Cloud." arXiv preprint.*

4. De S. et al. (2023). *"Approximate Computing for Energy Efficiency and Dynamic Quality Control." Eindhoven University of Technology Research Report.*

5. Zhang X. et al. (2025). *"ADCIM: Approximate Computing in Compute-in-Memory for Self-Attention." Journal of Systems Architecture.*

6. https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-023-00462-2.

7. Zhao Y. et al. (2022). *"EAFT: Balanced Energy-Aware and Fault-Tolerant Data Center Scheduling." Sustainability 14(5): 2781.*

8. Masdari M., Jafarian A., Ghahramani M. (2022). *"CEWOAVM: Cost-Effective and Energy-Aware VM Migration in Cloud Data Centers." Electronics, 11(21): 3639.*

9. Alisher Boltayev "*Savdo korxonalarida it texnologiyalarining roli va raqobatbardoshlikni oshirish usullari*" Raqamli iqtisodiyot" ilmiy-elektron jurnali maxsus soni. 2024 (999-1005).

10. Alisher Boltayev "*Revolutionizing trading enterprises: a comprehensive approach to enhancing management information systems*" "Raqamli iqtisodiyot" ilmiy-elektron jurnali maxsus soni. 2024 (58-64)