

DATA CENTERS

Enhancing Power-System Reliability and Energy Resilience in Mission-Critical Data Centers in Saudi Arabia

Waqas Ahmad 

Published October 2026

Abstract

Mission critical data centres are now becoming strategically important electrical loads as Saudi Arabia ramps up its cloud computing, AI, digital government, fintech and hyperscale infrastructure. Their operational requirements are extremely demanding in that computing services must be available in the face of utility disturbances, equipment faults, renewable variability, thermal stress and constraints on fuel and water. This review brings together the peer-reviewed research on data centre electrical integrity, demand flexibility, microgrids, energy storage, renewable integration, thermal management, predictive maintenance and power system resilience that has been published between 2020 and 2025 with a view to its specific relevance to the Saudi operating environment. The evidence shows that resilience is best achieved when redundancy is viewed as a coordinated multi-layered architecture rather than simply as duplicated equipment. Resilience can be achieved through a combination of dual utility paths, modular UPS, battery energy storage, dispatchable backup generation, microgrid controls, workload aware demand response and climate appropriate cooling, all of which can reduce the number of single points of failure and at the same time support grid stability. The Saudi context brings a number of additional design pressures in the form of very high ambient temperatures, water scarcity, the very rapid ramp up in penetration of PV and the need to balance low carbon operation with the requirement for assured ride through. The review therefore proposes a resilience by coordination framework which links electrical topology, energy resources, computing flexibility, thermal systems, maintenance analytics and grid services. There remain a number of research gaps which are specific to the Saudi context: reliability data sets, coupled electrical-thermal contingency models, long duration storage, cyber-physical failure propagation and validated operating strategies for high renewable penetration grids. The findings of this review form a research based basis for resilient data centre planning in the context of the Vision 2030 digital growth agenda.

1 Introduction

Data centres have gone from being massive consumers of commercial electricity to being mission critical infrastructure. The loss of a data centre can have a knock-on effect in terms of the disruption to banking, government services, telecoms, health platforms, industrial control and the myriad of cloud-based business processes that we now rely on. This is especially relevant in the context of Saudi Arabia where digitalisation is happening at pace and there's a huge expansion of renewable generation as well as a massive amount of new urban and industrial development. A modern data centre has to meet two requirements at the same time: it has to provide continuous, high quality power to the computing loads and it has to be able to operate in a power system which is itself becoming more dynamic, more converter based and more reliant on variable resources. Electrical continuity is no longer just a problem of having a backup generator.

Recent data centre research shows that reliability is very much a function of architecture. Having dual utility feeds, selectively redundant distribution paths, modular UPS and coordinated restoration will all reduce your exposure to single point failures, provided of course that your protective devices, transfer logic and operating procedures are all designed to maintain fault isolation [1]. At the same time data centres have an unusual degree of demand side flexibility. Your UPS batteries, your controllable cooling, your server power caps and your schedulable workloads can all be used to alter net demand on the system in seconds to hours. This means that you can in fact participate in frequency response and demand response without necessarily having to sacrifice service quality [2-4]. That's a complete reversal of the traditional view of the data centre as simply a load that needs to be protected by the grid. When properly controlled it can in fact become a resource that contributes to grid reliability.

The reliability issue gets worse as you add onsite generation and storage. The energy networks of the data centre are increasingly going to be a mix of PV, batteries, dispatchable generation, cooling and workload. Planning studies have demonstrated that reliability and operating economics are both affected by the uncertainty of renewable output, the computational load, component outages and the ability to shift workload during a contingency [8-11]. Storage can provide short duration ride through, frequency support, renewable smoothing and black start capability but the storage technologies themselves will have degradation, SoC constraints and failure modes which have to be taken into account in the resilience design [12,13].

Saudi Arabia is a very different application environment. You have very high solar availability which is obviously a huge opportunity for PV integration, but you also have very high ambient temperatures which in turn increases the cooling load at the same time as you can have elevated electrical loads. The hot arid climate creates water energy trade-offs in terms of heat rejection so the cooling strategy becomes an integral part of the electrical resilience rather than a separate issue of efficiency [14-19]. National renewable deployment will also change the frequency and voltage characteristics so there's a greater need for flexible loads and fast storage to maintain system stability [20-24]. There are already some Saudi specific studies which point to the benefits of hybrid PV/storage systems, demand side management and resilient microgrids in terms of improved reliability and reduced energy cost. The performance of these systems is however very much dependent on the design assumptions, the dispatch rules and the local conditions [25-30].

This review looks at all of those interacting elements. It doesn't treat the UPS, the renewables, the cooling and the demand response as separate issues. It's about how you can integrate them all into a layered architecture of resilience for the mission critical data centres in Saudi Arabia. The key idea is that you get better reliability if you design energy, computing and thermal flexibility as complementary assets with a clear focus on failure containment, recovery and grid support.

2 Aim and research objectives

The objective of this review is to provide a technically sound understanding of how mission critical data centres in Saudi Arabia can improve power system reliability and facility level energy resilience through the integration of renewable energy, storage, flexible computing loads and climate responsive cooling.

The four objectives of the review are: to assess the range of electrical architectures and operational measures that can be employed to protect critical data centre loads from utility outages, internal faults and power quality disturbances; to examine the potential for coordination of UPS batteries,

battery energy storage systems, dispatchable generation, renewable resources and microgrid controls across both short and long duration interruptions; to assess the role of workload scheduling, server power management and cooling flexibility as controllable resources for both internal resilience and for provision of services to the external grid; and to interpret the evidence in the context of Saudi Arabia where the issues of high temperature operation, water availability, increasing penetration of PV and the need for resilience metrics that are a composite of availability, recovery, cost and environmental performance are all significant. The review is designed to clearly separate the concepts of redundancy and resilience by focusing on fault containment, graceful degradation, recovery sequencing and protected operating reserves under uncertainty.

3 Review methodology

This was a structured integrative review using a methodology appropriate for an interdisciplinary topic which cuts across power engineering, data centre infrastructure, energy systems and thermal management. The review design was derived from the structure of the papers provided as references: a clearly defined scope, explicit search and inclusion/exclusion criteria, thematic categorisation, comparative synthesis, research gap analysis and a concluding design framework. The papers provided were used solely as references for structure and style of writing; the branding of the papers in terms of their publication and the labels they used for their own methodology were not carried over into the manuscript.

The searches were built around five concept groups: data centre or mission critical computing; reliability, resilience, availability, power quality and contingency; UPS, battery storage, microgrid, renewable energy and distributed generation; demand response, workload flexibility, frequency regulation and grid services; and cooling, thermal management, hot/arid climate and water energy interaction. Boolean combinations of the concepts within each group were constructed using AND/OR operators. The priority databases and publisher indexes included the Scopus indexed journal collections, the Web of Science indexed sources, IEEE Xplore, ScienceDirect, the Wiley/IET collections and other peer reviewed journal collections. The publication window was set to 2020-2025 to match the evidence horizon and to capture the more recent developments in hyperscale computing, inverter based resources, battery systems and grid interactive data centres.

Eligible studies were peer-reviewed journal articles or substantial English language review papers relating to at least one of the following: data centre electrical continuity; power system support from data centre flexibility; data centre microgrids; storage reliability; renewable integration; thermal energy coupling; or Saudi Arabian renewable and resilience applications. Studies which were solely about general building efficiency, generic cloud computing performance or information security resilience (with no energy or power system element) were excluded. Conference papers only, editorials, vendor material and unverifiable web content were also excluded from the core reference set. A study set in a non-Saudi climate or on a non-Saudi grid could be included provided the mechanism was transferable (e.g. hot arid cooling, frequency response, microgrid islanding, battery reliability, workload shifting).

The synthesis didn't produce any database hit numbers or screening totals. The studies were assessed qualitatively on a number of criteria: relevance, methodological transparency, reproducibility of the modelled architecture, clarity of the reliability and energy metrics and utility for mission critical operation. The evidence was then grouped into six analytical themes: electrical integrity and redundancy; storage and backup resources; grid interactive workload flexibility; coordination with

renewables and microgrids; thermal-electrical coupling; and digital maintenance and operational analytics. The Saudi specific literature was treated as a contextual layer across the themes rather than as a discrete body of evidence.

Thirty references were selected for the final manuscript, all from 2020-2025 and each checked to ensure a resolvable DOI could be found against publisher or bibliographic records. The set is a mixture of data centre specific work and more fundamental work on microgrid resilience and the integration of renewables in Saudi Arabia. This is a necessary mix as there is very little published work on data centre reliability in Saudi Arabia per se, though there are a number of general mechanisms which are of direct engineering relevance. The review therefore makes a distinction between the data which is actually observed in the data centres and the data which is reasoned to be applicable in Saudi Arabia. The thematic findings were then assessed for convergence and divergence. For instance, increased storage can improve ride-through but at the cost of increased degradation risk. Water saving cooling can lead to an increase in electrical demand. The final framework emerged from the interplay of the themes rather than from any one technology in particular.

4 Reliability and resilience requirements of mission-critical data centers

Reliability and resilience are related concepts but not the same thing. Reliability is about the probability that the electrical and thermal systems will perform their intended function under normal conditions. Resilience goes further in that it also takes into account the ability to absorb an abnormal event, maintain critical service and recover quickly. For a data centre the resilient design will therefore have to be able to deal with both the prevention of failure and with controlled degradation. An architecture that is highly redundant in that it stays online after a single fault, but which is not in any way resilient in that it cannot safely recover from a cascade of failures is very redundant indeed and yet not at all resilient.

Electrical integrity starts with a fault tolerant topology. Work on critical load distribution has highlighted the need for multiple utility paths, sectionalised buses, selective protection, maintainable redundancy and restoration logic that will not allow a local fault to escalate to a complete facility outage [1]. The UPS provides the link between the sub-second and minute time scales of grid disturbance and alternative supply, with batteries also being able to provide frequency response or peak management [2,3]. The battery strings themselves are however failure prone assets. Predictive techniques based on operational voltage and resistance data can be used to detect the onset of VRLA failure and thus reduce the chance that the nominal redundancy will in fact be hiding a degraded reserve capability [13].

A good resilience assessment must look at a number of timescales. The millisecond to second timescale covers things like voltage sags, protection operations, converter trips and frequency swings. The minute to hour timescale includes utility outages, cooling loss, generator starts, fuel constraints and battery discharge. The multi-hour to multi-day timescale is where you have the issues of renewable uncertainty, fuel supply, equipment repair, water availability and the need to relocate work. Each of these timescales will have different resource requirements and different control priorities. A fast UPS response is not going to be able to substitute for long duration energy adequacy and diesel autonomy is not going to be able to fix a bad fault selectivity or a failed static transfer switch. The mission critical design should therefore be viewed as a chain of dependencies where the weakest co-ordinated transition in that chain may be the one that ultimately determines the availability. The reliability metrics we use therefore need to go beyond simple annual availability.

Useful metrics would be loss of load probability, expected unserved critical energy, transition success, battery autonomy, generator start reliability, thermal ride through and recovery time. Scenario based testing should combine equipment failures with maintenance states, extreme weather and degraded communications as these conditions will expose the dependencies that a simple single component calculation will not. These metrics should be looked at on a critical service tier basis rather than aggregated load.

5 Saudi Arabian operating context

Saudi data centres are operating in a high growth digital environment with very challenging environmental conditions. The solar resource is a major attraction for PV generation and there is a wealth of work on the technical and economic potential of distributed PV across the country [26]. The issue with PV is that it is variable and inverter based so its contribution to resilience is highly dependent on storage, control, curtailment capability and the ability to island critical loads. The various studies on PV-battery and hybrid renewable systems in Saudi Arabia have shown that the cost-optimal system configuration is not necessarily the reliability-optimal one. A system that is designed to minimise energy cost can still have unacceptable loss of load performance and this will only be avoided if there is sufficient dispatchable capacity or sufficient storage included [28,29].

The National Grid is also heading towards a greater proportion of non-synchronous generation which in turn increases the value of fast frequency support and controllable demand. Work on modern power systems has shown that a reduction in rotational inertia can lead to an increased rate of change of frequency and a reduction in disturbance margins [23,24]. A similar study in the Saudi context looking at PV penetration in Hail identifies battery based support as a way of improving frequency stability [30]. Mission critical data centres can of course be a part of this new flexibility requirement via their UPS reserves, batteries, cooling modulation and workload scheduling, as long as the service level constraints are maintained.

Climate is just as important. Hot ambient conditions increase cooling power and reduce the annual value of simple air-side economisation. There's also hot arid data centre evidence that energy and water performance can go in opposite directions depending on the cooling technology you choose [14]. For the Saudi sites in particular, resilience planning must take into account electrical demand under extreme heat, water availability during extended events, dust and air quality exposure, chiller redundancy and the impact of thermal set points on IT reliability. Energy resilience in this scenario is entirely intertwined with thermal resilience. The factors mentioned all point towards a design which maintains reserve during the summer peaks and one which utilises local generation without assuming ideal weather. Regional planning must also take into account utility strength, fuel access, water logistics and the likelihood that a nearby facility will be impacted by a major grid disturbance.

6 Resilient power architecture for Saudi data centers

A resilient Saudi data centre should be viewed as a layered power architecture rather than a collection of backup devices. Where possible independent or physically diverse utility feeds will reduce common mode exposure. Diversity must be real: two feeds going through the same substation, the same corridor or the same protection scheme may not deliver the level of resilience suggested by a one line diagram. The internal medium and low voltage systems should be sectionalised, have selective coordination, have maintainable bus arrangements and have clearly defined transfer states so that equipment can be isolated without interrupting the critical IT loads [1].

The second layer is short duration ride-through. Modular UPS systems provide conditioned power during sags, switching events and the period prior to alternative sources coming online. UPS redundancy should be considered at the module, string, switchgear and control levels. Battery condition monitoring is of particular importance as capacity degradation can be masked during normal float operation. Data driven battery diagnostics can provide an earlier warning of failure enabling maintenance to be carried out prior to reserve capability being required [13]. Li-ion systems have a significantly higher energy density and cycling capability than traditional VRLA systems but their resilience case must include dependence on the battery management system, thermal propagation control, replacement strategy and state of charge availability at the instant of the disturbance.

The third layer is sustained on-site supply. Emergency generators are still of value in that they provide dispatchable power independent of solar conditions, but generator-only resilience is going to be constrained by generator start failures, condition of maintenance, fuel logistics and the length of the outage. Hybrid microgrid work has shown that a combination of PV, batteries and networked generators can deliver greater islanded resilience and lower life cycle cost than a diesel only solution when component reliability is represented in a realistic manner [25].

Layer 4 is microgrid control. Microgrid studies in data centres have shown that optimal operation will involve the co-optimisation of renewable generation, battery storage, conventional generation, cooling and workload scheduling [5-7]. Robust planning has also demonstrated that fault conditions and uncertain workloads should be factored into capacity selection at the planning stage rather than being dealt with as operational problems after the design is complete [9]. A microgrid controller for a mission critical facility would need to be able to manage a number of explicit operating modes: normal grid connected operation, disturbance ride-through, intentional islanding, degraded island operation, black start or sectional restart and controlled resynchronisation. Each of these modes would have associated requirements for reserve margins, battery depth of discharge, minimum generator commitment, cooling priorities and workload actions.

The fifth layer is computing flexibility. Some of the compute work can be deferred, power capped, migrated or rescheduled to create a controllable demand envelope [4,8,11]. During a power contingency this can be used to preserve power for the latency sensitive or safety critical services. It is important to note that workload control is not simply a mechanism for emergency load shedding. If it is incorporated into normal operation it can help to keep the batteries at a healthier reserve level, reduce peak demand and enable ancillary services. Work on frequency regulation has shown that coordinated control of IT and cooling loads can provide very rapid response to the grid [3] and static and dynamic data centre power modulation can improve grid frequency metrics [2]. The same controls can be resilience controls in the event of an external supply constraint.

Figure 1 puts all of these layers together into a Saudi mission critical architecture. Local protection and safe fallback states must be available if the higher level of optimisation is not available.

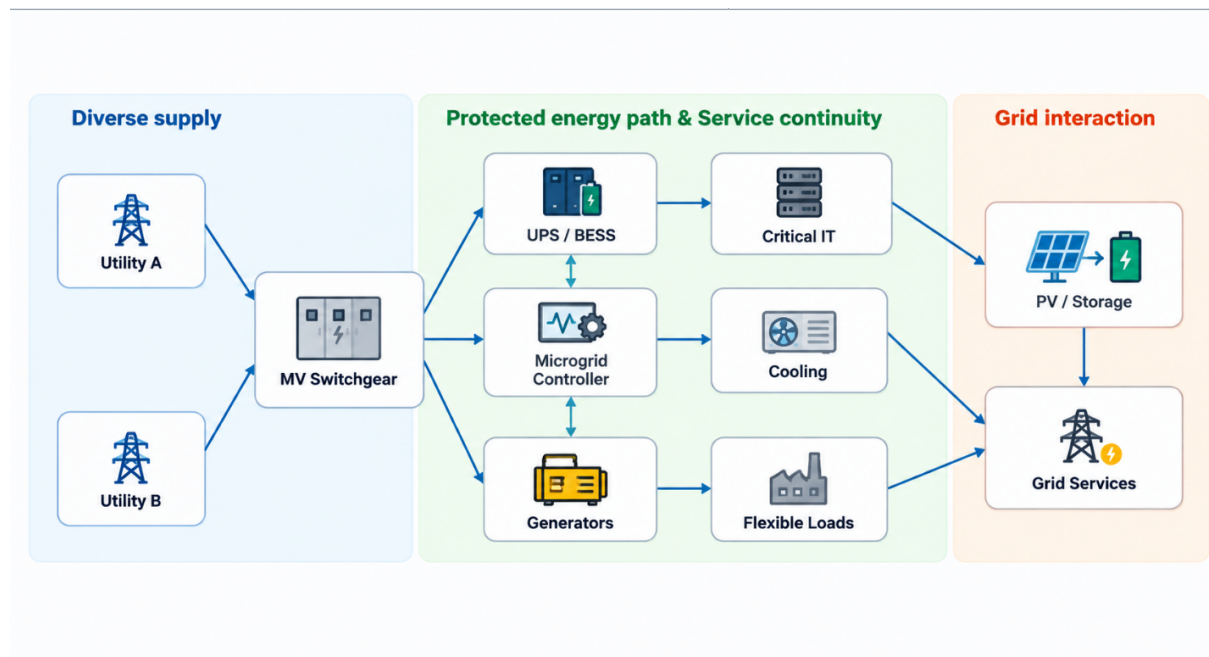


Figure 1. Layered resilient power architecture for mission-critical data centers in Saudi Arabia.

7 Renewable integration and grid-interactive resilience

Renewables turn the resilience problem from one of simple backup sizing to one of coordinated uncertainty management. PV can reduce the need for grid supply and fuel consumption during daylight outages, but it is not firm capacity unless you have storage, dispatchable generation or controllable demand. The Saudi PV literature is very supportive of the potential for widespread deployment, though the economics are highly variable depending on the region, tariff, system configuration and storage strategy [26]. The hybrid system studies in Al Jubail and Najran are no different in that they show the trade-off between cost and reliability. Adding in battery and dispatchable capacity can dramatically reduce the loss of load risk, but the least cost solution may not be sufficient to meet the continuity of supply requirements for a mission critical application [28,29].

For data centres the battery storage has three roles in terms of resilience. It can provide immediate ride-through during a transition from utility to generator (or vice versa). It can provide a buffer for the variability of renewables and thus provide a reserve for islanded operation. It can also provide support to the grid in terms of frequency and help to reduce local network stress under normal conditions. There is evidence from work on data centre parks and operational reliability that workload flexibility and storage coordination can improve both the economics of the facility and the performance of the wider power system [8-10]. The issue of course is reserve competition. A battery which has been discharged for energy arbitrage or for grid services may not have sufficient State of Charge to provide ride-through in the event of an outage. Mission critical controls should therefore have a resilience reserve floor which is a function of the forecasted risk, generator availability, weather and the expected duration of the outage.

Grid interactive operation must also take into account the evolving nature of systems with a high proportion of renewables. High levels of converter penetration can lead to reduced inertia and altered frequency and voltage characteristics [23,24]. Data centres can respond very quickly via their UPS inverters and controllable loads, but their participation must be constrained by service level agreements and electrical protection boundaries. A sensible approach would be to provide a small,

always available flexibility band for grid services and to protect that with a dedicated emergency reserve. That turns the facility from a passive protected load into an active participant in providing reliability, without opening the door to market forces which could in turn erode the resilience of the system. Forecast aware controls can be used to enhance this hierarchy by increasing the reserve prior to periods of high risk and reducing it when the utility is in a stable condition. The policies in place should be conservative in the face of forecast error and be coordinated with generator outages and cooling requirements.

8 Thermal-electrical coupling in hot-arid conditions

Cooling is a fundamental component of electrical resilience in that the loss of heat rejection can force a computing shutdown even when electrical generation is still available. At the local level, cooling demand is intrinsically linked to the weather and heat extremes can occur at the same time as grid peaks. Evidence from hot arid data centres shows that the choice of cooling system involves a direct trade-off between electricity and water use [14]. Evaporative assisted systems may reduce compressor power at the expense of increased water use, whereas dry or air cooled systems will reduce the onsite water demand but can result in higher electrical loads at high ambient temperature.

Research into free cooling and supply air optimisation shows that the climate dictates how much you can save. Air-side economisation is most effective in cooler, more humid climates and provides far less benefit in warm, dry climates [15]. Water-side free cooling can reduce the cooling energy required but this requires a very careful design of tower capacity, flow and temperature approach; in practice the actual performance achieved can be substantially below the ideal that would be predicted from the design [16]. Comparative studies of room, row and rack based cooling all point to the fact that the cooling architecture and the degree of airflow containment will have a significant impact on both thermal stability and facility energy consumption [17,18]. Life cycle environmental research also cautions against looking at cooling simply in terms of the electricity it consumes during operation as the impacts of equipment manufacture, refrigerants and water can all shift the overall environmental burden [19].

For the Saudi mission critical sites the resilience implication is very clear indeed: cooling must be in the same contingency model as the electrical supply. The battery autonomy calculations must include the cooling load as well as the IT load. Islanding strategies must define the level of allowable thermal set point relaxation, the level of staged cooling and the level of non-critical workload that can be shed. Where evaporative systems are used water storage or some form of alternative dry heat rejection may be required. Thermal sensing, airflow containment and predictive chiller maintenance will all help to extend the window of time available for a controlled response before the IT temperatures become the limiting factor.

9 Digital operations, predictive maintenance, and recovery

Resilience is a function of the state of the equipment at the time of the disturbance. A nominally redundant system can have hidden failures in the batteries, breakers, generators, sensors and controls which will create a latent common mode risk. Predictive maintenance therefore works in conjunction with architectural redundancy. Research into battery fault prediction has shown that operational monitoring can detect the precursors to failure well before a reserve event comes along and exposes them [13]. The same principles apply to the generator starting system, the cooling equipment, the transfer switches, the power converters and the protection devices.

Operational analytics should be a blend of condition data and resilience state. Rather than just reporting alarms you could have a digital operations platform that's able to calculate available ride-through time, generator readiness, battery reserve, cooling margin and the number of independent power paths currently in service. Maintenance activity should be reflected in real time as taking one UPS module or one bus section out of service will change the facility's contingency tolerance even if the customer load is the same. This is particularly pertinent in the high availability environment where the risk of maintenance induced outage can in fact be greater than the risk of random failure.

Recovery planning should be just as explicit. The aim after a major disturbance is not simply to get all the equipment back as quickly as possible, but to recover in an order which maintains stable power and cooling. Sectional restart, controlled load pickup, battery recharge limits and generator synchronisation all need to be planned in conjunction with the restoration of the workload. Exercises and digital simulations should be used to test degraded states, communication failures and operator handover so that resilience is demonstrated in operation rather than just by the number of pieces of equipment. Recovery is something that has to be validated repeatedly.

10 Critical synthesis: A resilience-by-coordination framework

The evidence reviewed is consistent with a resilience by coordination approach which comprises six interdependent domains. The first is electrical integrity (diverse utility connections, sectionalised distribution, selective protection and maintainable redundancy). The second is energy assurance (UPS, batteries, dispatchable generation, renewables and fuel/energy availability). The third is computing flexibility (workload prioritisation, deferral, migration and server power control). The fourth is thermal assurance (redundant cooling paths, appropriate climate heat rejection, water planning and thermal ride-through). The fifth is operational intelligence (condition monitoring, predictive maintenance, reserve estimation and mode aware controls). The sixth is grid interaction (frequency support, demand response and coordination of renewables within the protected resilience boundaries).

These domains should be designed against a sequence of disturbance states rather than a single availability target. In normal operation the aim is efficiency with reserve. In the event of a grid disturbance the fast controls are there to maintain voltage, frequency and IT continuity. If the disturbance continues then the facility will go into islanded operation, re-prioritise the workloads and dispatch storage and generation according to the expected duration of the event. In the case of a prolonged event energy adequacy, fuel, water, repair access and thermal conditions will become the dominant factors. Resync and recovery must be such as to avoid any new transients or simultaneous load pickup.

This framework changes the way we value redundancy. An extra generator might actually provide less resilience than better protection selectivity if both of the existing generators are on the same bus. A bigger battery might be of less value than better state of charge management if market dispatch is such that it's constantly depleting the reserve. More efficient cooling may reduce the annual energy use but could in fact reduce resilience if that cooling is dependent on water which would be unavailable during a long term emergency. The right unit of analysis is therefore the service pathway from the external grid through to the computing output, taking into account the thermal and control dependencies.

Table 1 sets out the core resilience functions of the major resources. Table 2 then takes that synthesis and translates it into Saudi design priorities. The key cross-cutting principle is protected flexibility.

Resources can be used in economic optimisation and as grid services only to the extent that they do so within the constraints of the required contingency reserve, fault isolation and service level obligations. That principle resolves the apparent paradox of data centres being both extremely inflexible critical loads and at the same time being a very valuable source of flexibility. They can be both, as long as the flexibility is explicitly subject to the resilience requirements. Coordination is the design metric.

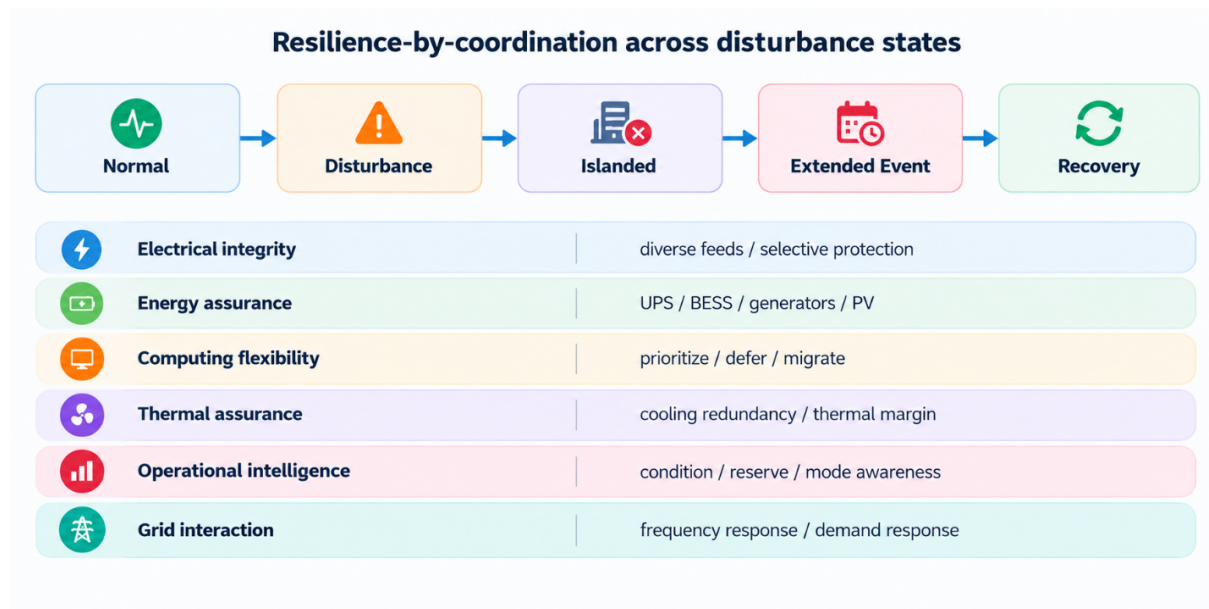


Figure 2. Resilience-by-coordination framework across operating and disturbance states.

Table 1. Resilience functions and limitations of principal data-center energy resources.

Resource / control	Primary resilience function	Key limitation to manage
Diverse utility feeds	Reduces common supply-point exposure	Shared routes or substations can create hidden common mode
UPS and batteries	Sub-second ride-through, power quality, reserve	Degradation, state of charge, BMS and thermal risk
Dispatchable generators	Long-duration firm supply	Start reliability, fuel logistics, maintenance state
Photovoltaics	Fuel displacement and daytime island support	Variability; not firm without storage or dispatchable backup
Workload flexibility	Reduces critical demand and enables grid services	QoS, latency, migration and application constraints
Cooling flexibility	Extends thermal ride-through and reduces peak demand	Ambient heat, water dependence, equipment limits
Predictive analytics	Reveals latent failures and available reserve	Sensor quality, model drift, cyber/communications dependency

Table 2. Saudi-specific resilience priorities derived from the reviewed evidence.

Saudi design priority	Recommended resilience response
Extreme ambient temperature	Size island-mode cooling explicitly; include thermal margin in autonomy calculations

Saudi design priority	Recommended resilience response
Water scarcity	Evaluate dry, hybrid, and evaporative cooling with both WUE and emergency water availability
High solar resource	Integrate PV with protected storage reserve and dispatchable backup rather than treating PV as firm capacity
Growing inverter-based generation	Use fast storage and bounded demand flexibility for frequency support where technically permitted
Rapid digital-load growth	Plan utility capacity, onsite resources, and workload flexibility together under uncertain expansion
Dust and harsh environment	Strengthen filtration, equipment derating, inspection, and condition-based maintenance
Critical national services	Use explicit recovery-time, loss-of-load, and common-mode-failure criteria beyond annual PUE

11 Research gaps and future directions

There are several gaps which prevent the direct application of the global literature to Saudi mission critical data centres. Firstly there is very little public reliability data on failure rates, outage durations, generator performance, battery degradation and cooling incidents in Saudi facilities. The probabilistic models that do exist are therefore reliant on international data which may not reflect the actual local conditions of temperature, dust, maintenance and utility. An anonymised national reliability dataset would be of immense value in enabling an evidence based design.

Secondly, electrical and thermal contingency models are very rarely coupled. Most of the power studies will have a fixed or very simplistic load for the cooling demand, and the cooling studies in turn don't normally model islanded generation, battery reserve or protection constraints. Future work should be to have a time resolved resilience model which incorporates heat accumulation, workload response, chiller staging, battery SoC and generator dispatch. This coupling is particularly important in the context of the Saudi summers.

Thirdly, long duration energy resilience is still very much in its infancy. Batteries are good for the short and medium term, but you're looking at a very different situation if you're talking about multi-day outages. You've got to think about fuel, you've got to think about the renewable forecasts, you've got to think about storage duration and you've got to think about how much workload you can relocate. There should be some comparative studies looking at combinations of batteries, hydrogen, thermal storage, renewable overbuild and dispatchable generation under a Saudi resource profile.

Fourthly, the cyber-physical interdependencies require greater attention. The microgrid controllers, the battery management systems, the workload schedulers and the building controls all add to the level of coordination but also create a raft of common software and communications dependencies. Resilience studies need to look at what safe fallback behaviour is available in the event of optimisation failure or communications failure. Grid interactive data centres also need some validated rules around how much flexibility they can reserve. Future work should look at just how much of the UPS, battery and workload capacity can be made available to the Saudi grid without increasing the internal loss of load risk (particularly under high levels of renewables and under extreme temperature conditions). The work should also look at the various resilience investments in

terms of economic metrics (avoided downtime costs etc) so that the various forms of redundancy, storage, cooling and digital controls can be properly compared.

12 Practical, managerial, and policy implications

For the designer the key point is that resilience requirements need to be translated into operating states and dependency maps prior to any equipment capacity being selected. The single line diagram needs to be checked against common mode failures, maintenance outages and failed transfers. Storage sizing should be on both the power and energy basis and cooling demand must be included in the islanding calculations. The contribution of renewables to fuel displacement and resilience should be captured, but they should not be credited as firm capacity unless there is an explicit adequacy model.

For the operator the reserve visibility is absolutely critical. The dashboards need to tell you not just the PUE or the energy cost but the available UPS runtime, the battery health, the generator readiness, the cooling margin and the current state of redundancy. Market participation and demand response should of course respect the protected reserve floors. Preventive and predictive maintenance should be scheduled taking the temporary loss of redundancy as a risk factor.

For Saudi policymakers and grid operators, data centres that are mission critical could become high quality flexibility resources if the connection standards are set up to recognise their fast controls whilst still allowing for the continuity obligations to be met. The incentives for storage, resilient microgrids and demand response should be based on verified capability rather than just installed capacity. Standardised definitions of outage severity, ride-through, recovery time and renewable backed islanding would enable benchmarking across the growing data centre sector. Procurement should reflect this.

13 Limitations

This review is constrained by the paucity of published Saudi data centre operational data and by the fact that the range of reliability metrics used in the literature is extremely diverse. A large number of the studies rely on simulation rather than long term field data and results from other climates or electricity markets cannot simply be transferred without modification. The 2020-2025 timeframe is of course biased towards more recent evidence and excludes the earlier, foundational work. The review is therefore intended to be an engineering review and a research framework rather than a substitute for site specific fault studies, protection coordination studies, thermal modelling or probabilistic reliability assessment. The body of evidence reviewed is sufficient to identify the recurring engineering mechanisms and to establish testable priorities for future Saudi field studies, pilot projects and for the development of standardised resilience reporting.

14 Conclusion

Mission critical data centres in Saudi Arabia will require a much broader concept of resilience than the conventional approach to backup power. The literature review makes it clear that resilience is the result of a coordinated approach to electrical topology, storage, dispatchable generation, renewables, cooling, computing flexibility, maintenance intelligence and interaction with the wider power system. Dual supply paths and UPS are still the bedrock of any design but their value is contingent upon fault selectivity, equipment condition, control logic and the success of the transition between modes.

Batteries are of course a very significant asset in terms of both ride through and grid support, but they must be treated as resilience assets in terms of their state of charge and health rather than simply as a financial asset.

Saudi Arabia has very high solar potential which makes the case for PV integration very strong. The hot arid climate and the water constraints make thermal design a primary resilience issue. The evidence from the Saudi renewables systems is that lowest energy cost does not necessarily equate to acceptable reliability. Mission critical facilities therefore need to have explicit loss of load and recovery criteria, resilience reserve floors and climate specific assumptions. Hybrid architectures (storage, renewables, dispatchable generation and flexible workloads) are far more defensible than any single technology.

The key contribution of the review is the resilience-by-coordination framework which links six domains: electrical integrity, energy assurance, computing flexibility, thermal assurance, operational intelligence and grid interaction. The data centre is viewed as both a protected critical load and a potentially valuable flexibility resource. The two roles are compatible so long as the grid services and optimisation are constrained by protected emergency reserves and service level constraints.

Future research in Saudi Arabia should be focused on field reliability datasets, coupled electrical-thermal contingency modelling, long duration storage, cyber-physical fallback behaviour and verified participation in frequency and demand response. As the Kingdom's digital infrastructure grows, resilient data centres can provide more than just uninterrupted computing; they can be a valuable asset to a more flexible and resilient national power system. There is a clear system benefit to this. If data centres can maintain auditable emergency reserves then their flexible capacity can be used to provide frequency control, demand response and to support renewable integration during normal operation. That is a very significant opportunity in the Saudi context where digital load is growing and renewables are expanding at the same time. The future facilities should therefore be planned as resilient cyber-physical energy assets rather than just as buildings with generators.

References

- [1] Parise G, Allegri M, Parise L. Electrical integrity resilience of data centers and critical loads. *IEEE Transactions on Industry Applications*. 2020;56(4):3397–3402. <https://doi.org/10.1109/TIA.2020.2986288>
- [2] Al Kez D, Foley AM, Mueen SM, Morrow DJ. Manipulation of static and dynamic data center power responses to support grid operations. *IEEE Access*. 2020;8:182078–182091. <https://doi.org/10.1109/ACCESS.2020.3028548>
- [3] Fu Y, Han X, Baker K, Zuo W. Assessments of data centers for provision of frequency regulation. *Applied Energy*. 2020;277:115621. <https://doi.org/10.1016/j.apenergy.2020.115621>
- [4] Zhang Y, Wilson DC, Paschalidis IC, Coskun AK. HPC data center participation in demand response: An adaptive policy with QoS assurance. *IEEE Transactions on Sustainable Computing*. 2022;7(1):157–171. <https://doi.org/10.1109/TSUSC.2021.3077254>
- [5] Wang D, Xie C, Wu R, Lai CS, Li X, Zhao Z, et al. Optimal energy scheduling for data center with energy nets including CCHP and demand response. *IEEE Access*. 2021;9:6137–6151. <https://doi.org/10.1109/ACCESS.2020.3049066>
- [6] Tian M, Zhang H, Wu H. Energy management for data centre microgrids considering co-optimisation of workloads and waste heat. *IET Energy Systems Integration*. 2022;4(1):43–53. <https://doi.org/10.1049/esi2.12044>
- [7] Zhu Y, Wang J, Bi K, Sun Q, Zong Y, Zong C. Energy optimal dispatch of the data center microgrid based on stochastic model predictive control. *Frontiers in Energy Research*. 2022;10:863292. <https://doi.org/10.3389/fenrg.2022.863292>

-
- [8] Zhang S, Gao J, Qian H, Yu J, Luo S, Guo Q. Power system operational reliability assessment based on the data center energy consumption elastic space. *Frontiers in Energy Research*. 2024;12:1370861. <https://doi.org/10.3389/fenrg.2024.1370861>
- [9] Wang Z, Wang Y, Ji H, Hasanien HM, Zhao J, Yu L, et al. Distributionally robust planning for data center park considering operational economy and reliability. *Energy*. 2024;290:130185. <https://doi.org/10.1016/j.energy.2023.130185>
- [10] Chen H, Gao P, Liu K, Chen L, Huang W. Joint expansion planning for data centers and distribution networks based on conditional value-at-risk theory considering low carbon characteristics. *Electric Power Systems Research*. 2024;229:110162. <https://doi.org/10.1016/j.epsr.2024.110162>
- [11] Zhou S, Zhou M, Wu Z, Wang Y, Li G. Energy-aware coordinated operation strategy of geographically distributed data centers. *International Journal of Electrical Power & Energy Systems*. 2024;159:110032. <https://doi.org/10.1016/j.ijepes.2024.110032>
- [12] Chen X, Zhang M, Jiang S, Gou H, Zhou P, Yang R, et al. Energy reliability enhancement of a data center/wind hybrid DC network using superconducting magnetic energy storage. *Energy*. 2023;263:125622. <https://doi.org/10.1016/j.energy.2022.125622>
- [13] Li X, Pang A, Yang W, Zhao Q. VRLA battery fault prediction for data center based on random forest model and feature enhancement method. *Journal of Energy Storage*. 2023;72:108666. <https://doi.org/10.1016/j.est.2023.108666>
- [14] Karimi L, Yacuel L, Degraft-Johnson J, Ashby J, Green M, Renner M, et al. Water-energy tradeoffs in data centers: A case study in hot-arid climates. *Resources, Conservation and Recycling*. 2022;181:106194. <https://doi.org/10.1016/j.resconrec.2022.106194>
- [15] Silva-Llanca L, Ponce C, Bermudez E, Martinez D, Diaz AJ, Aguirre F. Improving energy and water consumption of a data center via air free-cooling economization: The effect weather on its performance. *Energy Conversion and Management*. 2023;292:117344. <https://doi.org/10.1016/j.enconman.2023.117344>
- [16] Mi R, Bai X, Xu X, Ren F. Energy performance evaluation in a data center with water-side free cooling. *Energy and Buildings*. 2023;295:113278. <https://doi.org/10.1016/j.enbuild.2023.113278>
- [17] Cho J, Lim S. Balanced comparative assessment of thermal performance and energy efficiency for three cooling solutions in data centers. *Energy*. 2023;285:129370. <https://doi.org/10.1016/j.energy.2023.129370>
- [18] Zhou J, Kanbur BB, Le DV, Tan R, Duan F. Multi-criteria assessments of increasing supply air temperature in tropical data center. *Energy*. 2023;271:127043. <https://doi.org/10.1016/j.energy.2023.127043>
- [19] Isler-Kaya A, Karaosmanoglu F. Life cycle assessment of a climate-friendly data center cooling device. *Energy and Buildings*. 2023;288:113006. <https://doi.org/10.1016/j.enbuild.2023.113006>
- [20] Wang Y, Rousis AO, Strbac G. On microgrids and resilience: A comprehensive review on modeling and operational strategies. *Renewable and Sustainable Energy Reviews*. 2020;134:110313. <https://doi.org/10.1016/j.rser.2020.110313>
- [21] Kumar S, Saket RK, Dheer DK, Holm-Nielsen JB, Sanjeevikumar P. Reliability enhancement of electrical power system including impacts of renewable energy sources: A comprehensive review. *IET Generation, Transmission & Distribution*. 2020;14(10):1799–1815. <https://doi.org/10.1049/iet-gtd.2019.1402>
- [22] Jiang J, Peyghami S, Coates C, Blaabjerg F. A comprehensive study on reliability performance of photovoltaic-battery-based microgrids under different energy management strategies. *Journal of Energy Storage*. 2021;43:103051. <https://doi.org/10.1016/j.est.2021.103051>
- [23] Shair J, Li H, Hu J, Xie X. Power system stability issues, classifications and research prospects in the context of high-penetration of renewables and power electronics. *Renewable and Sustainable Energy Reviews*. 2021;145:111111. <https://doi.org/10.1016/j.rser.2021.111111>
- [24] Obi M, Slay T, Bass RB. Distributed energy resource aggregation using customer-owned equipment: A review of literature and standards. *Energy Reports*. 2020;6:2358–2369. <https://doi.org/10.1016/j.egyr.2020.08.035>
- [25] Marqusee J, Becker W, Ericson S. Resilience and economics of microgrids with PV, battery storage, and networked diesel generators. *Advances in Applied Energy*. 2021;3:100049. <https://doi.org/10.1016/j.adapen.2021.100049>
- [26] Al-Hanoot AK, Mokhlis H, Mekhilef S, Alghoul M, Shareef H, Samatar AM. Distributed PV systems in Saudi Arabia: Current status, challenges, and prospects. *Energy Strategy Reviews*. 2024;55:101535. <https://doi.org/10.1016/j.esr.2024.101535>
-

-
- [27] Eltamaly AM. A novel energy storage and demand side management for entire green smart grid system for NEOM City in Saudi Arabia. *Energy Storage*. 2024;6(1):e515. <https://doi.org/10.1002/est2.515>
- [28] Alghamdi AS. Optimizing energy costs and reliability: A multi-objective framework with learning-enhanced manta ray foraging for hybrid PV/battery systems. *Energy*. 2024;291:130346. <https://doi.org/10.1016/j.energy.2024.130346>
- [29] Al Dawsari S, Anayi F, Packianather M. Techno-economic analysis of hybrid renewable energy systems for cost reduction and reliability improvement using dwarf mongoose optimization algorithm. *Energy*. 2024;313:133653. <https://doi.org/10.1016/j.energy.2024.133653>
- [30] Kahouli O. Frequency stability analysis of power grid during photovoltaic generation in Hail region (Arabia Saudi). *Frontiers in Energy Research*. 2025;12:1323154. <https://doi.org/10.3389/fenrg.2024.1323154>