Trade-off Assessment of Load-Following Performance and Operational Burden with Reactor Unit Granularity for a 100 MWe AI Data Center

Tae Jun Park, Jeong Ik Lee*
Dept. Nuclear & Quantum Eng., KAIST, 373-1, Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea
**Corresponding author: jeongiklee@kaist.ac.kr*

***Keywords: AI Data Center, Small Modular Reactor, Microreactor, Load Following, Unit Granularity.**

1. Introduction

AI data centers require large and highly reliable power supplies. Google and Amazon are pursuing multiple advanced-reactor and Small Modular Reactor (SMR) projects to support data-center electricity demand [1,2]. Reactor capacity selection must consider not only rated power but also the minimum power, ramp rate, low-power operating time, and repeated power maneuvers imposed by the actual AI data-center load.

Long-term facility-level power data remain limited. Choukse et al. reported rapid power variations during AI training and a 5–15 min utility scheduling interval [3], while Vercellino et al. combined H100 workload measurements with a bottom-up facility model to release one-year, one-minute whole-facility profiles for 20–80% utilization [4,5]. Previous studies have examined nuclear-powered data-center availability [6], flexible dispatch of multi-unit SMRs [7], non-synchronous unit-specific load-following control [8], LF (load-following)-SMR scheduling [9], microreactor requirements for high-performance computing (HPC) [10], and dynamic SMR–data-center matching [11]. However, the effect of unit granularity and post-downpower recovery on the trade-off between direct-load service and unit-level dynamic burden has not been quantified for the same 100 MWe total capacity across N=1–10. This study quantifies that trade-off using a generic architecture-screening framework.

2. Methodology

2.1. AI data-center load and analysis basis

The reference demand is a 100 MWe AI data center. This point is selected because large-load interconnection requests are expanding into the tens-to-hundreds-of-MW range [12] and because it permits a direct comparison from 1×100 MWe to 10×10 MWe at fixed total capacity. The long-term loads are the 20%, 40%, 60%, and 80% utilization time series from Vercellino et al. [4,5]. The annual maximum of each series is mapped to 100 MWe so that its temporal shape and scenario-specific variability are retained.

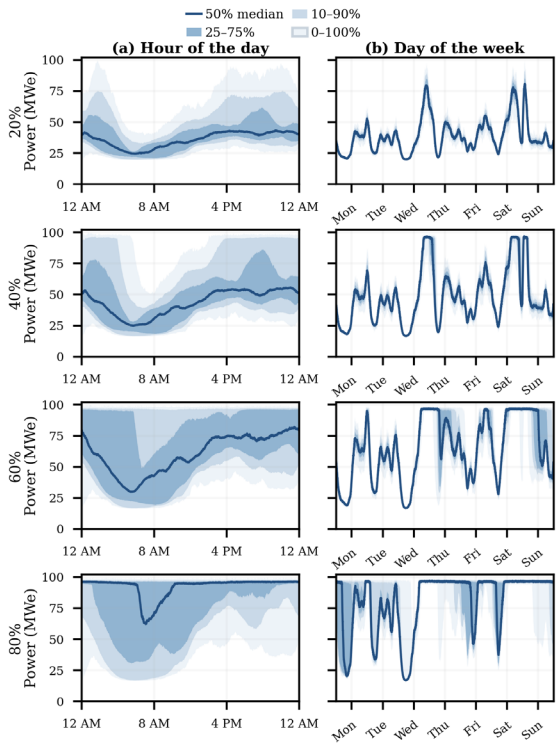


Fig. 1. Time-of-day power distributions from the one-year, one-minute colocation time series of Vercellino et al., scaled to a 100 MWe peak demand.

Figure 1 applies the percentile representation used by Vercellino et al.; the calculations use the full one-year, one-minute time series rather than the percentile curves. Reactor-level dispatch is updated using nonoverlapping 5 min averages. This interval is not assumed to be a universal reactor response time; it uses the lower bound of the scheduling interval reported by Choukse et al. to retain short-term variability [3]. The facility simulation by Vercellino et al. excludes cooling, power-distribution losses, and other non-IT auxiliary loads [4]. The profiles are used as load-shape scaling for architecture comparison rather than as actual hyperscale site-meter demand.

2.2. Reactor architecture and staged dispatch

Total nuclear capacity is fixed at 100 MWe and divided into N=1–10 units. Rated unit power is P_{unit}=100/N MWe, and N=10 represents approximately the 10 MWe scale of the Holos-Quad microreactor [13]. In the non-synchronous staged dispatch, as many units as

possible remain at rated power while one regulating unit supplies the residual demand. This unit-specific dispatch follows the motivation of non-synchronous multi-unit operation, which avoids forcing all units to follow the same normalized trajectory [8]. Minimum stable power is set to $0.2P_{unit}$ as a permissive screening lower bound based on the 20–40% Rated Thermal Power (RTP) extended low-power range discussed by the IAEA [14].

A unit that has reduced power is not allowed to increase output during the effective recovery period T_{rec}. T_{rec} is not a universal reactor-specific constant; it is a screening parameter representing the delay before a unit can resume full regulating duty, and 0–9 h is examined. The range is informed by the hour-scale stable-power restriction used in the LF-SMR study by Yang et al. [9]. Because ramp capability varies across reactor technologies, no common ramp constraint is imposed; the required ramp is instead calculated as an architecture output.

2.3. Flexibility metrics

The IAEA identifies minimum power, power-change rate, reduced-power duration, and load-cycle occurrence/depth as key dimensions of flexible operation [14]. The OECD/NEA likewise discusses ramping and daily load cycling as core characteristics [15]. These sources do not prescribe equations for the present AI time-series problem, so this study converts those dimensions into four quantitative metrics.

The first metric, direct-load coverage, measures how closely nuclear output directly follows data-center demand. For scenario u, the annual mismatch between demand $D_u(k)$ and nuclear supply $P_u(k)$ is normalized by total demand.

$$C_u = 100\left[1 - \frac{\Sigma_k |D_u(k) - P_u(k)|}{\Sigma_k D_u(k)}\right] \quad (1)$$

The minimum across the four scenarios is used as the robust coverage. It is an energy-based matching metric, not availability or loss-of-load probability.

The second metric, required unit ramp, corresponds to power-change rate. For unit i, the output change over consecutive 5 min dispatch intervals is normalized by rated unit power.

$$r_i(k) = 100\frac{|P_i(k) - P_i(k-1)|}{P_{unit}\Delta t} \quad (2)$$

To reduce sensitivity to a single extreme event, the P99 of nonzero ramp events is used, and the maximum across the four scenarios is defined as the robust requirement.

The third metric, low-power residence, is the fraction of online unit-hours with $0<P<0.4P_{unit}$. The 40% value is a screening threshold based on the IAEA discussion of the 0–40% RTP range as a more challenging region for flexible operation [14]; it is not a universal operating limit.

$$F_u = 100\frac{H_{low,u}}{H_{on,u}} \quad (3)$$

The fourth metric, maneuvering index, measures cumulative repeated output variation. A 100→0→100% sequence is counted as one equivalent cycle. It is not a fatigue usage factor, but a screening indicator for comparing repeated power-control burden across architectures.

$$M_u = \frac{1}{2N}\Sigma_{i,k} |p_i(k) - p_i(k-1)| \quad (4)$$

2.4. Pareto-based architecture selection

The primary benefit of unit subdivision is direct-load coverage, and the primary burden is required unit ramp. At each T_{rec}, solutions with lower coverage and equal or higher ramp are removed to form the nondominated Pareto front. The two objectives are normalized to 0–1, and the N with the maximum perpendicular distance from the line connecting the two end points is selected as a knee reference without assigning preference weights [16]. Low-power residence and maneuvering are supporting metrics used to assess operability around the knee. For T_{rec}≥5 h, the 5 min full-unit-transition ceiling (=20% P_{unit}/min) affects the Pareto geometry, so the corresponding knees are treated as screening results only.

3. Results and Discussion

3.1. Direct-load coverage

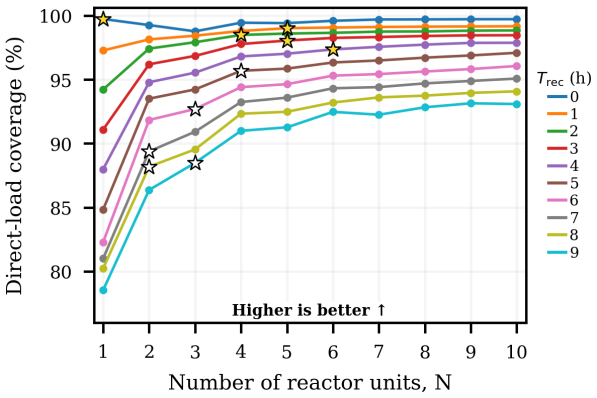


Fig. 2. Scenario-robust direct-load coverage. Stars indicate coverage–ramp Pareto knees.

In Figure 2, coverage at T_{rec}=0 h is 99.73% for N=1, and unit subdivision provides almost no additional benefit. At T_{rec}=3 h, coverage increases from 91.09% to 98.46% as N increases from 1 to 10; at 4 h, it increases from 87.98% to 97.87%. The Pareto knees for T_{rec}=1–4 h occur at N=5, 4, 5, and 6, respectively, while the additional coverage from each knee to N=10 is only 0.15–0.52 percentage points. Thus, a multi-unit architecture recovers direct-load service under a recovery restriction, but the benefit shows diminishing returns beyond an intermediate N.

3.2. Required unit ramp

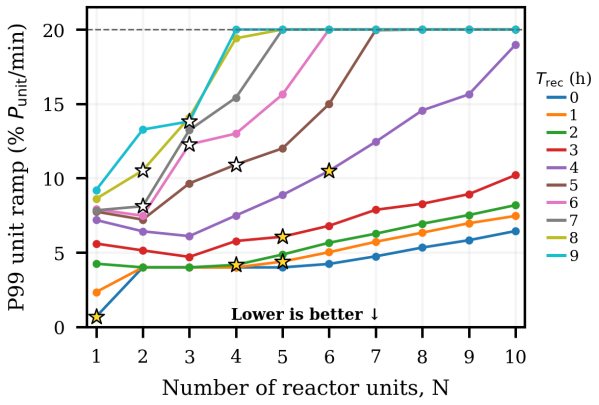


Fig. 3. Scenario-robust P99 unit-ramp requirement.

Figure 3 shows that the required unit ramp generally increases with N because the same MW-scale adjustment is concentrated into units with smaller rated power, which increases the normalized ramp. Increasing T_{rec} further concentrates upward duty on units that have completed recovery. At the T_{rec}=1–4 h knees, the P99 requirement is 4.17–10.49% P_{unit}/min, rising to 7.46–18.98% P_{unit}/min at N=10. The 20% P_{unit}/min plateau in some T_{rec}=5–9 h curves is not convergence of a physical reactor ramp rate. It is the full-unit-transition ceiling corresponding to a 0↔100% P_{unit} changes within one 5 min interval (100/5=20%/min). As recovery lengthens, fewer units remain available for upward response, so near-full-unit transitions enter the upper 1% of nonzero events at smaller N, shifting the plateau to the left.

3.3. Low-power residence

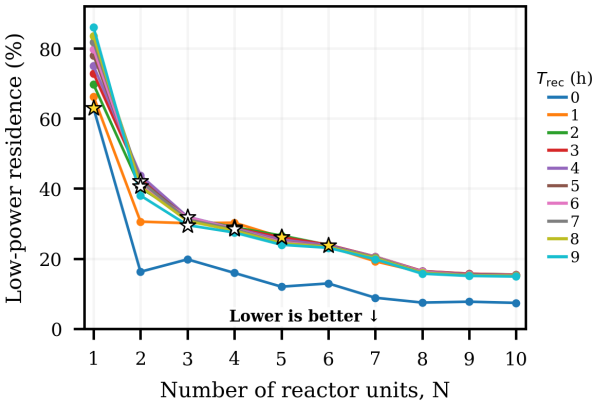


Fig. 4. Scenario-robust low-power residence below 40% of unit rated power.

The low-power residence in Figure 4 decreases with unit subdivision because a larger fleet can allocate online units near rated power or place units offline. At T_{rec}=3 h, it falls from 72.86% to 15.39% as N increases from 1 to 10. Reactor technologies for which prolonged low-power operation is particularly undesirable may favor an N above the Pareto knee.

3.4. Maneuvering burden

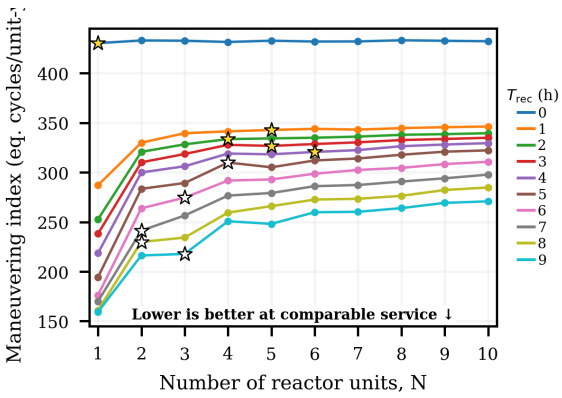


Fig. 5. Scenario-robust maneuvering index.

In Figure 5, the T_{rec}=0 h curve lies above the other recovery cases because, without a recovery restriction, the architecture executes more short-term reversals in the AI data-center load. For N=1, the maneuvering index decreases from about 430 to 287, 238, and 159 cycles/unit-y as T_{rec} increases from 0 to 1, 3, and 9 h, while coverage simultaneously falls from 99.73% to 97.27%, 91.09%, and 78.53%. A lower maneuvering index is not inherently better and should be compared among architectures that provide similar service levels.

3.5. Pareto-based unit-count selection

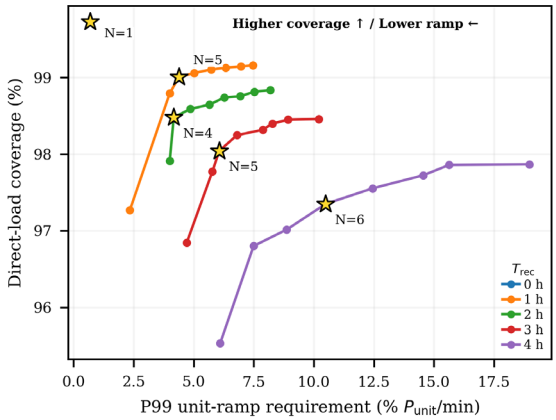


Fig. 6. Coverage–P99 ramp Pareto fronts for T_{rec}=0–4 h.

Figure 6 shows the Pareto fronts between direct-load coverage and required unit ramp for T_{rec}=0–4 h. At T_{rec}=0 h, N=1 simultaneously provides high coverage and the lowest ramp, so there is little practical trade-off. At T_{rec}=1–4 h, higher service requires greater unit ramp, and the knees occur at N=5, 4, 5, and 6, respectively. For example, at T_{rec}=3 h, increasing N from 5 to 10 raises coverage only from 98.04% to 98.46% (+0.42 percentage points), whereas the P99 ramp increases from 6.06% to 10.21% P_{unit}/min.

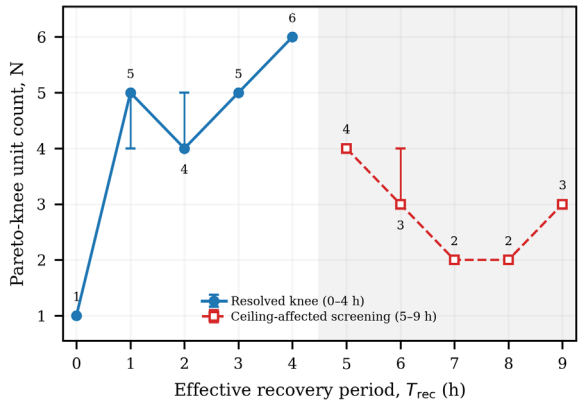


Fig. 7. Pareto-knee unit count versus recovery period.

Figure 7 summarizes the Pareto knee as a function of recovery period. The difference between the maximum-distance criterion and a normalized ideal-point criterion is shown by the error bars. Both methods retain an intermediate granularity of N=4–6 for T_{rec}=1–4 h. For T_{rec}≥5 h, the 5 min transition ceiling affects the Pareto geometry, so the shaded-region results are treated as ceiling-affected screening results rather than quantitative optima.

The knee is not a universal optimum. It marks an architecture at which the service benefit of further modularization begins to diminish sharply. If direct-load coverage is the overriding objective, subdivision beyond the knee remains possible, but the additional service gain is small. Reactor technologies that strongly avoid low-power operation may favor a larger N, whereas technologies with limited ramp or repeated maneuvering capability may favor the knee or a smaller N.

The analysis uses ideal staged dispatch and an effective T_{rec} and does not establish the feasibility of a specific reactor design. Intra-step PCS dynamics, reactor-specific xenon and thermal constraints, reserve and storage requirements, and the economics of increasing unit count are not included. The knees in Figs. 6 and 7 serve as architecture references for subsequent reactor/PCS dynamic assessment and for allocating the roles of storage and workload management.

4. Conclusions

This study maps the publicly available one-year, one-minute AI facility profiles of Vercellino et al. to a 100 MWe peak demand and compares N=1–10 over T_{rec}=0–9 h. With no recovery restriction, N=1 provides about 99.7% direct-load coverage and the lowest required ramp. For T_{rec}=1–4 h, multi-unit architectures recover coverage but increase unit-level ramp, and the Pareto knee lies at an intermediate granularity of N=4–6. Further subdivision from the knee to N=10 improves coverage by only 0.15–0.52 percentage points, while the P99 ramp increases by 3.1–8.5% P_{unit}/min.

From a design standpoint, additional subdivision beyond the knee remains reasonable when direct-load coverage is the overriding objective, although the service gain is limited. Reactor technologies that strongly avoid low-power operation may favor an N above the knee, whereas technologies with limited unit ramp or repeated maneuvering capability may favor the knee or a smaller N. For T_{rec}≥5 h, the P99 approaches the 5 min transition ceiling; unit subdivision alone is insufficient, and reactor recovery, PCS decoupling, storage, or workload management should be considered together. The results do not define a universal optimum; they provide an application-side architecture-screening map that can be compared with the actual dynamic capability of candidate SMR and microreactor systems.

Future work will couple reactor-specific xenon and thermal-hydraulic behavior with PCS dynamics to test the physical basis of T_{rec} and the required ramp. It will also examine coordinated operation with storage and workload management, cycle-depth-dependent fatigue, and the economics and maintainability associated with increasing unit count.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.: RS-2024-00436693).

REFERENCES

[1] M. Terrell, “New nuclear clean energy agreement with Kairos Power,” Google, Oct. 14, 2024.
[2] Amazon, “Amazon signs agreements for innovative nuclear energy projects to address growing energy demands,” Oct. 16, 2024.
[3] E. Choukse et al., “Power Stabilization for AI Training Datacenters,” arXiv:2508.14318, 2025.
[4] R. Vercellino, J. Willard, G. Campos, W. da Silva Pereira, O. Hull, M. Selensky, and J. Mueller, “Measurement of Generative AI Workload Power Profiles for Whole-Facility Data Center Infrastructure Planning,” arXiv:2604.07345, 2026.
[5] R. Vercellino, J. Willard, G. Campos, W. da Silva Pereira, O. Hull, M. Selensky, and J. Mueller, “Dataset of Generative AI Workload Power Profiles”, NLR Data Catalog, 2026.
[6] B. Hanna, A. Abou-Jaoude, N. Guaita, P. Talbot, and C. Lohse, “Navigating Economies of Scale and Multiples for Nuclear-Powered Data Centers and Other Applications with High Service Availability Needs,” Energies, vol. 17, 2024.
[7] S. Alhadhrami, G. J. Soto, and B. Lindley, “Dispatch analysis of flexible power operation with multi-unit small modular reactors,” Energy, vol. 280, 2023.
[8] H.-H. Byun and M.-S. Yim, “Genetic algorithm-based non-synchronous unit-specific optimal load-following control of multi-unit small modular reactor,” Energy, vol. 314, 2025.
[9] Y. Yang, D. Wang, J. Shi, C. Wu, and Z. Han, “Uncertainty-Aware Day-Ahead Datacenter Workload Planning with Load-Following Small Modular Reactors,” ACM SIGENERGY Energy Informatics Review, vol. 5, 2025.
[10] A. J. H. Lee, L. Wodrich, D. Kalinichenko, C. S. Brooks, and T. Kozlowski, “Modeling Microreactor Requirements for High-Performance Computing,” Nuclear Technology, vol. 210, 2024.

[11] Z. You, Y. Wu, J. Wang, Z. Wei, and J. Ling, "Dynamic equilibrium mechanism: Integrating small modular reactors and optimized cooling for sustainable data centers," Annals of Nuclear Energy, vol. 224, 2025.
[12] G. Krishnamoorthy, V. Westfallen, and K. McKenna, "The Role of the U.S. Electric Distribution System in Serving Data Center and Other Large Loads", NLR/TP-6A40-96783, 2026.
[13] A. Moisseytsev and C. Filippone, "Load Following Analysis of the Holos-Quad 10MWe Micro-Reactor with Plant Dynamics Code", ANL/NSE-21/32, Argonne National Laboratory, 2022.
[14] IAEA, Non-baseload Operation in Nuclear Power Plants: Load Following and Frequency Control Modes of Flexible Operation, IAEA Nuclear Energy Series No. NP-T-3.23, 2018.
[15] OECD Nuclear Energy Agency, Technical and Economic Aspects of Load Following with Nuclear Power Plants, 2011.
[16] G. Yu, Y. Jin, and M. Olhofer, "Benchmark Problems and Performance Indicators for Search of Knee Points in Multiobjective Optimization," IEEE Transactions on Cybernetics, vol. 50, 2020.