ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 6 Issue 3 July 2026
Paper Id: JETA-V6I3P104/ Page No: 25-34

Review Article

Energy-Aware Bare-Metal Provisioning in Virtualized Data Centers

Ratan Raj Anandeshi
Campbellsville University, Kentucky

Received Date: 18 July 2026 Revised Date: 25 July 2026 Accepted Date: 30 July 2026

Abstract: *Energy-conscious bare-metal provisioning has become an important research question since virtualized data centers are now full of mixed workloads, whose performance, isolation, and energy usage vary radically across virtual machines, containers, and direct hardware allocation. Traditional consolidation-based resource management assumes that more virtualization and higher resource utilization will always improve efficiency; however, more recent work shows this does not hold when boot latency, migration overhead, thermal characteristics, I/O interference, and workload heterogeneity are factored in. This paper reviews the current state of research on this topic in energy-aware bare-metal provisioning in virtualized data-centers and has a special focus on provisioning delay, architecture choice, power-aware scheduling, placement optimization, thermal interrelations and hybrid orchestration between bare-metal and virtualized resources. Key topics include energy modeling, dynamic resource consolidation, VM placement, bare-metal reservation, hybrid resource management, and the activation-energy trade-off. Reported research suggests that energy efficiency depends heavily on server utilization, provisioning granularity, interference sensitivity, data center-wide orchestration policies, and the ability to reserve non-virtualized nodes for latency-sensitive workloads.*

Keywords: *Bare-Metal Provisioning, Cloud Data Centers, Energy Efficiency, Resource Orchestration, Virtualized Infrastructure*

I. INTRODUCTION

Modern data centers no longer rely on a single resource-abstraction model. Virtual machines, containers and bare metal servers now co-locate within production systems and architectures that must meet competing objectives such as latency, throughput, isolation, elasticity, and power efficiency. Initial research in the area of cloud energy management focused on server consolidation and utilization-aware scheduling since idle and underutilized machines were identified as one of the main sources of avoidable energy waste [1]. That view was subsequently expanded by later surveys, which indicated that reliability, migration cost, thermal coupling, and performance degradation make any easy connection between increased consolidation and reduced power draw [2], [3]. Formal energy-modelling research added another significant factor to the analysis when it was shown that properly evaluating energy policies requires clear models that link workload behaviour to infrastructure state and overall power use [4]. Recent attention has revealed that bare-metal reservation and provisioning time should also be included in the analysis since infrastructure efficiency is not limited to steady-state utilization but also how quickly and selectively physical nodes can be activated, reserved, and released [5]. Taken together, this work shifts the focus from narrow VM consolidation toward a broader question of how heterogeneous provisioning modes are to be co-ordinated in an energy-conscious data center.

The issue has gained significance due to the growing support of direct hardware exposure by cloud providers and enterprise platforms along with the traditional virtualization. The studies carried out in relation to Metal-as-a-Service systems prove that the provisioning and booting delays of bare-metal instances are not incidental operational details, but, instead, nontrivial design variables [6]. Surveys of server architecture recommendation literature indicate that bare metal, containers, and KVM virtual machines have significantly different startup and performance configurations, in which prior claims about the importance of static allocation are becoming unsustainable [7]. Workloads with intensive I/O behaviour, strong data-locality requirements, or high interference sensitivity may consume less aggregate energy on a temporarily reserved bare-metal node than on a heavily virtualized host which needs to repeat through migration, or manage oversubscription, or recover its performance, etc. Energy-aware provisioning therefore cannot thus be reduced to reducing the number of active-servers. The appropriate question in the case of heterogeneous data centers is what mode of provisioning provides the desired service level at minimum total energy and control overhead.

The greater research environment indicates, too, that virtualization is still essential, although not always the best. Considerable returns have been achieved in most cloud environments through task-consolidation heuristics, dynamic VM consolidation, and power-aware placement strategies [8]-[10]. Subsequent research proposed I/O-aware scheduling, global energy-conscientious placement, energy-proportional VM scheduling, and predictive anti-correlated placement to handle

This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/2.0/)

heterogeneous demands more effectively [11]-[14]. This agenda was further developed with reinforcement-learning schemes and thermal-aware scheduling schemes that tried to adjust the choice of allocation in real-time conforming to varying workloads and cooling loads [15], [16]. Meanwhile, containerized and hybrid-cloud research studies have shown that the overhead of migration, orchestration organization, and the type of workload may invert anticipated savings particularly in scenarios where applications are based on multiple layers of abstraction [17]-[19]. This synthesis of results renders the concept of bare-metal provisioning analytically pertinent even in facilities that are highly virtualized since physical-node reservation may serve as a corrective mechanism in the situations under which the overheads caused by virtualization may eliminate anticipated efficiency benefits.

Even in the current state of extensive development, some aspects of the problem are yet to be solved. One, most of the literature measures placement, consolidation, or orchestration within largely virtualized environments, but a direct analysis of bare-metal provisioning has received limited attention as a parameter of energy-management has simultaneously few notes. Second, provisioning-time metrics like boot delay, reservation window and server activation cost are normally discussed independently of run-time metrics, including migration overhead, thermal stress, energy proportionality. Third, most published appraisals are based on simulation or workload traces without considering operational constraints, e.g. image distribution delay, orchestration reactivity, or mixed-fleet fragmentation. This review seeks to synthesize evidence reported in journals pertinent to the energy conscious bare-metal provisioning in virtualized data centers and re-frame that evidence in an conceptually unified framework that encompasses the reservation policy, provisioning delay, workload suitability, and energy-performance trade-offs. In the sections that follow, the previous literature, emerging patterns, comparative findings, future studies, and the wider implications of considering bare metal as a responsible energy-conscious choice of virtualized infrastructures are discussed.

II. LITERATURE REVIEW

One major thread in the literature looks at energy efficiency through virtualization-aware consolidation and placement. The concept of energy-efficient utilization was initiated by Lee and Zomaya [8], who viewed it as a task-consolidation problem where to accomplish the task both active and idle energy is to be taken into account. Beloglazov and Buyya [9] have applied that reasoning to dynamic consolidation, which necessitates the management of the impact of SLA without decreasing the count of active hosts. The field of VM placement and migration grew later in the survey by Silva Filho et al which demonstrated that resource balancing, network effects and migration expenses are properties that all interplay so that optimization is inadequate in single-objective analysis [3]. This was supported by a parallel survey conducted on reliability and energy efficiency where the findings indicated that energy minimization may compromise resilience and service continuity on failures and very dynamic demand [2]. This body of work, as a whole, examines virtualization as the primary control mechanism of energy, but it also reflects the downside of consolidation where interference, reliability, and migration penalties become highly optimistic.

The second stream involves modelling, prediction and systematic assessment. Formal models like E-mc2 [4] and subsequent review of cloud-simulation energy models suggest that policy analysis should be based on clear models of power, defining the behaviour of CPU, memory, storage and network. This modelling agenda matters to bare-metal provisioning since the activation costs associated with physical-node are different with energy and timing costs that cannot be well modelled using abstractions such as VMs alone. Metal-as-a-Service environments predictive studies indicate that provisioning and booting times can have a significant impact on the quality of orchestration [6]. An architecture recommendation study on services of IaaS platforms contributes to the addition of knowledge: Firstly, bare metal, containers, and virtual machines differ in performance containment, but also in startup time and applicability of operation [7]. These findings indicate that energy-conscious provisioning involves the use of models that traverse steady-state power, activation, orchestration, and workload-sensitive interference.

A third group of literature looks at more advanced scheduling and placement systems within virtualized systems. There are power-aware placement [10], I/O-aware VM scheduling [11], global energy-aware placement [12], energy-proportional scheduling [13], and predictive anti-correlated placement [14], all of which strive to make resource mapping under heterogeneous conditions more efficient. Taken together, the contribution of these studies is more about method than adding small steps: the energy consumption is demonstrated to be based on placement quality, pattern of interference, and hardware profile as opposed to consolidation depth. Other more recent works based on reinforcement learning [15], thermal-aware scheduling [16], and collaborative VM optimization [17] also suggest that adaptive control has become an even more important subject in efficient management. That notwithstanding, these studies tend to assume that the only resources worth optimizing are the virtualized ones. This assumption breaks down in situations where, to evade a latency-sensitive, interference-prone, or high-I/O portfolio, the workload may have to spend a reconfiguration fee, then transition to a reserved bare-metal pool.

A fourth, and particularly relevant, line of work deals with heterogeneous and mixed infrastructures more directly. Bare-metal reservation analysis [5] focuses on the trade-off between reactivity and energy efficiency and demonstrates that maintaining pre-reserved nodes to enhance the responsiveness may backfire when reservation policy is mis-configured. The HeporCloud study found that workload-sensitive schedulers can independently achieve large energy reductions and yet achieve sublinear performance reductions in heterogeneous data centers. Containerized-datacentre studies also indicate that consolidation and migration approach of the abstraction layer should be cautious of the low-lying abstraction layer [17], [18]. The current evidence on the matter has been incorporated in a recent review of energy, performance, and cost-efficient cloud data centers [20], which argues that the resource-management decisions need to be analysed on the hardware, orchestration, and application levels. The general literature so sees the tendency to a more comprehensive understanding of energy-aware provisioning where bare metal is not only not an aberration but also not a luxury option, but rather a part of heterogeneous data-center energy management.

Table 1: Summary of Key Findings

Ref	Focus	Key Findings
[6]	Metal-as-a-Service provisioning prediction	Provisioning and booting times for bare-metal instances can be modelled from platform-monitoring data, indicating that activation latency is a first-order orchestration variable rather than a negligible overhead.
[7]	Server architecture recommendation across bare metal, containers, and VMs	Bare metal, containers, and KVM virtual machines exhibit materially different startup and performance characteristics, supporting differentiated infrastructure selection rather than uniform virtualization.
[8]	Energy-efficient task consolidation	Energy-aware heuristics that account for active and idle energy improve resource utilization efficiency in cloud systems and establish consolidation as a foundational energy-management mechanism.
[9]	Dynamic VM consolidation	Adaptive VM consolidation can reduce energy consumption while controlling SLA violations, but migration decisions remain sensitive to workload volatility and host-state changes.
[10]	Power-aware performance-guaranteed VM placement	Joint treatment of power and application-performance requirements improves placement quality relative to methods that optimize energy alone.
[11]	I/O-aware VM scheduling	I/O virtualization overhead can erode expected energy gains, making collective-I/O aware scheduling preferable for mixed workloads in virtualized clouds.
[12]	Global energy-aware placement	Placement policies that account for system-wide effects outperform purely local assignment, especially in heterogeneous cloud data centers.
[13]	Energy-proportional VM scheduling	Server energy proportionality and measured efficiency profiles can guide VM assignment more effectively than saturation-oriented packing alone.
[14]	Predictive anti-correlated placement	Resource-usage prediction and anti-correlation improve placement decisions by reducing contention and service violations while lowering energy demand.
[15]	Reinforcement-learning VM consolidation	Learning-based consolidation can improve dynamic distribution of VMs and enhance energy-performance balance under uncertain and changing workloads.
[16]	Power- and thermal-aware VM scheduling	Joint optimization of power and thermal states improves scheduling quality and highlights cooling-aware control as a necessary extension of power-aware provisioning.
[17]	Collaborative VM scheduling	Multi-factor VM scheduling can improve energy efficiency through coordinated optimization rather than isolated host-level decisions.
[18]	Containerized-data-center consolidation	Migration and consolidation outcomes differ markedly among VMs, containers, and applications running inside VMs, implying that abstraction-layer choice has direct energy consequences.
[19]	Hybrid heterogeneous cloud orchestration	Workload-aware orchestration across heterogeneous resources can save substantial energy while preserving performance, supporting mixed-fleet resource management.

A trend is decisive as seen in Table 1; the literature has shifted out of generic consolidation and has oriented in the direction of a greater differentiation of control over the processes of activation, placement, orchestration and heterogeneity.

That development underlines the assumption of this paper to study bare-metal provisioning as a part of resource decisions on an energy-conscious continuum and not its own functional niche.

III. METHODOLOGY

The field uses a combination of methods: survey analysis, simulation-based evaluation, predictive modelling, empirical trace analysis, and architecture-based experimentation. The surveys and taxonomies [1]-[3], [20] offer the widest of all views in classifying energy-aware work into placement, consolidation, scheduling, thermal control, and orchestration of hybrids. These studies can be useful in determining the similarity of concerns, but in general such broad reviews tend to confuse the difference between provisioning decision and steady-state scheduling decisions. Compared to this, however, closer experiments on bare-metal reservation [5], boot-time prediction [6], and multi-mode server architecture [7] reveal the workings of operational variables that are usually apparently concealed in VM-based studies. The literature thus possesses an ontological split, between the various works on control that focus on the dynamism of placement and infrastructure provisioning, and those works aiming at the provisioning cost of activating infrastructure, reservation policy, and provisioning latency.

Another recurring pattern is the move from host-level energy models toward cross-layer decision frameworks. The studies of the early consolidation mostly concerned the reduction of the active host [8], [9]. Subsequent research used performance guarantees [10], I/O behaviour [11], thermal conditions [16], or prediction-based anti-correlation [14]. The problem was further extended by hybrid orchestration studies where it was demonstrated that even the abstraction layer affects the energy-performance trade-off [18], [19]. This trend indicates that energy conscious bare-metal provisioning fits in a layered decision model with no fewer than four interconnected steps workload characterization, provisioning-mode selection, runtime placement or consolidation, and release or reconfiguration. Energy efficiency suffers upon optimization of any of those stages separately. A scheduling policy can be effective in a VM pool, and yet globally worse than it would have been in a physical node that was reserved in the first place.

The other methodological pattern is evaluation strategy. Numerous designs are based on the use of simulation and historical trace due to the challenging aspect of execution of controlled experiments of full scale of production data centers [4], [14], [15], [19]. This method has the advantage of being able to compare many scenarios, but can typically obscure image-distribution delay, cost of transitioning through a power state, and fragmentation of orchestration. This limitation is partly resolved by the research on MaaS boot-time prediction [6] and bare-metal reservation [5], where time-to-readiness and pre-reservation policy are represented as quantifiable variables. The more powerful methodological agenda would be to combine the latter variables with the more sophisticated optimization schemes already obtained in virtualized scheduling. This integration is still not complete in the existing literature, which is one of the reasons why the provisioning-time and the run-time policies are often researched separately.

The methodological demonstrations hence favor a centralized premise of which provisioning of energy-conscious virtualization data centers is arranged by decision limits as opposed to resource type by itself. The involved limits include admission, profiling, mode selection, activation, scheduling, thermal stabilization, and deprovisioning. The use of bare metal would be beneficial in cases where predictable performance, low interference, or high startup amortization is more important than the flexibility benefits of virtualization. Virtualization will be beneficial in cases where the elasticity in terms of consolidation governs the cost of activation and in cases where migration cost becomes prohibitive. The main theoretical point to note is that energy-conscious bare-metal provisioning should be viewed as selective physical-resource admission within a heterogeneous control plane as opposed to anti-virtualization design.

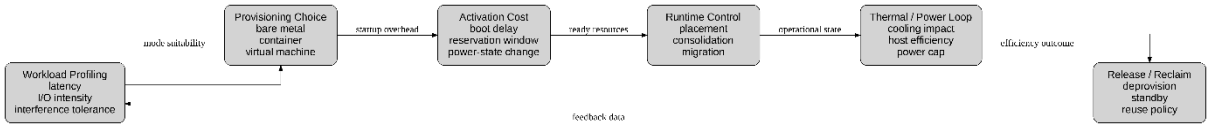

Figure 1: Conceptual Framework

IV. DISCUSSION

According to the reported literature, energy-conscious provisioning performance is based on the balance between the consolidation benefits and the activation cost or interference cost. Early work on energy saving in cloud data centers found that idle servers are a major source of wasted energy, and that consolidation can help bring that down (first-order control) [8], [9]. Innovatively placed performance and global-state knowledge integrated into VM allocation enhanced that image later as performance and state-of-the-world consciousness were embedded in placement studies [10], [12]. However, the same literature informs us that consolidation is not always a good thing. Foreseen savings can be offset in I/O intensive workloads, thermal hotspots, and control loops which use up resources during migration [11], [16]. With this kind of effects

taken into consideration, bare-metal provisioning becomes an analytically interesting option due to the possibility of a direct physical allocation that evades the constant runtime penalties that will accrue within over-optimized virtualized pools.

Of particular concern to this conclusion are provisioning-time studies. It is observed in bare-metal reservation analysis [5] that reactivity and energy efficiency are in conflict: maintaining machines in physical form is beneficial to response time yet may leave machines idle or partially unoccupied often. The above result will be complemented by MaaS provisioning prediction [6], which seeks to prove the possibility of modelling boot-time uncertainty and hence use it in making policy decisions. The reason why such policy is important is further explained in architecture selection work [7]. When bare metal is materially superior in terms of initiating or operational characteristics of a specified mix of workloads, then deploying a workload to a VM cluster under the guise of apparent flexible virtual resources may escalate overall power consumption costs when the cost of an interim status before successfully executing service startup is factored, or the cost of a slow startup transfer, or more intensified consolidation interference are accounted. This literature suggests in turn a more general energy accounting framework according to which activation latency and provisioning readiness should be treated as real energy variables, not just background operational detail.

The literature on comparative VM-placement provides a finer point, as opposed to contradiction. Power-aware and performance-guaranteed placement [10], anti-correlated prediction-based placement, reinforcement-learning consolidation, and collaborative VM optimization [14], [15], and [17] all achieve material efficiency gains over simple baselines. The results of such findings substantiate the validity of the fact that virtualized infrastructures are highly versatile and capable of providing robust energy savings in the majority of demand patterns. The gains related to it, however, are contingent upon the quality of prediction, process of workload characterization, and mobility or interference costs. In other words, the same things that make advanced virtual resource control work well also make the case for selective bare-metal provisioning stronger, not weaker. When the literature acknowledges that the quality of the placement is important, then it is rational to consider a "bare metal" as yet another placement choice in the optimization space, particularly on workloads that happen to be giving rise to ill-sharing behaviour.

This interpretation is supported by studies that are proportional to thermal and energy. EASE-infused scheduling [13] focuses on the fact that host efficiency varies with the servers' pool, whereas thermal sensitive scheduling [16] demonstrates that local locations may affect the cooling and hotspot behavior. These results suggest that the decision between bare-metal node as opposed to a virtualized host is not only a matter of the preference of abstraction but also a matter of fleet heterogeneity and thermal topology. A node with a low physical placement, whose operation is close to an optimal efficiency point, can be more competitive than a more nominal -utilized virtualized cluster whose virtualization host nodes acquire thermally inefficient states. This is among the reasons why global reasoning is better in comparison to host-local reasoning in various studies [12], [16]. Bare-metal provisioning makes the most sense when the controller needs to avoid scheduling energy-sensitive workloads on already-loaded virtual hosts or those with thermal issues on overloaded virtualized networks.

Containerized studies and hybrid studies enhance the argument by indicating that the cost of migration and orchestration varies with the level of abstraction. Khan et al. work [18] resource-consolidation allows, in certain conditions, to be more energy- and performance-efficient than VM migration, with the workload-aware orchestration of heterogeneous hybrid resources yielding a significant efficient workload, as reported by HeporCloud [19]. The implication of such outcomes is not that containers or hybrid scheduler replace bare metal. Rather, the jointly occurring evidence allows concluding that energy-conscious infrastructure management needs to consider the entire spectrum of provisioning forms. Bare metal is especially applicable because the characteristics of a workload required to be placed directly (i.e. expected residence time and interference characteristics) make sense, but containers and VMs also are good when elasticity and sharing are priorities. The overall analytical lesson is that one provisioning layer cannot be assumed to be the best in all classes of workloads.

One major restriction in the interrelated literature is the division between the provisioning policy and the operational feedback. Bare-metal reservation and provisioning-delay studies optimize pre-runtime choice whereas many studies have used placement, consolidation or orchestration after resources are instantiated. This division brings about a blind spot. The workload can be assigned to a VM because of immediate availability and is then subject to the repetitive migration or throttling due to interference and then ends up consuming more energy than a more appropriate slower physical allocation would have consumed. On the other hand, excessive reservation of bare-metal nodes will reduce the utilization to an extent of negating the anticipated benefit. The literature suggests that a successful controller should have a connection between the classification of admission-time and runtime feedback of thermal status, frequency of migration, realized usage and realized energy proportionality. This type of cross-layer integration is not developed in the published journal literature.

Another issue worth addressing is how results are measured across studies. The published literature reports on total energy, the violation of SLA, the number of migrations, the duration of the period during which they were provisioning, the

completion of the workload, or the degradation of the performance [5], [6], [10], [15], [19] differently. Even though all metrics are informative, the field still lacks a consistent evaluation framework for energy-aware bare-metal provisioning in virtualized data centers. More practical composite assessment would also entail time-to-ready, energy-per-successful-deployment, the energy-recovery time of migration, interference-corrected throughput, thermal penalty, reservation ineffectiveness. In the absence of such measures, the term energy-efficient will be prone to partial understanding. A VM-only policy can look efficient when judged by how many active hosts it brings down, and the mixed policy can seem to be less efficient even though it can represent a smaller total lifecycle energy of the interference-sensitive workloads. This is one of the most evident research gaps that is observable in the reviewed corpus.

Table 2: Method Comparison

Ref	Method	Strengths	Limitations
[6]	Provisioning-time prediction using platform-monitoring data	Captures bare-metal activation delay explicitly and improves reservation planning	Focused mainly on readiness estimation rather than full lifecycle energy optimization
[7]	Empirical architecture recommendation across resource types	Compares bare metal, containers, and VMs within a unified selection framework	Limited scope relative to full-scale heterogeneous production fleets
[9]	Adaptive heuristic VM consolidation	Strong baseline for reducing active-host count while considering SLA impact	Migration cost can accumulate under volatile workloads
[10]	Power-aware and performance-constrained placement optimization	Balances energy and service quality more directly than pure consolidation	Primarily centered on virtualized resources
[11]	I/O-aware scheduling policy	Exposes a major source of hidden inefficiency in virtualized environments	Less suited to non-I/O-dominant workload classes
[12]	Global energy-aware placement strategy	Incorporates broader system state and avoids overly local decisions	Increased control complexity and data requirements
[13]	Energy-proportional scheduling based on measured host behaviour	Exploits server heterogeneity and avoids uniform assumptions about efficiency	Requires prior measurement and stable host characterization
[14]	Predictive anti-correlated VM placement	Reduces contention and service violations through workload prediction	Prediction error can degrade effectiveness
[15]	Reinforcement-learning consolidation	Adapts to dynamic environments and uncertain demand patterns	Training cost and policy interpretability remain concerns
[16]	Power- and thermal-aware scheduling	Integrates cooling-aware reasoning into resource management	Thermal models can be difficult to calibrate accurately
[19]	Hybrid heterogeneous orchestration	Demonstrates practical gains from workload-aware mixed-resource control	Evaluation assumptions may not cover all provisioning-delay realities

Table 2 reveals that a large percentage of methodological advances has been achieved, yet the majority of methods continue to optimize on a limited abstraction layer. The techniques that unambiguously combine provisioning mode, cost of activation, and runtime adaptation are rather limited.

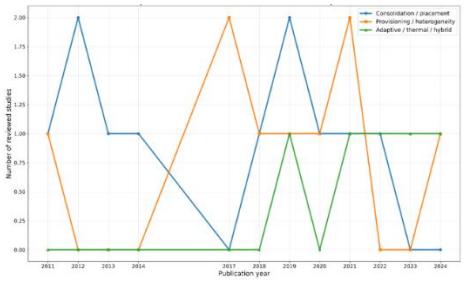

Figure 2: Distribution of Reviewed Journal Studies By Year

The overview of how the read journal literature is shifting in the scenarios of early consolidation and energy modelling to the alternative of prediction, thermal mindfulness, hybrid orchestration, and bare-metal-mindful provisioning is summarized in figure 2. The discussion is supported by the graph that demonstrates that the subsequent work considers heterogeneity, and activation behaviour as fundamental energy-management issues instead of fringe operational specifics.

Table 3: Results Comparison

Ref	System	Metric	Outcome
[5]	Cloud bare-metal reservation environment	Reactivity versus energy efficiency	Shows that faster response from pre-reserved physical nodes can reduce reactivity cost but may weaken energy savings if reservation is excessive.
[6]	Metal-as-a-Service platform	Provisioning and booting time	Demonstrates that startup delay can be predicted and therefore incorporated into resource-selection policy.
[7]	OpenStack-based mixed server environment	Startup and benchmarked performance	Finds clear performance and startup differences across bare metal, containers, and KVM virtual machines.
[10]	Virtualized cloud placement system	Power and application performance	Reports improved placement through joint power-performance treatment relative to energy-only mapping.
[11]	I/O-virtualized cloud environment	Energy loss from I/O virtualization overhead	Indicates that I/O-aware scheduling can recover efficiency lost to virtualization overhead.
[13]	Virtualized cloud data center	Energy efficiency and proportionality	Shows that host-specific energy proportionality improves scheduling decisions.
[14]	Predictive VM placement environment	Energy use and service violations	Reports lower energy demand and fewer violations through anti-correlated predictive placement.
[15]	Cloud data-center consolidation controller	Energy-performance balance under dynamic load	Demonstrates that reinforcement learning can improve VM distribution in uncertain environments.
[16]	Cloud data center with thermal coupling	Power and thermal optimization	Shows benefit from joint thermal and power-aware scheduling.
[18]	Containerized data center	Energy and performance efficiency under migration choices	Finds differing efficiency patterns across VM, container, and application migration, underscoring abstraction-layer sensitivity.
[19]	Hybrid heterogeneous cloud data center	Energy saving and workload performance	Reports up to substantial energy savings while maintaining acceptable performance under workload-aware orchestration.

Table 3 also supports one of the key findings of this review: the positive outcomes are the most stable when energy management includes heterogeneity, cost of activation and the nature of workload and not only use of the host. The fact that conclusion is linked directly to the argument of considering bare-metal provisioning an active energy-management tool within virtualized data centers.

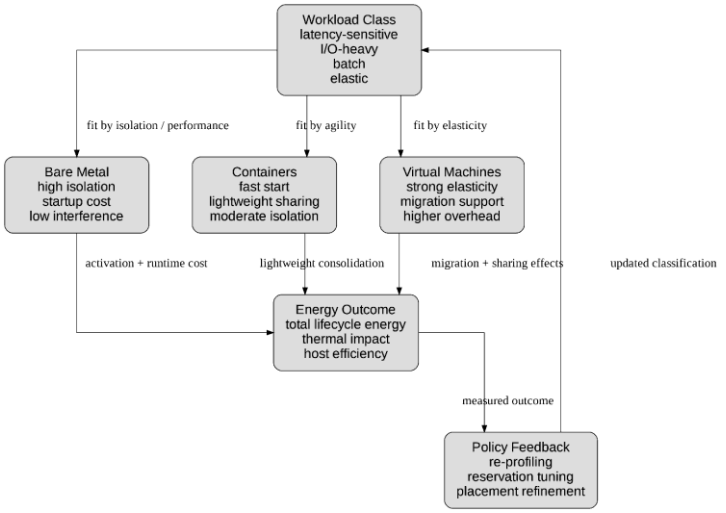

Figure 3: Relationship Diagram For Provisioning-Mode Selection in Heterogeneous Data Centers

Figure 3 plots the causal relationship between workload characteristics, the choice of provisioning and the results of the energy. The diagram points out the reason why bare metal, containers, and virtual machines should be considered as related options regulated by the fitness of workloads, startup latency, and the side effects of the consolidation but not by strict infrastructure preference.

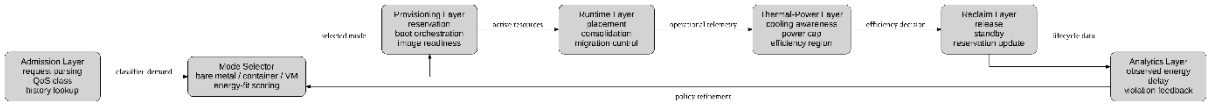

Figure 4: Integrated Model For Energy-Aware Bare-Metal Provisioning in Virtualized Data Centers

Figure 4 incorporates the literature into a cycle of control flow of admission, reservation of bare metal, virtualization scheduling, thermal monitoring and reclamation. This model contributes to the primary argument that provisioning-time decisions, as well as run-time decisions, need to sustain each other over an extended infrastructure lifecycle in case energy efficiency is to be considered.

V. FUTURE DIRECTIONS

One clear direction for future work is tighter integration between provisioning-time decisions and run-time control. The significance of bare-metal reservation windows, boot-time prediction, VM consolidation, thermal scheduling, and hybrid orchestration have already been demonstrated using journal studies [5], [6], [9], [16], [19]. Nonetheless, to a large extent, this work is compartmentalized. Future study ought to look at the controllers which make intelligent decisions simultaneously on whether a workload should be on bare metal, a container or a VM and updating those decisions as telemetry continues to come in. These designs would not be limited to initial placement logic but instead would be lifecycle-aware orchestration as it considers energy over both admission, execution, and release instead of looking at it over just one stage.

A second area that needs work is how results are measured. Up-to-date literature tells of useful measures, but a typical benchmark package that would match the heterogeneous provisioning is yet to be evaluated. More work should be done in the future, which involves time-to-ready, reservation inefficiency, migration energy recovery time, interference-adjusted throughput, thermal overhead/thermal penalty, and deprovisioning cost besides total energy. With these measures in place there would be a possibility of comparison between a bare-metal-first policy and a VM-first or container-first policy based on common assumptions. Even in the absence of that step, evaluation is still prone to close interpretations where one policy appears better due to the omission of important lifecycle costs in the measurement frame.

The third opportunity is predictive and learning-based orchestration which includes resource-mode selection as a direct input. Adaptive potential of VM consolidation has already been shown in terms of reinforcement-learning studies [15]. Similar efforts are now to be made in the case of the increased action space of heterogeneous infrastructures, such as in physical-node reservation, in warm-standby policy and in fragmentation of a mixed fleet. This orientation is difficult in a methodological manner since the exploration cost in the production settings can be substantial. Nevertheless, controlled simulation supported by actual trace data and confirmed power models may offer a robust middle path. These studies would be of interest especially in clouds with workloads classes that reoccur too that learned provisioning templates are reasonable.

One should also pay more attention to interdisciplinary work. Thermal engineering, analysis of hardware architecture, operations research, and service-economics analysis also find more and more applications in energy-aware provisioning. It is already indicated by the research on selection of server architecture [7], thermal-conscious scheduling [16], and energy-performance-cost-trade-offs [20]. It is also the case that future scholarship ought to unite the use of system modelling with more empirical constraints to deployment, including, but not limited to supply-chain variability, cooling topology, carbon-conscious scheduling, policies that govern hardware lifecycle management. This integration would aid in the shift of the field away to searching in abstract data centers like algorithmic optimization and into operation decision support of heterogeneous facilities in the real world.

VI. CONCLUSION

This review shows that energy-conscious bare-metal provisioning in virtualized data centers is not a side issue but a natural outcome of how modern infrastructure is becoming increasingly heterogeneous, and of the understanding that the overhead of virtualization, the migration cost, the boot delay as well as the thermal behaviour have significant impact on the overall lifecycle energy. Pre-consolidation and location experiments laid the groundwork stipulating critical tenets towards saving on wastes by ensuring better utilization [8]-[10]. Subsequent research on predicting provisioning, the mixed-resource architecture, resource mode-proportional scheduling, thermal control, and hybrid orchestration broadened this perspective and demonstrated that resource mode-proportional workload operation is an efficient operating mode [5]-[7], [13], [16], [19].

A consistent finding across the literature is that reducing the number of active hosts is only a partial measure of energy efficiency. When the allocation policies take into consideration the readiness delay, performance isolation, interference, migration side effects, and thermal state, favourable results are attained. Bare metal is useful in such an environment as direct hardware reservation can sometimes reduce full-lifecycle energy of workloads of choice where such allocations would otherwise seem less accommodating than virtualized allocations would seem to be. Meanwhile, indiscriminate physical-node reservation can undermine savings, and this is why bare metal should not be considered a default or an exemption, it should be part of a greater heterogeneous control policy.

The key gaps of the research are apparent. There is still limited cross-layer research on the mode selection at admission time and runtime feedback, metric standards remain incomplete and product-scale testing of heterogeneous provisioning policies is in progress. That notwithstanding, it can be seen that the trend of the literature is clear cut. Data center control will become more energy conscious in the future, and less reliant on more exclusive optimization of virtualization and more based on the orchestration coordination of bare metal and containers and virtual machines as part of coherent and feedback-driven orchestration frameworks.

- **Interest Conflicts:** The author declares that there is no conflict of interest concerning the publishing of this paper.

VII. REFERENCES

[1] Kaur, T., & Chana, I. (2015). Energy efficiency techniques in cloud computing: A survey and taxonomy. ACM Computing Surveys, 48(2), Article 22.

[2] Sharma, Y., Javadi, B., Si, W., & Sun, D. (2016). Reliability and energy efficiency in cloud computing systems: Survey and taxonomy. Journal of Network and Computer Applications, 74, 66–85.

[3] Silva Filho, M. C., Monteiro, C. C., Inácio, P. R. M., & Freire, M. M. (2018). Approaches for optimizing virtual machine placement and migration in cloud environments: A survey. Journal of Parallel and Distributed Computing, 111, 222–250.

[4] Castañé, G. G., Núñez, A., Llopis, P., & Carretero, J. (2013). E-mc2: A formal framework for energy modelling in cloud computing. Simulation Modelling Practice and Theory, 39, 56–75.

[5] de Assunção, M. D., & Lefèvre, L. (2018). Bare-metal reservation for cloud: An analysis of the trade-off between reactivity and energy efficiency. Cluster Computing, 21(2), 1289–1300.

[6] Sîrbu, A., Pop, C., Şerbănescu, C., & Pop, F. (2017). Predicting provisioning and booting times in a Metal-as-a-Service system. Future Generation Computer Systems, 72, 180–192.

[7] Yamato, Y. (2017). Performance-aware server architecture recommendation and automatic performance verification technology on IaaS cloud. Service Oriented Computing and Applications, 11(2), 121–135.

[8] Lee, Y. C., & Zomaya, A. Y. (2012). Energy efficient utilization of resources in cloud computing systems. The Journal of Supercomputing, 60(2), 268–280.

[9] Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. Concurrency and Computation: Practice and Experience, 24(13), 1397–1420.

[10] Zhao, H., Wang, J., Liu, F., Wang, Q., Zhang, W., & Zheng, Q. (2018). Power-aware and performance-guaranteed virtual machine placement in the cloud. IEEE Transactions on Parallel and Distributed Systems, 29(6), 1385–1400.

[11] Xiao, P., & Liu, D. (2013). An energy-efficient virtual machine scheduler with I/O collective mechanism in resource virtualisation environments. International Journal of Networking and Virtual Organisations, 13(4), 311–326.

[12] Balaji, R. (2025) 'Autonomous SRE: SLO-Aware Multi-Agent Root-Cause Analysis and Verified Self-Healing for Cloud-Native Systems', ESP Journal of Engineering & Technology Advancements, 5(4), pp. 211-221. doi: 10.56472/25832646/JETA-V5I4P128.
[13] Feng, H., Deng, Y., & Li, J. (2021). A global-energy-aware virtual machine placement strategy for cloud data centers. Journal of Systems Architecture, 116, Article 102048.
[14] Balaji, R. (2024) 'Threat-Model-Driven Zero-Trust Authorization for Microservice APIs in Kubernetes-Based Cloud-Native Applications', ESP Journal of Engineering & Technology Advancements, 4(4), pp. 223–234. doi: 10.56472/25832646/JETA-V4I4P128.
[15] Qiu, Y., Jiang, C., Wang, Y., Ou, D., Li, Y., & Wan, J. (2019). Energy aware virtual machine scheduling in data centers. Energies, 12(4), Article 646.
[16] Shaw, R., Howley, E., & Barrett, E. (2019). An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions. Simulation Modelling Practice and Theory, 93, 322–342.
[17] Shaw, R., Howley, E., & Barrett, E. (2022). Applying reinforcement learning towards automating energy efficient virtual machine consolidation in cloud data centers. Information Systems, 107, Article 101722.
[18] Chen, R., Liu, B., Lin, W., Lin, J., Cheng, H., & Li, K. (2023). Power and thermal-aware virtual machine scheduling optimization in cloud data center. Future Generation Computer Systems, 145, 578–589.
[19] Wang, B., Liu, F., Lin, W., Ma, Z., & Xu, D. (2021). Energy-efficient collaborative optimization for VM scheduling in cloud computing. Computer Networks, 201, Article 108565.
[20] Khan, A. A., Zakarya, M., Buyya, R., Khan, R., Khan, M., & Rana, O. F. (2021). An energy and performance aware consolidation technique for containerized datacenters. IEEE Transactions on Cloud Computing, 9(4), 1305–1322.
[21] Khan, A. A., Zakarya, M., Rahman, I. U., Khan, R., & Buyya, R. (2021). HeporCloud: An energy and performance efficient resource orchestrator for hybrid heterogeneous cloud computing environments. Journal of Network and Computer Applications, 173, Article 102869.
[22] Katal, A., Dahiya, S., & Choudhury, T. (2023). Energy efficiency in cloud computing data centers: A survey on software technologies. Cluster Computing, 26(3), 1845–1875.