Figure 3a shows estimated total capital cost of three different storage technologies, measured in dollars per kilowatt ($/kW), which is the standard measure of power plant costs.[3] The figure shows the initial power component associated with power conversion equipment, which could be thought of as building a storage plant with no energy storage capacity (duration), or only the power part of the plants illustrated in Figure 2. As duration is added, the costs increase at a rate that is assumed to be constant and is equal to the cost for each additional hour of storage duration (dollars per kilowatt-hour) multiplied by the number of hours.[4] This is also reflected in the cost per incremental hour, shown in Figure 3b, which starts with the first hour and includes the power cost, and then each additional hour requires only the energy-related costs.

We use these three technologies to illustrate the important difference between energy- and power-related costs. For example, Li-ion batteries have a relatively low power-related cost but a high energy (duration) cost, which results in the highest rate of increase in capital cost as a function of duration. Our pumped storage curve represents a technology with higher power-related costs but a lower cost per unit of incremental energy. For pumped storage in particular, this curve is only illustrative, as pumped storage costs are typically site-specific and can leverage significant economies of scale for both power- and energy-related costs. This also means pumped storage typically has a declining incremental cost as a function of duration (which would produce a downward slope in Figure 3b as opposed to the constant incremental value illustrated); so, the total system increase is not actually linear as shown in Figure 3a. For this reason, we include a dashed line to capture the considerable uncertainty around these cost assumptions, particularly as pumped storage historically is constructed with 8 or more hours of capacity. Finally, we include a boundary case of hydrogen storage with the highest power-related costs but that also uses an underground formation that assumes close to zero cost for additional duration after the initial development.

Figure 3 is simplified and used to demonstrate the basic relationships between storage duration and capital costs. This relationship will strongly influence the overall economic performance of storage technologies used for different applications; therefore, understanding the impact of duration on overall storage value is critical.

Figure 3 includes only the initial capital costs. The total life-cycle cost of storage technologies includes several other important components that vary by technology. Technologies with shorter calendar lives or higher cycling-introduced degradation will require more frequent replacement or refurbishment of key components. Variable operation and maintenance costs will also vary by technology, while round-trip efficiency impacts the cost of charging electricity needed to provide different services. These factors are considered when evaluating the total economic performance, as discussed in Section 3.3.

[3] Note that all costs are measured in terms of the AC rating, as the grid uses AC power.

[4] Ideally, this measurement represents usable energy, after accounting for state of charge limitations, conversion to AC, and other factors.

3.2 Storage Benefits and Values

Storage can provide an array of services that can largely be represented by four general classes (Table 1) that capture over 95% of the costs of operating the bulk power system.[5]

Table 1. Four Major Categories of Bulk Power System Storage Services

Service	Description
Capacity	Firm capacity
Energy	Energy shifting/dispatch efficiency/avoided curtailment
Transmission	Avoided capacity, congestion relief
Ancillary services	Operating reserves, voltage support

Note that Table 1 does not explicitly list RE-specific applications, such as "renewable firming" or "renewable time-shifting." These applications are specific cases of the more general applications listed and are therefore already captured in Table 1.

Likewise, Table 1 captures some applications that can be provided by behind-the-meter storage. For example, firm capacity and energy shifting value is reflected in tariffs by demand charges and time-of-use rates. However, the table does not include several additional values provided by distribution- or customer-sited storage, including avoided upgrades and local reliability and resiliency. We focus here exclusively on utility-scale storage; other analyses within the Storage Futures Study examine the potential value, costs, and potential adoption of behind-the-meter storage.

3.3 Storage Economic Performance Metrics

The simplest economic performance metric commonly applied to generation technologies is the levelized cost of energy (LCOE). It measures the delivered cost of energy, including both fixed and variable costs, and it also includes the impact of financing, expected life, and expected annual energy production. A similar metric is the levelized cost of storage (LCOS). It includes all fixed and variable cost components over the life of the storage plant, including charging energy and the impact of round-trip efficiency. The limitations of LCOE and LCOS as a stand-alone performance metrics are widely documented, but the most obvious is that they provide no indication of the value of the energy or other services potentially provided (10). This is particularly problematic when comparing storage technologies that provide fundamentally different services (e.g., short-duration storage that provides only operating reserves) to longer-duration technologies (e.g., pumped storage that provide multiple services, including firm capacity, time-shifting, and operating reserves).

To properly evaluate the economic performance of storage, metrics that consider both costs and benefits must be used. The actual metric used depends on the perspective of the owner or operator (which may not be the same entity). Vertically integrated utilities and other regulated entities typically use a least-cost planning approach, which is sometimes referred to as integrated

[5] By the bulk power system, we mean the high voltage transmission system and generators but not the distribution network. The 95% value is derived from data in PJM (8) and ISO-NE (9).

resource planning (11). This approach compares various resources over an extended (i.e., multidecade) period to derive a least-cost mix while considering reliability and various policy constraints. While the final performance metric is expressed in terms of a cost (e.g., a net present cost or even an LCOE of the entire system), the value of various services provided by the entire system is embedded in this cost. For example, storage acting as a peaking plant can reduce operating costs across the generation fleet, and this benefit is reflected in a reduced system cost.

A second approach—which an independent power producer might use—is to evaluate the economic performance of storage in isolation, and to then compare its life-cycle costs to life-cycle revenue to determine whether this results in a satisfactory rate of return or other economic performance metric. This second approach can be easier to calculate, and we use it in several examples in this work to illustrate the cost-competitiveness of batteries for several applications.

Evaluating the economic competitiveness of storage in either a least-cost portfolio or as a stand-alone investment requires an additional analysis element: determination of the optimal duration of storage. This analysis is unique to storage and compares the incremental costs of storage duration to the incremental benefits of the added duration, thus ensuring the value of adding more hours exceeds the costs. This type of evaluation of duration drives the markets for energy storage, and in the following sections, we use the framework of the four phases to examine the relationships of storage duration, value, and applications.

3.4 Competing Flexibility Technologies and Approaches

This report focuses on economic drivers of new storage deployment. In many cases, such as the use of storage to provide peaking capacity, storage is compared to a traditional gas-fired generator. However, the flexibility and value that storage provides can also be supplied by other competing technologies, including demand response, managed charging of electric vehicles, and other sources of flexible supply or demand. The framework presented here could be applied to other approaches to provide grid flexibility and could demonstrate variations in the competitiveness or storage or the overall market potential.

4 Phase 1: Short-Duration Storage for Providing Operating Reserves

After the minimal deployment of storage that occurred after the 1980s, interest in storage was renewed in the early 2000s with the convergence of several events. One was the creation of wholesale markets. These markets eventually included several operating reserve products that provided storage an opportunity to directly compete and demonstrate its potential value compared to resources that have traditionally provided these services (12).

While operating reserves consist of numerous services and market products, they all represent the ability of a generator or aggregated set of generators to increase output (provide “upward” reserves) or decrease output (provide “downward” reserves) (13). These reserves are provided in response to random variations in supply and demand at various time scales. The distinctions between different reserve services can be characterized by three factors:

- *How much* reflects the quantity of power potentially needed by the system, or how much headroom is needed from the set of plants providing this service; this is measured in power capacity (megawatts [MW]).
- *How fast* reflects the response rate needed or how quickly the set of plants providing the services are required to move from one setpoint to another (MW/second) and is a combination of the time needed to initiate a response to the reserve event and ramp rate.
- *How long* is the duration for which the plants must hold the new output, and for an energy storage device, represents the amount of stored energy (megawatt-hours [MWh]).

The application of these three factors to a single plant is illustrated in Figure 4, which shows the output of a generator that is operating below maximum output and able to provide some reserve capacity based on its operating limits.

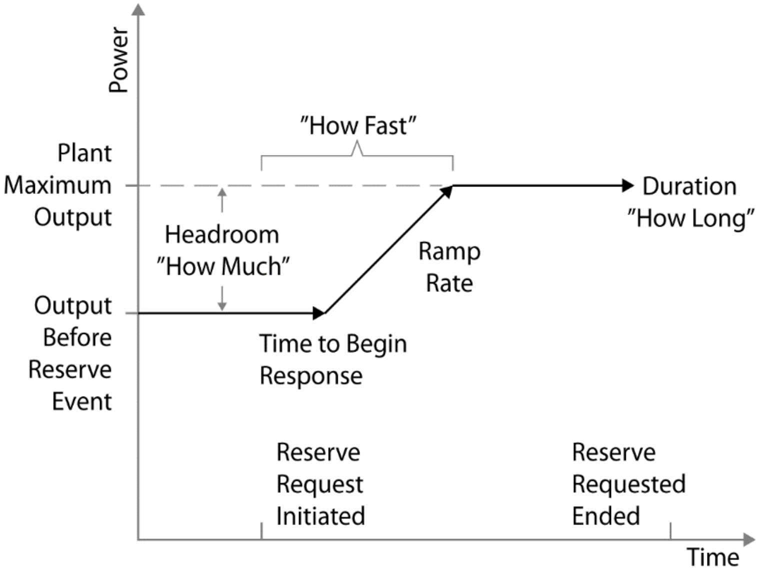


Figure 4. The three characteristics of operating reserves

Though specific operating reserve products have different names depending on the region, we focus here on two major classes that exist in all market regions in the United States and which offer higher value:

- *Spinning contingency reserves* are used to respond rapidly to address the failure of large power plants or associated transmission lines.
- *Regulating reserves* are used to address smaller, random variations in supply or demand.[6]

We do not consider non-spinning or supplemental reserves, which are slower to respond and require a long-duration response of multiple hours.

In brief, "how much" of each reserve produce is required is determined by each balancing area.[7] How much of this can be provided by any individual storage asset is determined by its power capacity and operating state, meaning it must operate at less than full output (holding headroom) and hold sufficient energy to respond to a reserve call or event.

Each balancing area also establishes rules for the response rate ("how fast") required for generators to participate in the provision of operating reserves. For example, an operator might require a generator to increase output in 10 minutes for the provision of spinning contingency reserves.

Rules also establish the length of time a unit must be held at the increased output, typically in the range of 15–30 minutes. As an example, spinning contingency markets may require a resource providing this service to hold output for at least 30 minutes (reflecting the time needed to bring up additional generation capacity) (15). For a 1-MW storage device to be able to provide this service, it would need to have 30 minutes of discharge capacity or 0.5 MWh of stored energy. A device with less energy capacity (duration) could still provide this service, but with a lower power rating. For example, a 1-MW device with 15 minutes of capacity (0.25 MWh) would need to discharge at 0.5 MW to supply power for the 30-minute interval.

Value of Phase 1 Services

Historical market values can be used to estimate an approximate value for energy storage providing various services. Prices for operating reserves are often measured in units of capacity available during 1 hour (MW-hr). This is not a unit of energy—it represents *capacity* that is available for a response over a period of time.[8] A facility providing spinning contingency reserves is paid for this provision even if there are no calls for providing energy; more simply stated, the plant is paid for doing nothing other than being ready to respond and then responding if called to do so. The average spinning contingency reserves market prices in 2019 ranged from

[6] Though regulating reserves are sometimes referred to as "frequency regulation," the North American Electric Reliability Corporation glossary defines frequency regulation to include both governor response (frequency response) and the service described in this section (14). To avoid potential confusion, we use the term regulating reserves.

[7] The balancing area is the entity responsible for balancing supply and demand, including the provision of operating reserves. Depending on location, this can be a market operator (independent system operator/regional transmission organization) or a vertically integrated utility.

[8] Some regions use cost per kW-month. This is also is similar to how payments in capacity markets may be measured in kW-yr, or the provision of 1 kW of capacity for a 1-year period.

$3/MW-hr to $27/MW-hr, with a national weighted average of about $11/MW-hr. A storage plant providing contingency reserves would essentially "idle" in a charged state, waiting for a response, and then be paid for the whole time. When called, the plant would then discharge until the end of the event and then recharge as soon as possible so that it could return to a charged state and provide reserves again.

Regulating reserves are more complicated. Unlike contingency reserves, which are rarely used, provision of regulating reserves requires a unit to change output fairly frequently in response to small, random variations in demand (16).[9] A storage plant providing regulating reserves would sit at a condition with a high state of charge, but it would continuously increase and decrease output in response to grid needs. This frequent, shallow cycling would incur additional costs that are due to battery degradation, as well as costs for make-up energy to compensate for losses associated with the potential substantial energy throughput. However, prices for regulating reserves are typically higher than those for contingency reserves; average 2019 regulating reserve prices in market regions ranged from about $6/MW-hr to $32/MW-hr for combined up and down reserves, and the weighted national average was about $15/MW-hr.[10] This capacity-related payment is often supplemented by a payment associated with increasing or decreasing output.[11]

In either case, the device could provide operating reserve services close to 100% of the time, limited mainly by periodic maintenance. Such provision of services nearly 100% of the time has enabled a cost-effective entry point for energy storage, even before recent cost declines.

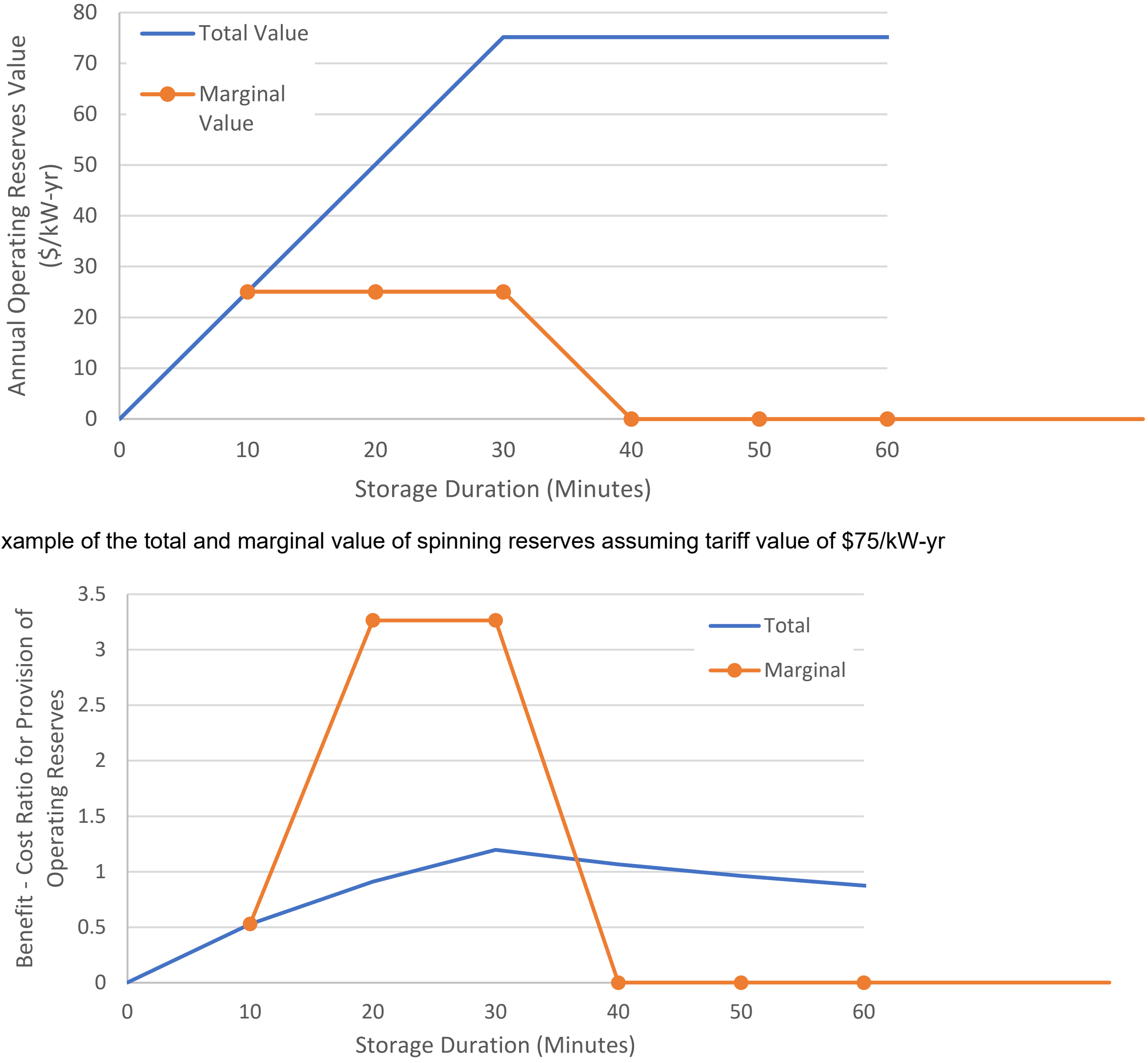
A key aspect of these services is that once the minimum duration requirement ("how long") is met, there is no additional value for additional duration (stored energy) for a given market service. If the duration requirement were 30 minutes, a 1-MW battery with 1 hour of storage would receive no additional value compared to a 1-MW battery with 30 minutes of storage. Figure 5a shows a specific example of the value of a device providing spinning reserves under the Arizona Public Service tariff of $6.26/kW-month ($75/kW-yr)[12] for spinning reserves and a 30-minute requirement. A 10-minute battery would need to be derated to one-third of its capacity to provide 30 minutes of service, and therefore this battery would receive about $25/kW-year (with 100% availability). Each dot on the "marginal" curve shows the value of an incremental 10 minutes of storage. The total value increases as storage duration is added until the battery reaches 30 minutes of duration, at which point the value of adding additional duration for this service is zero.

[9] As a result, a fourth parameter associated with operating reserves might be described as "how often" or the frequency at which a reserve service is called. It ranges from to a few times month for spinning contingency reserves to nearly continuously for regulating reserves.

[10] Higher prices for regulating reserves partially reflect wear and tear and degraded performance of the thermal plants that often provide these services. Data includes CAISO, ERCOT, ISO-NE, MISO, NYISO, PJM, and SPP (13).

[11] This is sometimes referred to as a mileage payment. In addition, regulating reserve prices may vary based on the speed and accuracy of a plant following a regulations signal, with an example being PJM, which has two regulating reserve markets separated by speed of response (13).

[12] "Arizona Public Service Company Pro Forma Open Access Transmission Tariff," http://www.oatioasis.com/AZPS/AZPSdocs/APS_OATT_Volume_2_20170601.pdf



(a) Example of the total and marginal value of spinning reserves assuming tariff value of $75/kW-yr

(b) Example of total and marginal B/C ratio assuming 90% battery availabliity and 2020 Li-Ion cost estimates

Figure 5. Example of the value of battery storage providing operating reserves

Figure 5b shows the total and marginal benefit-cost (B/C) ratios for a Li-ion battery providing this service, assuming 2020 battery cost estimates from Figure 3 (7) and 90% device availability. The figure shows that because of the high initial cost, a 10-minute device has a B/C ratio of less than one. However, the B/C ratio of adding a second 10 minutes is much higher, as it is considering the incremental benefits of just the additional energy. Adding a third 10 minutes of duration (for a total of 30 minutes) maximizes the B/C ratio of project as a whole. Adding another 10 minutes to the project (to a total of 40 minutes) would still provide a B/C ratio of greater than 1 for the project as a whole even though the marginal B/C ratio would be zero. So, examining a 40-minute project in isolation could appear to be a reasonable investment. However, increasing the duration from 30 to 40 minutes actually lowers the overall return on the project and would actually result in a nonoptimal investment, which is also reflected in the zero marginal value curve after 30 minutes. This change in marginal value is why it is critical to compare marginal costs and benefits as a function of duration.

The ability of storage to provide cost-competitive operating reserves has resulted in significant deployments of energy storage in the United States of 1 hour or less (Table 2).[13]

Table 2. Phase 1 Utility-Scale (>0.5 MW) Storage Deployment with 1 hour or less capacity, 2011–2019

Region	Deployment (MW)
Alaska & Hawaii	27
California	139
Non-CAISO Western Interconnection	29
Texas	108
PJM	182
New York & New England	66
Other Regions in the Eastern Interconnection	171
Total	**721**

The capacity in Table 2 includes 656 MW of Li-ion batteries, 47 MW of flywheels, and 18 MW of other battery types.

Limits to Phase 1: Total Reserve Requirements

Phase 1 is limited by the total amount of high-value operating reserves needed in the U.S. power system, as summarized in Figure 6.[14] Regulating reserves requirements are driven by the size of normal variability in net load,[15] and contingency reserves are driven by the size of the largest expected power plant or transmission line failure in each region. The total requirement for these two services in the conterminous United States is about 18 GW; for comparison, peak demand is more than 600 GW. This means these markets have the potential to quickly saturate.[16]

Beyond existing operating reserve markets, monetization of existing frequency response requirements could add to the size (cumulative deployment of storage) of Phase 1. Frequency response, which is the ability of generators to respond rapidly and automatically to changes in frequency, is typically provided by generators equipped with governors. This is currently a market (i.e., a compensated) service only in the ERCOT region, but interest in a frequency response market is growing in other regions (13). Estimating the potential value of frequency response is difficult; regulating reserves market prices may be a reasonable lower bound proxy given the higher response rate needed for frequency response. However, like other reserve types described above, the total potential market is limited, with the total frequency response

[13] Data from EIA Form 860 for the year 2019. https://www.eia.gov/electricity/data/eia860/

[14] Data are derived from sources described in (13). For the requirement in nonmarket regions, we multiply the percentage requirement of a large utility in that region by the total peak demand of the larger region in which it is located. This means we use the requirements of a single utility as the proxy for the larger region as a whole.

[15] This loosely represents the largest very rapid and unpredictable change in either load or VRE that occurs in a few minutes. This is typically small (a few percent), as most changes occur over longer timescales (16).

[16] The chart in Figure 6 does not include non-spinning or replacement reserves, which may have multihour requirements and could potentially be served by storage in Phase 2 or beyond. The value of these services is historically much lower than the reservices considered here (13).

requirement in the entire United States being about 8.2 GW (13). Overall, this total results in a technical potential of short-duration operating reserves of less than 26 GW (Figure 6), with an economic opportunity for storage in Phase 1 likely being substantially less than this, particularly with competition from demand response including controlled EV charging, and existing PSH.

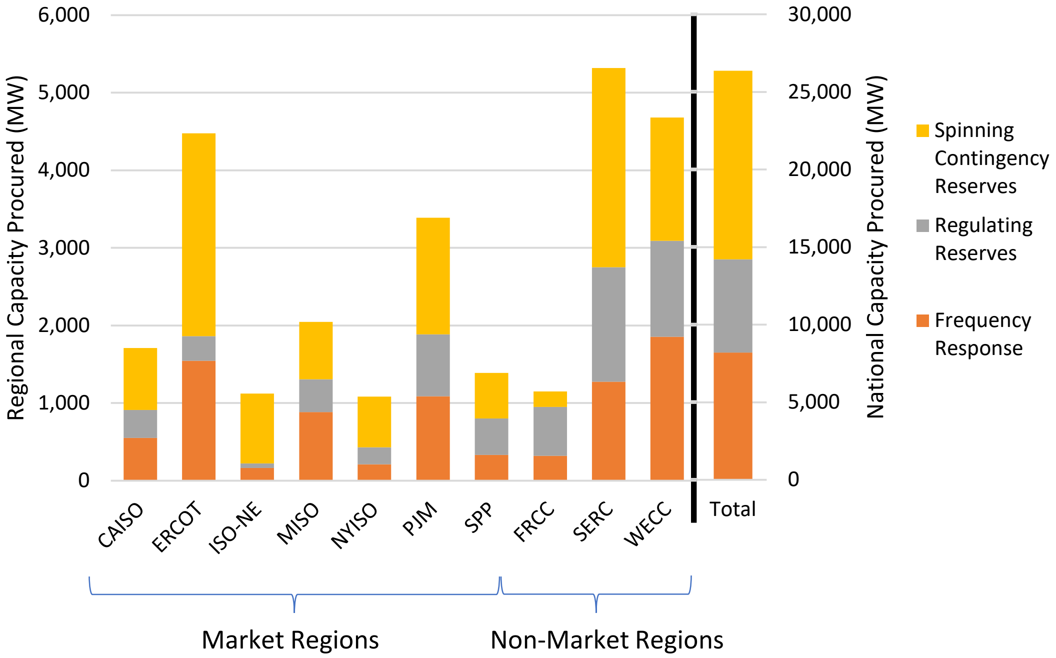


Figure 6. Current U.S. grid requirements for high-value operating reserve products potentially served by energy storage in Phase 1

Two additional factors could extend Phase 1. The first is additional market products, including a flexibility/ramping reserve product that has been created to address additional variability and uncertainty in the minutes-and-beyond time scales resulting from VRE deployment (13). This product has been introduced in a limited number of regions, and it typically requires a slower response rate (i.e., lower ramp rate) and longer duration (i.e., the resource holding output for longer) than regulating reserves.

The second factor that could extend Phase 1 is potential growth in regulating reserves that may result from increased deployment of VRE resources. However, several studies have found that much of the increase in variability occurs in time scales in the minutes-or-longer time scales (17, 18), and it may drive the creation and use of a flexible ramping reserve, as opposed to very large increases in regulating reserves.[17] So overall, there is no clear evidence for a very large growth in operating reserve requirements as a result of large VRE deployments. Overall, though it is limited, Phase 1 represents an important entry point for storage, particularly in regions with little prior storage deployment and for services where storage can offer higher value.

[17] Deployment of VRE resources is unlikely to affect the requirements for contingency reserves and frequency response, where requirements are typically set based on the sizes of the largest likely system failures (13).

5 Phase 2: The Rise of Battery Peaking Power Plants

As Phase 1 operating reserve markets saturate and declining battery prices create new opportunities, we transition to Phase 2: the deployment of batteries with about 2–6 hours of duration for providing peaking capacity. Peaking capacity is used to meet short periods of peak demand on hot summer days, or in some locations, in periods of extreme cold. Peaking capacity is typically provided by simple-cycle gas turbines, older gas steam plants, or internal-combustion generators (1). However, the continued decline in the costs of Li-ion batteries has increased their competitiveness over traditional sources, and Phase 2 has already begun in some locations (19).

In Phase 1, we consider short-duration storage providing only a single service because of the power constraints of the battery. This means that for a battery providing upward reserves, the entire power capacity of the battery is dedicated to the possibility of needing to increase output. Given the continuous need for this capacity when providing operating reserves, even if additional energy capacity (duration) were added, this duration would be unable to be used for other services as long as the device is providing operating reserves. As a result, it is difficult for devices aimed primarily at providing operating reserves to provide additional services.

Alternatively, a battery peaking plant typically provides multiple services, including provision of physical capacity (capacity credit), the value of energy time-shifting, and operating reserves during certain periods. A battery peaking plant can provide both capacity and energy shifting services simultaneously because the periods of highest prices (when the battery will discharge to maximize revenue or minimize system costs) are very highly correlated to periods of highest demand when the system needs reliable capacity (20). Periods of low prices (when the battery will charge) are also periods of low demand, and therefore when large amounts of spare capacity are available and the risk of an outage is low. Therefore, these two services—capacity and time-shifting—do not double count either the energy or power capacity of the battery and can be "stacked." [18]

A multihour battery could also provide reserves in addition to capacity and energy shifting. When a battery is charging, it can provide upward reserves, as long as the battery has charged enough to meet the reserve duration requirement *and* has reserved sufficient energy to meet its capacity obligation. It can also provide upward reserves while "idle" in the period between charge and discharge.

As a result, there can be considerable overlap of Phase 1 and Phase 2, when batteries deployed in Phase 2 also provide operating reserves for additional revenue. At low enough storage costs, some regions may largely skip Phase 1 by using 2-hour to 6-hour devices to provide energy, capacity, and operating reserves. However, the same considerations of market saturation apply, particularly as batteries deployed in Phase 1 may reduce overall reserve prices. Therefore, we focus on the value of peaking capacity as the primary driver of the transition to Phase 2, though

[18] The concept of combining multiple services, or "value stacking" is not unique to storage, and many generation resources provide multiple service and thus inherently value stack, although this term is seldom used when talking about traditional generation capacity.