Optimal Purchases of Ancillary Services

An important feature of ancillary services is that the major categories constitute a hierarchy of declining qualities, measured in terms of their usefulness for maintaining the reliability and security of the transmission system. Although these categories are given different names in different jurisdictions, a typical list is the following:

1. Regulation. Generation units equipped with automatic generation control (AGC) can follow load variations on the time scale of seconds. This capability is limited, however, by the usual requirement that an AGC unit must return to its set point every few minutes, typically 10 minutes.
2. Spinning Reserve. The incremental generation that an active unit can ramp to within ten minutes and then sustain, typically for 30 to120 minutes, is considered spinning reserve. Some quick-release hydro units and curtailable loads can also serve this function.
3. Non-spinning Reserve. The incremental generation that can be achieved by units with slower responses, and those requiring startup, is considered non-spinning reserve. Jurisdictions differ as to the ramping time allowed, varying from 10 to 30 minutes. Fast-start combustion turbines can serve this function.
4. Replacement or Operating Reserve. Incremental generation that can be obtained in the next hour to replace spinning and non-spinning reserves used in the current hour is classified as replacement or operating reserve. (FERC does not consider replacement reserve to be an ancillary service.)

In real-time operations, one can view these categories as a cascade of options that the system operator (SO) can invoke to cap the real-time price. Regulation is the preferred alternative to meet small short variations, and to meet larger variations with incs and decs offered as supplemental energy bids in the balancing market. But as these preferred options expire during a sustained surge of demand, the second alternative is to instruct spinning reserve units to ramp up, and if that fails to instruct non-spinning reserve units to start up; and finally, as these units are called they must be replaced by operating reserves if the required reserve margins are to be maintained.

The definitions of the last three categories are unduly narrow in that they focus entirely on incremental generation. This restriction is a holdover from an earlier era. In competitive markets it is often equally necessary to maintain decremental reserves. A further peculiarity is that there is no category of reserves that is explicitly intended to provide load following on a large scale. In principle, supplementary offers of increments and decrements in the real-time balancing market are intended to serve this function. In practice, however, the system operator (in jurisdictions such as California) has been reluctant to count on there being sufficient bids in the balancing market in shoulder hours when the ramp rate must be quite high to meet the sharply increasing or decreasing load. The result has been that instead of the usual allowance of 2 or 3% regulation reserves, the system operator has acquired five times as much, thereby enabling the AGC units to provide load following.

The reserve requirements are generally stated in terms of a percentage of the prediction demand load, and indeed it is typically demanders who reimburse the system operator for the reserves that it purchases. These percentages are partially restricted by the regional reliability councils and NERC. For example, the WSCC requires that the sum of the spinning and non-spinning reserves must be at least 7% of the load served by thermal units, and 5% of the portion of the load served by hydro units. A typical pattern is, say, 3% regulation (in non-shoulder hours), 3.5% spin, 3.5% non-spin, and 3% replacement. These percentages are derived from prior experience in vertically integrated utilities and tight power pools, and will likely change as experience is acquired in the newly restructured competitive markets.

In the initial months, however, these reserve requirements were implemented in a way that revealed the importance of the system operator's (SO's) active role in acquiring reserves at least cost. First we describe the initial implementation and the problems that ensued, and then we describe a better one.

The Initial Implementation

The initial implementation started from the recognition that bids were typically offered for the best quality service for which a unit might be accepted. For example, a unit equipped with AGC might be offered for regulation; failing that, for spin; failing that, for non-spin, and failing that for replacement. This was reinforced by the FERC requirement that bids rejected for higher a quality ancillary service must (if the bidder chooses) to be consider also for lower quality services. These considerations resulted in the procedure of sequential auctions, in which the auctions for regulation, spin, non-spin, and replacement were conducted in sequence.

The major problem resulting from this implementation was that the prices of the lower quality reserves were often higher than the prices of higher qualities. In California, for instance, on several occasions the price of replacement reserve exceed $1000/MW in hours when regulation and spin were acquired at prices below $30/MW. And of course these high prices for the lower qualities produced a perceived shortage of offers for the higher qualities, at least from those suppliers not subject to FERC-imposed price caps.

The explanation for this problem is evidently the fact that the SO set firm requirements for each reserve category, and then conducted the auctions in sequence. Such a procedure makes it possible that a unit rejected for regulation because its offered price was $20 when the clearing price was $10, could be accepted for spin at a clearing price of $30. Thus, there is no assurance that the clearing prices for successively lower qualities are declining.

A Better Implementation

A better implementation is based on the recognition that each reserve category can substitute for the next lower quality. Thus, the SO should purchase additional regulation to substitute for spin whenever spin is more costly, additional spin to substitute for non-spin whenever non-spin is more costly, and so on. This purchasing policy precludes separate auctions conducted in sequence for the various ancillary services. In particular, it requires that the SO take full advantage of the substitutions allowed of higher qualities for lower qualities. One way to formulate this policy is to re-interpret the requirements as cumulative requirements, as in the following set of constraints:

Reserve	Minimal Requirement
Regulation	> 3 %
Regulation + Spin	> 6.5 %
Regulation + Spin + Non-Spin	> 9.5 %
Regulation + Spin + Non-Spin + Replacement	> 12.5 %

If the SO adopts the policy of accepting the bids that minimize its total cost of ancillary services subject to these four constraints, then necessarily the clearing prices will decline as the qualities decline.

The Active Role of the System Operator

The preferred implementation describe above is indicative of the more active role that the system operator must take in order to meet its reserve requirements at least cost. Taking advantage of possible substitution of higher qualities for lower qualities might be termed the "smart buyer" approach to purchases of ancillary services. A consequence is that the SO's demand for any one quality responds to the prices at which it can buy substitutes. As in any case that demand is sensitive to price, this dampens the market power of large suppliers. More generally, the SO can take more active measures to mitigate market power. One approach is to adjust its reserve requirements (the percentages in the table above) in response to the prices offered. For instance, WSCC allows the SO discretion in its choice of the quantities of regulation and replacement, and it restricts only the sum of the spin and non-spin requirements. Paying $1000 for replacement reserve is unnecessary if it is reasonably sure that supplemental energy bids in the balancing market can be obtained at a fraction of this price. In California this consideration led to a cap on the price of replacement reserve of $500 initially, and then later $250.

Real-Time Operations

An important consideration during real-time operations is the merit order in which supplementary bids and ancillary services are ranked. A unit reserved for, say, spin has a specified "reservation price" indicating that the supplier prefers not to be called to ramp up if the real-time spot price at which it will be paid for generation is less. Unlike a supplemental energy bid, however, this reservation price cannot be used directly to establish the spinning unit's ranking in the merit order. This is because activation of a spinning unit reduces the reserves available to meet requirements and therefore requires replacement. To find its place in the merit order the SO must take account of the reservation price *plus* the cost of activating sufficient replacement.

A supplier with a unit reserved for spin would, of course, like to "double dip," being paid both for reserving capacity and for energy generated when called in the merit order, so the natural tendency is to insist that it should be called for incremental energy whenever its reservation price is next in the merit order. The SO must insist, nevertheless, that by purchasing the option on the unit's capacity it has acquired the privilege of exercising that option only when the full costs of doing so are less than the alternatives offered by supplementary energy bids in the real-time market that do not incur the cost of replacement.