Title

Oversizing and Part-Load Problems

Authors

Meier, Alan

Merrill, Wyatt

Publication Date

2026-08-03

Oversizing and Part-Load Problems

Alan Meier, Lawrence Berkeley National Laboratory
Wyatt Merrill, U.S. Department of Energy

ABSTRACT

Oversizing, the common engineering practice of specifying devices with capacity exceeding the actual load requirement, is a widespread practice across virtually all building technologies end-use categories, including HVAC, electrical systems, lighting, appliances, and plug loads. This practice, driven by factors like design uncertainty, institutional pressures, and risk aversion, results in wasted capital investment, control difficulties, and excessive energy consumption due to inefficient part-load operation.

Part-load operation, where devices run below maximum capacity, is the dominant operating mode in most energy systems and presents a complex design challenge. Solutions to match output to load fall into three broad categories: constraining the output, adjusting the device's internal behavior, and linking output to energy storage or other waste-heat reuse applications. The energy implications of part-load are critical, as efficiency often drops sharply as load decreases across a wide diversity of devices.

To quantify the extent of this problem, we derive a dimensionless Part-Load Metric (PLM) based on device efficiency and its frequency distribution of operating hours at various output levels. The PLM quantifies the deviation of a device's actual efficiency from its maximum design efficiency. This metric also serves as a measure of "capital inefficiency," enabling engineers to compare the impact of different part-load solutions and providing a unified framework for evaluating performance across various devices and systems.

Introduction: Converting Energy to Services and Sizing

Devices transform energy into services sought by a user. Motors transform electrical energy into mechanical motion, lights transform electricity into illumination, and boilers transform chemical energy into heat. These transformations—and the devices undertaking them—are extraordinarily diverse in application, size, and complexity. Designing devices to reliably and economically accomplish these transformations even as services sought differ over time is a fundamental engineering challenge.

Despite the diversity of transformations, these devices share several common features (Figure 1). Each device has an *input*, that is, a flow of energy used for the transformation. The input will typically be electricity or some sort of fuel. Each device has an *output*, that is, the product of the device's transformation. Each device has a *capacity*, that is, its maximum output, typically defined as the peak output that a device can deliver in specified conditions. This output is sometimes called the "design" or "nominal" capacity. The output is used to meet some sort of need for it, that is, the *load*. The load corresponds to practical needs, such as to make up for heat loss in a building or to the power required by the pump to lift water. Ultimately, the output provides a *service*, such as thermal comfort or mobility. While Figure 1 depicts the device as a simple "box", it can be as complex as an electric grid.

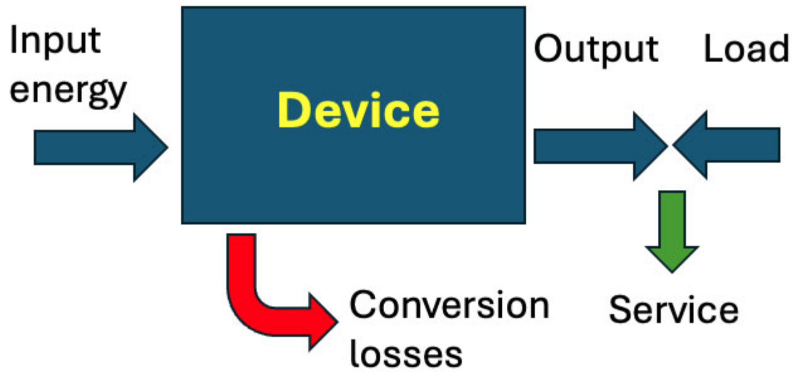


Figure 1. Characteristics of a typical device. The output of the device meets a load and enables the provision of a service. For example, electricity is supplied to an air conditioning unit (input energy), some amount of that input is lost through inefficiencies in the compressor, fan motor, electrical resistance, etc. (conversion losses), some output is provided in the form of thermal energy, and whether that output is appropriately sized depends on the building's demand at that time (load). Ultimately, a service is rendered (lowering temperature and humidity), which is less than the input energy due to conversion losses and may or may not be oversized for the real-time load.

Engineers design devices to have sufficient capacity for whatever *load* can be reasonably expected for the *service* it provides. Each application has its own guidelines for sizing to assure adequate output for all—or nearly all—anticipated loads. A heating system, for example, is sized to meet the building's heating requirements for the coldest day. A pump is sized to meet maximum anticipated flow. In an electrical panel or building transformer, the total capacity (typically expressed in terms of amps or watts) is sized to meet an expectation of the peak demand of the building(s) they service. A device with insufficient capacity, that is an *undersized* device, provides inadequate services, leading to reduced productivity, discomfort, and other kinds of costs (including litigation). Thus, a more common engineering outcome is oversizing.[1] In addition to the higher upfront capital investment, some of which may never be utilized, oversizing has three visible drawbacks operationally: it typically causes excessive energy use, is more difficult to control, and often requires greater maintenance. Nevertheless, oversizing is a widespread phenomenon across nearly all building technologies.

Oversizing occurs for understandable reasons, such as uncertainty in design conditions or to accommodate anticipated growth in load. Oversizing can also be caused by institutional pressures, such as allowing each component engineer to independently err on the side of caution, leading to an accumulated oversizing at the system level. Incomplete information can also play a role: oversizing almost always occurs when engineers and contractors rely on "rules of thumb" rather than recommended calculations. To our knowledge, no comprehensive review has been undertaken of oversizing across all applications and sectors, but reviews of specific applications have been conducted. Oversizing in building HVAC systems has been widely investigated. One

[1] Also called "over-provisioning" in data centers and certain other applications.

survey found that 40% of HVAC units in commercial buildings are oversized by at least 25%, with 10% being greatly oversized (more than 50%) (Djunaedy et al. 2011). Another detailed study of HVAC systems in seven commercial buildings found oversizing ranging from 14 - 247% for cooling and 165 - 630% for heating (Woradechjumroen et al. 2014). The luminous efficacy of lamps decays sharply under dimming conditions (NGLIA 2019). Studies of electrical panels suggest that they typically draw a small fraction of their rated capacities (Less and Walker 2026; Murphy et al. 2026). A survey of wastewater treatment facilities in Spain found the majority of large or medium-sized facilities were oversized 30 - 60% and smaller plants were even higher (Johnson et al. 2021). Oversizing is an acute problem in data centers, both with respect to the computing infrastructure and the services required to support them (Takci et al. 2025; Van Geet and Sickinger 2024). Engineers often restrict the sizing of the infrastructure to ensure a continuously perfect quality of service, which leads to an overestimation of the needed equipment (Benaissa et al. 2022). In summary, oversizing exists in virtually every type of application and sector and is especially relevant to buildings.

Part-Load Operation

When the service rendered by the device is constant, the procedures for sizing are straightforward. For example, a light fixture providing a constant level of illumination and single-speed fans provide a constant level of air movement. A more common situation is *part-load operation*,[2] that is, where a device runs at less than its maximum capacity (that same light fixture on a dimmer switch, or variable-speed fan). Part-load operation is the dominant operating mode in most energy systems and involves devices that are responsible for most of the world's energy consumption. Balancing the device's output with the need for the service is the essence of the part-load design challenge. Engineers must incorporate this flexibility within the constraints of acceptable costs, efficiency, and quality of service. It is a "problem" because failure to account for part-load operation leads to discomfort, manufacturing inconsistencies, increased maintenance, excess energy consumption, and wasted capital investment.

Strategies to Solve the Part-Load Problem

There are many creative solutions to match a device's output to the service being demanded by the end user, but there are few attempts to link these solutions across the whole spectrum of engineering applications. Despite the huge diversity of devices, part-load solutions fall into three broad categories:

- Constrain the output of the device
- Adjust the internal behavior of the device
- Link the device's output to energy storage or other applications

Constraining the device's output typically takes the form of brakes, throttles, dampers, resistors, and thermal dilution. Adjusting a device's internal behavior can be accomplished by cycling, staging, and variable-output operation. Excessive output (or insufficient output) can be

[2] The term "part-load" is widely used in the HVAC industry but has different names in other applications. For example, in wind turbines it is "Region II operation". For diesel and gas generators, the term "underloading" is sometimes used. The lighting industry uses "task tuning", "high-end trim", and "institutional tuning."

managed by diverting it to (or recalling it from) an energy storage device or utilizing the excess for ancillary applications (i.e. waste heat harvesting). Even though the solutions may appear different across devices, the strategies are fundamentally the same. For example, "thermal dilution" is used to modulate output in air conditioners (where it is called "terminal re-heat") and in hot water distribution (where it is called "tempering"). Figure 2 illustrates the commonality of these two approaches.

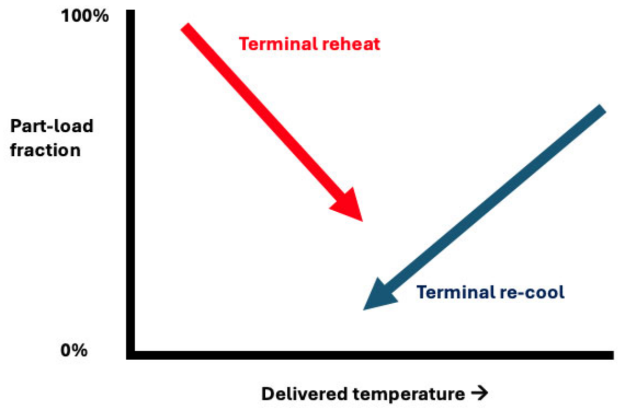


Figure 2. Impact of delivery temperature on part-load for terminal reheating and terminal re-cooling (tempering). A cooling system (e.g., an air conditioner) operating at peak efficiency (100%) will operate at lower part-load when the delivery temperature is raised by the heater. Similarly, a faucet operates at peak efficiency until cold water is added to lower the temperature and lower part-load efficiency.

Technological advances may enable new part-load solutions to be transferred from one application to another. This paper's contribution is partly the recognition of the cross-cutting nature of part-load problems.

Energy Implications of Part-Load Operation

The solutions described above demonstrate that a system can be designed to overcome the part-load problem in many ways. The goals of these solutions typically include reducing investment and improving quality and reliability of service. Saving energy is another important reason to improve part-load performance because higher part-load operation generally raises efficiency. Figure 3 shows the relationship between efficiencies and part-load for mechanical, electrical, and electronic devices. Some devices retain their efficiencies over a wide range of part-loads, while others drop almost proportionally with the level of output. Almost all devices exhibit a sharp drop in efficiency below about 20% part-load (and many are unable to reliably operate below that level).

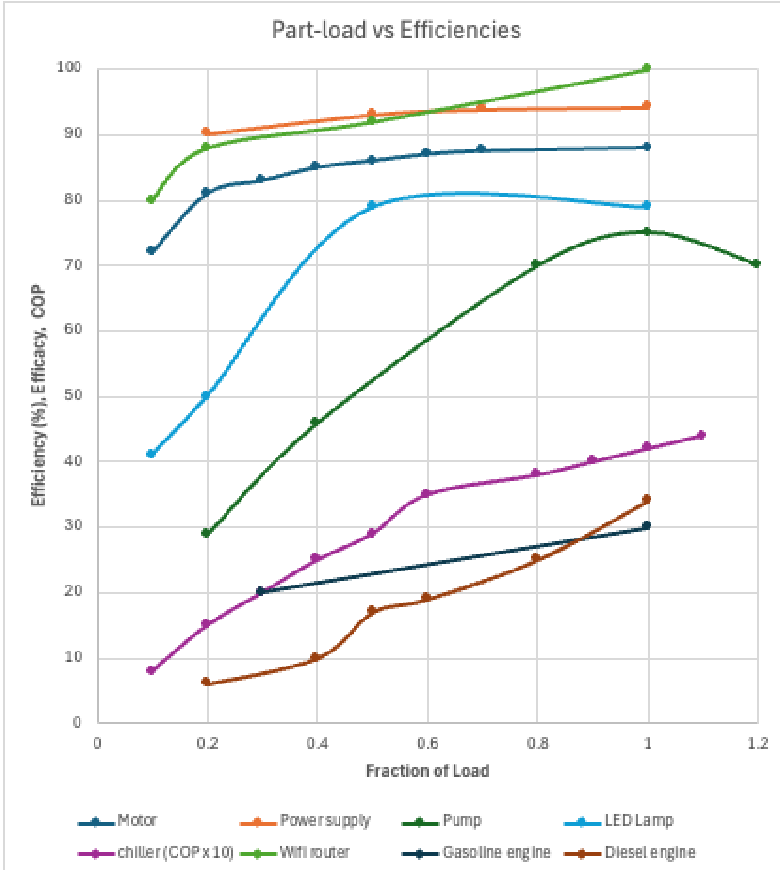


Figure 3. Relationships between part-load and efficiency for selected devices. This is a mixture of laboratory and field measurements. Some values have been normalized to facilitate comparison.

The energy losses are obvious for some part-load solutions—when a mechanical brake or an electrical resistor create intense heat whenever they constrain output—but other losses may be more subtle. For example, thermal dilution will result in both direct energy losses and a decrease of potential work (that is, exergy). Sometimes the energy used to apply the final control (e.g., terminal reheat or re-cool) may occur elsewhere, and the system efficiency calculation may omit it.

A generic representation of the part-load efficiency curve for a device is displayed in Figure 4. Benchmark measurements of efficiency for regulations and engineering calculations are typically performed at 100% of rated capacity and rarely below 50% because these test conditions capture normal operating conditions. However, these conditions become less relevant where the device is greatly oversized for the actual services it provides. For example, if a pump is oversized by 40%, then the device will experience a maximum load of about 71% of its capacity. When the load falls to 0.5 of peak, the part-load factor falls to about 36%. The oversizing phenomenon means that benchmark tests at lower part-loads become increasingly relevant.

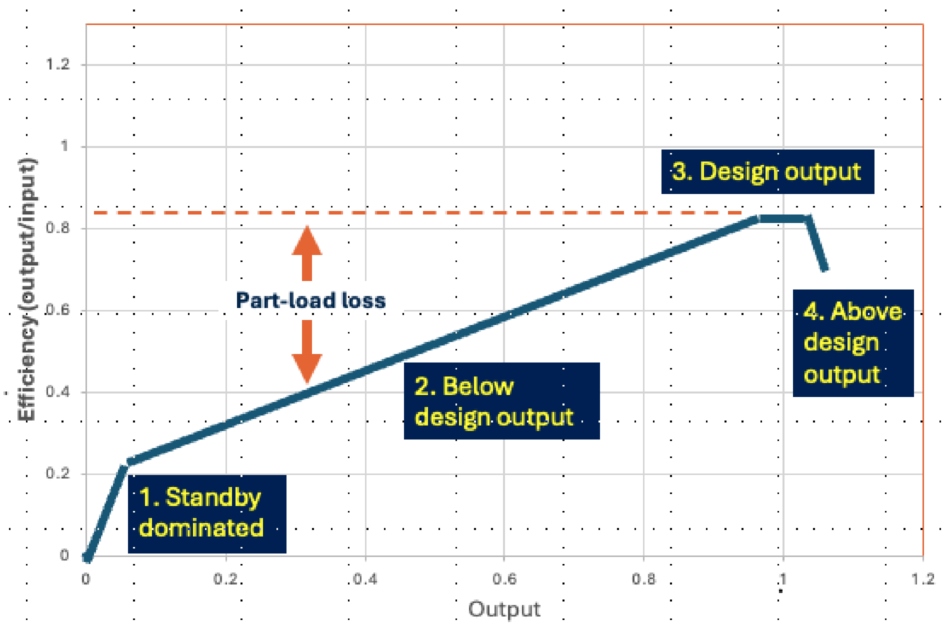


Figure 4. A simplified efficiency map showing the four key performance regions. The lengths and slopes of the efficiency segments depend on the system (and might not even exist. The arrow identifies the loss in efficiency arising from part-load operation at a specific output. The segments should not be confused with operating time in that region.

Most systems will have several part-load conditions and associated part-load efficiencies (or efficacies, depending on the nature of the output). The relationship between output and efficiency is shown in Figure 4. Four general regions of part-load behavior will appear:

1. Standby-dominated
2. Below design output
3. Design output
4. Above design output

In region 1, zero or very low output, standby power consumption dominates and, as a result, the efficiency is very low. In region 2, with rising output, device energy efficiency rises. In region 3 output efficiency peaks at design output. This is a kind of tautology because engineers try to maximize efficiency at design output. Finally in region 4, efficiency falls when the device is operated above design conditions. Overdriving equipment is a common feature (occurring in internal combustion engines, variable-speed motors and batteries) but at the cost of depressed efficiency and increased maintenance.

A unique mode exists at zero output (and therefore zero efficiency) which is typically called standby or idle mode. Devices perform various management and maintenance functions in the standby mode. Power supplies often remain energized, drawing some power even when there is no load. Millions of heat pumps are equipped with crankcase heaters to protect the compressors during the off cycle. Data centers have particularly high standby power consumption to maintain their redundant systems. Standby power consumption ranges from milliwatts to Megawatts and, since it occurs often for large fractions of a device's operating cycle, standby can represent a significant fraction of a device's total energy consumption. Here, too, the literature reports relatively information about the times devices reside in their standby modes. These unique characteristics and their crosscutting nature have prompted separate attention by policymakers and industry.

The difference in efficiencies (between design and part-load) is also shown in Figure 4. The vertical arrow represents the difference at a specific level of output (in this case, 0.33 of design output).

Derivation of a Part-Load Metric

In this section, we derive a metric quantifying the extent of part-load operation. It is based on both the part-load losses described above and the time the device operates at part-load. A metric can help expose wastes of both energy and capital. Some existing energy estimation procedures already estimate real-world energy consumption by approximating the distribution with a few points. For example, the Pump Energy Index, uses a pre-determined, time-weighted load profile to calculate a "true weighted efficiency" (Dahl 2018) that accurately predicts energy consumption and savings. However, this metric is limited to very narrow applications (e.g., large pumps in municipal facilities). Our goal is to develop a metric with broader applicability and more easily recognize "high" and "low" values of part-load usage across a wide range of devices and systems.

The metric is based on a distribution of the number of operating hours in each interval of output. This is graphically depicted in Figure 5. Detailed operating data (output, duration, loads, efficiency) are increasingly available as the costs of sensors, data collection and analysis fall. Increasingly, energy management systems collect these kinds of data with high resolution, sometimes with intervals below five minutes.

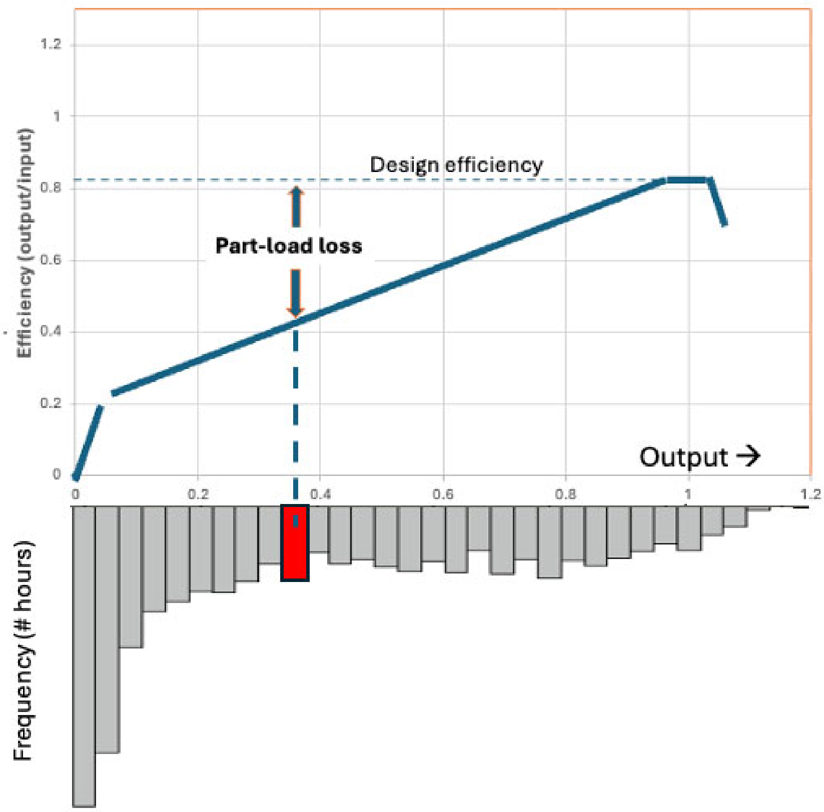


Figure 5. Superposition of efficiency and frequency curves to enable calculation of part load energy at each level of output.

The losses—or gap—attributable to part-load operation at a given level of output—or bin—is the difference in output between the design and actual conditions multiplied by the frequency (e.g., number of hours). These output losses need to be further divided by their respective efficiencies to obtain input energy losses. The input energy losses can be expressed as:

Input energy losses L_b attributed to part-load operation at output level b are given by

$$L_b = H_b \times S_b \times \left(\frac{1}{\eta_b} - \frac{1}{\eta_{design}}\right)$$

where H_b is the time that the device operates at level b (e.g. hours), S_b is the output at level b (e.g., fraction of design load), η_{design} is the design efficiency, and η_b is the efficiency at output level b. The cumulative *input* losses can be calculated by summing the individual losses at each part-load bin for the entire year (or other period). That is,

$$L_{cum} = \sum_{b=0}^{S_{max}} H_b \times S_b \left(\frac{1}{\eta_b} - \frac{1}{\eta_{design}} \right)$$

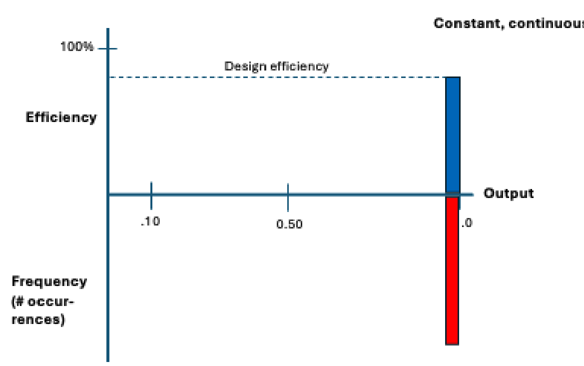
where S_{max} is the maximum output. (S_{max} may exceed the design capacity. This situation corresponds to a device being overdriven and, in some cases will be as much as 20% above design capacity.) When the observed efficiencies are everywhere equal to the design efficiency, the cumulative part-load losses are zero. (This occurs when the device operates at a single level of output.) More commonly, the observed efficiencies everywhere are less than the design efficiency (that is $\eta_{design} > \eta_b$); in that case, the cumulative losses will be a value between zero and L_{cum}.

A dimensionless metric of performance can be obtained by dividing this value by the cumulative load assuming design efficiency:

$$PLM = \frac{\sum_{b=0}^{S_{max}} H_b \times S_b \left(\frac{1}{\eta_b} - \frac{1}{\eta_{design}} \right)}{\sum_{b=0}^{S_{max}} H_b \times S_b \left(\frac{1}{\eta_{design}} \right)}$$

This ratio corresponds to the extent to which a device's actual efficiency departs from the design efficiency and ranges from 0 upwards, where zero corresponds to perfect alignment (e.g., no departure from design). Put another way, engineers should strive to achieve the lowest PLMs feasible.

Several use cases can be graphically examined to understand the limits of the metric. The following examples in Figure 6 illustrate the different efficiency maps and frequency distributions.

Case	PLM	Distributions
Case 1. The device operates continuously at constant load This device operates at only one level of output and continuously. In this case, $\boldsymbol{\eta}_{design} = \boldsymbol{\eta}_b$. So the numerator is zero and the metric is zero.	0	

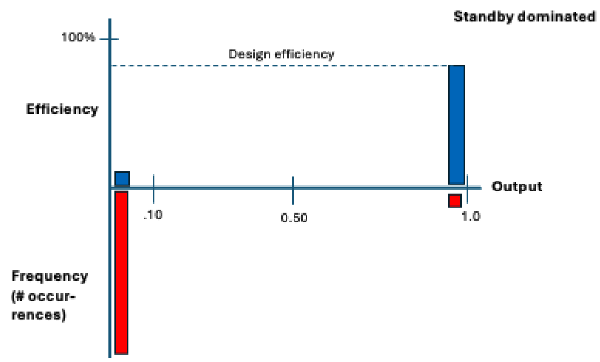
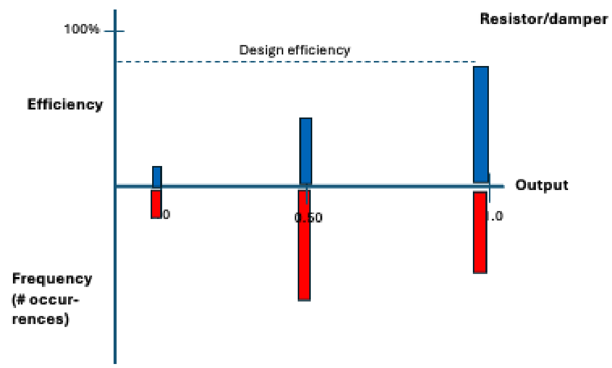
Case 2. The device operates nearly all the time at very low or zero output. This device operates in only two modes: full-load, which happens only briefly, and an extremely low part-load mode, which happens most of the time. This corresponds to many devices that spend most of their time in standby mode.	2.5	
Case 3. The device operates at several levels of output, with varying amounts of time in each condition. The device operates in three modes. It corresponds to a two-speed fan with a standby mode. It operates most frequently at about 50% of design capacity. A resistor or damper may be used to modulate its output.	0.1	

Figure 6. Part-load metrics (PLM) and distributions for three common operating behaviors of devices.

A PLM was calculated for a fourth operating behavior: a staged system consisting of two large, water pumps. In this case, we drew upon actual measurements (Johnson et al. 2021). The system consists of two nearly identical pumps in a municipal water treatment facility. Output and efficiency were measured every 5 minutes. The data "cloud" of efficiencies reflects small variations in instantaneous conditions and measurement uncertainties (for example, cycling across measurement intervals). The system's peak efficiency occurs at about 60% output because the first-stage pump is slightly more efficient than the second. The PLM is approximately 0.34. The distributions of operating times and efficiencies are shown in Figure 7.

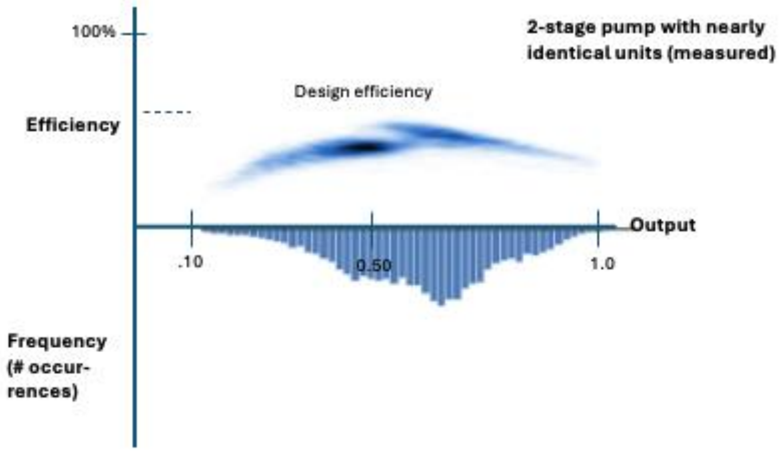


Figure 7. Monitored data from a large water treatment plant in Massachusetts with two, staged pumps.

The PLM derived above describes the extent to which existing capacity is under-utilized. A PLM near zero corresponds to little deviation from the maximum efficiency that can be achieved with the existing equipment continuously operating at design efficiency. Higher PLMs occur if a device operates significant portions of the time at part-load, that is higher deviations from full-load efficiency. The part-load metric is a measure of "capital inefficiency," that is, the extent to which a device's capacity is wasted. If a device operates continuously at its design efficiency, then there is no wasted investment. If the engineer has installed excess capacity that is never used to meet the load, then the PLM will be higher.

Discussion

Oversizing—and the excessive part-load operation that it causes—occurs in almost every category of devices and it has three consequences. First and foremost, oversizing is an investment problem, because a cheaper device with lower capacity could meet anticipated loads. Second, oversizing causes devices to operate more hours and at lower part-loads than originally intended, leading to difficulties in control and reliability. Finally, oversizing wastes energy and increases operating costs because devices operate at efficiencies lower than anticipated. For these reasons, engineers need to more carefully estimate the requirements for a device's service and understand the technical drawbacks arising from specifying equipment whose capacity will never be fully required.

Part-load operation is a necessary feature of a device's behavior to match its output with the load. Part-load operation occurs in nearly every kind of device; it takes place in mechanical, electrical, and thermal systems. In spite of the huge diversity of devices, part-load solutions fall into three broad categories:

- Constrain the output of the device
- Adjust the internal behavior of the device
- Link the device's output to energy storage or other applications

Even though the solutions may appear different across devices, the strategies are fundamentally the same. Technological advances may enable new part-load solutions to be transferred from one application to another. This paper's contribution is the recognition of the cross-cutting nature of part-load problems.

Conclusions

This paper proposes a metric for evaluating the extent of a part-load problem in a device. The metric enables an engineer to compare the impacts of different part-load solutions. For example, is staging superior to variable capacity in a particular application? The metric is sufficiently general so that it can be applied to a wide range of situations and compared. Should we expect the PLMs for a furnace fired by natural gas and an electric heat pump to differ? More broadly, how do PLMs compare across different applications? These questions can only be answered with higher resolution measurements. Fortunately, they are becoming increasingly easy to collect.

Several research opportunities emerge from this paper that can ultimately reduce both investments and operating costs. These are listed below.

1. In the laboratory, test energy efficiencies of devices at much lower part-loads than today. In that way, engineers can better estimate the energy impacts of low part-load operation.
2. In the field, monitor output and efficiency at greater time resolution. These measurements can contribute to a better understanding of the size and scope of oversizing and the part-load problem.
3. Design systems to more readily accommodate smaller, low-capacity stages. (Each industry has a different name for these stages: auxiliary, "pony," standby, etc.) These stages can be more closely integrated with primary stages and efficiently operate for long periods. This functionality could be the norm rather than the exception for modulating devices.
4. Treat standby power use (i.e., energy consumption with zero output) as an important characteristic of part-load operation. Standby often represents a significant fraction of a device's total energy use. Some technologies can enable reductions in standby power for broad classes of devices.
5. Acknowledge that oversizing is not solely an engineering problem—risk aversion creates strong institutional pressures to oversize equipment. These pressures include insufficient time to perform a careful load calculation, avoiding user call-backs, inadequate information, and fear of liability. The cost of oversizing is typically invisible to the client, which further encourages the design engineer to err in favor of larger capacities. Given the inevitably of oversizing, technical strategies that maintain high energy efficiency and reliability at low part-load and institutional mechanisms to encourage such strategies are critical. These strategies must consider the twin perspectives of efficient use of capital and energy.

Acknowledgment

This work was supported by the Office of Critical Minerals and Energy Innovation, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

Benaissa, Manal, Georges Da Costa, and Jean-Marc Nicod. 2022. "Standalone Data-Center Sizing Combating the Over-Provisioning of the IT and Electrical Parts." *2022 International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW)*, November, 57–62. https://doi.org/10.1109/SBAC-PADW56527.2022.00019.

Dahl, Trygve. 2018. "The Theory and Application of True Weighted Efficiency—A New Metric to Evaluate Pump Energy Efficiency Considering Multiple Operating Conditions." Paper presented at 47th Turbomachinery & 34th Pump Symposia. September.

Djunaedy, Ery, Kevin Van Den Wymelenberg, Brad Acker, and Harshana Thimmana. 2011. "Oversizing of HVAC System: Signatures and Penalties." *Energy and Buildings* 43 (2–3): 468–75. https://doi.org/10.1016/j.enbuild.2010.10.011.

Johnson, Hilary A., Kevin P. Simon, and Alexander H. Slocum. 2021. "Data Analytics and Pump Control in a Wastewater Treatment Plant." *Applied Energy* 299 (October): 117289. https://doi.org/10.1016/j.apenergy.2021.117289.

Less, Brennan D., and Iain S. Walker. 2026. "Characterizing Peak Electricity Demand for U.S. Households: An Assessment of End-Use Loads and Demand Factors." *Applied Energy* 410 (May): 127422. https://doi.org/10.1016/j.apenergy.2026.127422.

Murphy, Sean, Lixi Liu, Brennan Less, et al. 2026. "Characterizing Electrical Panel Capacity, Breaker Space, and Loads in U.S. Single-Family Homes." *Journal of Building Engineering* 120 (February): 115576. https://doi.org/10.1016/j.jobe.2026.115576.

Next Generation Lighting Industry Alliance (NGLIA) LED Systems Reliability Consortium. 2019. "Dim-to-Warm LED Lighting: Initial Benchmarks." DOI:10.13140/RG.2.2.10417.53609

Takci, Mehmet Türker, Meysam Qadrdan, Jon Summers, and Jonas Gustafsson. 2025. "Data Centres as a Source of Flexibility for Power Systems." *Energy Reports* 13 (June): 3661–71. https://doi.org/10.1016/j.egyr.2025.03.020.

Van Geet, Otto, and David Sickinger. 2024. *Best Practices Guide for Energy-Efficient Data Center Design*. NREL/TP--5R00-89843; DOE/GO--102024-6283. National Renewable Energy Laboratory (NREL), Golden, CO (United States). https://doi.org/10.2172/2417618.

Woradechjumroen, Denchai, Yuebin Yu, Haorong Li, Daihong Yu, and Huojun Yang. 2014. "Analysis of HVAC System Oversizing in Commercial Buildings through Field Measurements." *Energy and Buildings* 69 (February): 131–43. https://doi.org/10.1016/j.enbuild.2013.10.015.