Technical Disclosure Commons

Defensive Publications Series

April 17, 2018

Capacity planning for computing clusters

Lu Huang

Ana Radovanovic

Radovanovic Ye

Alexandre Duarte

Walfredo Cirne Filho

Capacity planning for computing clusters

ABSTRACT

This disclosure provides techniques that enable elastic computing service providers to optimize allocation of computing resources across multiple tiers of service level objectives, while allowing for periods of oversubscription to be met by cross-tier movement of resources. Further, the techniques enable providers of computing services to plan their computing investments given rates of oversubscription and anticipated demand across tiers. Service providers frequently plan investments based on peak, rather than average, demand and may therefore be left with surplus capacity during periods of below-peak demand. While such surplus can be resold at lower levels of service guarantees, such reselling increases demand (oversubscription). By accounting for oversubscriptions and inflations of service level objectives across tiers, the techniques of this disclosure enable optimal multi-tier resource allocation.

KEYWORDS

- Elastic computing
- Computing cluster
- Resource allocation
- Service level objective
- Oversubscription

BACKGROUND

Elastic computing, in which a customer can dynamically deploy computing resources based on current demand, has gained popularity due to its attractive economics, e.g., reduced investment on fixed computing capacity on part of the customer. A provider of elastic

computing services is able to cater to variation in individual customer demand by planning for aggregate demand, e.g., statistical multiplexing of computing resources.

However, service providers generally plan and invest for peak aggregate demand, e.g., by reserving resources in anticipation of a surge in user demand. Such reservation of capacity can result in low utilization, and to the extent that utilization is low, represents inefficient use of capital. To alleviate low utilization, service providers resell unused resources at zero or reduced service level objectives (SLOs). Nevertheless, lower SLOs in turn increase demand, potentially to an extent beyond resources made available by the over-provisioning for higher SLOs. Therefore, it is appropriate for service providers to plan for their entire load, considering demand at all SLO levels. The presence of multiple-availability SLOs adds complexity to the capacity-planning exercise, such that cluster administrators often simply decide for a single-availability SLO, needlessly promoting workload that might work even with weaker guarantees.

DESCRIPTION

This disclosure presents formal techniques to provision computing capacity for clusters that offer multiple-availability SLOs. The techniques also ensure provision of the minimum capacity needed to support a given load, e.g., given a specific set of tiered demands. This disclosure utilizes the following definitions in this context:

- *Raw capacity* refers to the computing resources, e.g., processing power (CPU / GPU), memory, storage, etc., allocated to a computing pool. For example, raw capacity may be equivalent to hardware resources available to the computing cluster. By default, the entire raw capacity of a multi-tier (or multiple-availability SLO) computing cluster is assigned to the highest tier.

- *Designated capacity* refers to the vector of fractional raw capacities designated to support the load at each tier of a multi-tier computing cluster. The designation of raw capacities to different tiers is a policy decision, e.g., taken by a cluster administrator.

Example: In a six-tier computing cluster, the designated capacity is as per Table 1 below.

Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Tier 6
0.50	0.15	0	0.15	0.20	0

Table 1: Vector of designated capacity

The above table indicates that 50% of the raw capacity is designated to support the load at tier 1, 15% is designated to tier 2, zero percent to tier 3, etc.

- *Oversubscription fraction* refers to the fraction by which the designated capacity of a tier can be inflated in order to generate bonus capacity for that tier. The oversubscription fraction is based on analysis of load characteristics. By definition, the capacity designated to a particular tier cannot generate bonus capacity for higher tiers. Bonus capacity for a given tier depends on the fraction of its capacity that can be oversubscribed to load at same or lower tier. Oversubscription fraction is represented by a matrix, such that a cell in the matrix of oversubscription fractions indicates the fractional bonus capacity that can be generated at the column tier of the cell by the raw capacity designated to the row tier of the cell.

Example: In a six-tier computing cluster, the oversubscription fraction is as per Table 2 below.

	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Tier 6
Tier 1	0.03	0.05	0.02	0.10	0.20	0.25
Tier 2	0.00	0.10	0.00	0.10	0.00	0.25

Tier 3	0.00	0.00	0.09	0.10	0.20	0.25
Tier 4	0.00	0.00	0.00	0.22	0.00	0.25
Tier 5	0.00	0.00	0.00	0.00	0.20	0.25
Tier 6	0.00	0.00	0.00	0.00	0.00	0.25

Table 2: Oversubscription fractions for a cluster with six tiers

In the example of Table 2, if the cluster administrator designates 100 units of a particular computing resource to tier 3, then such designation generates 9 units of bonus capacity at tier 3, 10 units of bonus capacity at tier 4, 20 units of bonus capacity at tier 5, 25 units of bonus capacity at tier 6, and zero units of bonus capacity (by definition) at tiers 1 and 2. Oversubscription fraction varies with the pool, tier and duration, e.g., shorter durations typically are associated with larger oversubscription fractions. Bonus capacity arises under certain circumstances, e.g.,

1. when load patterns make resources available at lower tiers, e.g., the diurnal cycle consistently frees up resources during the night;
2. when there is confidence that allocated resources are not going to be fully used within the medium time horizon, e.g., four months; etc.

Oversubscription fractions are refreshed periodically, e.g., between once a week to once every few months. Once increased for a specific time period, oversubscription fractions cannot be decreased for that time period.

- *SLO inflation* of a tier is the multiplicative inverse of the SLO availability of that tier.

Example: In a six-tier computing cluster, the SLO inflation is as per Table 3 below.

Tier 1	**Tier 2**	**Tier 3**	**Tier 4**	**Tier 5**	**Tier 6**

1.00	1.20	1.40	1.60	1.80	2.00

Table 3: Vector of SLO inflation

In the example of Table 3, the inflation fraction 2.00 for tier 6 indicates that this tier is unavailable ½ = 50% of the time. On the other hand, load running on tier 1 is covered by a strict 100% availability.

SLO inflation is also represented as a matrix such that diagonal elements of the matrix are the multiplicative inverses of the SLO availabilities for the tiers, and off-diagonal elements are zero. The dimensions of the SLO inflation matrix are identical to the dimensions of the matrix of oversubscription fractions.

Example: The SLO inflation matrix for example 3 is as follows.

	Tier 1	**Tier 2**	**Tier 3**	**Tier 4**	**Tier 5**	**Tier 6**
Tier 1	1.00	0.00	0.00	0.00	0.00	0.00
Tier 2	0.00	1.20	0.00	0.00	0.00	0.00
Tier 3	0.00	0.00	1.40	0.00	0.00	0.00
Tier 4	0.00	0.00	0.00	1.60	0.00	0.00
Tier 5	0.00	0.00	0.00	0.00	1.80	0.00
Tier 6	0.00	0.00	0.00	0.00	0.00	2.00

Table 4: SLO inflation represented as a matrix

- *Tiered capacity* is the capacity at each tier after accounting for adjustments to raw capacity due to SLO inflation and oversubscription. Tiered capacity is the capacity that is used by the cluster admission control system.

The techniques of this disclosure enable providers of elastic computing services to calculate the tiered capacity based on raw capacity, oversubscription fraction and SLO inflation, using the following equations.

```
Capacity multiplier =                                          (1)
Designated capacity x
 (SLO inflation + Oversubscription fraction)

Tiered capacity = Capacity multiplier x Raw capacity           (2)
```

In the equations above, operators such as multiply (x) and add (+) are matrix operators, since the underlying variables (e.g., SLO inflation, oversubscription fraction, etc.) are matrices.

Example

Given designated capacity of Table 1, oversubscription fraction of Table 2 and SLO inflation matrix of Table 4, the capacity multiplier is calculated as

$$\text{Capacity multiplier}$$

$$= \begin{bmatrix} 0.50 & 0.15 & 0.00 & 0.15 & 0.20 & 0.00 \end{bmatrix} \times$$

$$\left(\begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} + \begin{bmatrix} 0.03 & 0.05 & 0.02 & 0.10 & 0.20 & 0.25 \\ 0.00 & 0.10 & 0.00 & 0.10 & 0.00 & 0.25 \\ 0.00 & 0.00 & 0.09 & 0.10 & 0.20 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.22 & 0.00 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.20 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.25 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 0.515 & 0.220 & 0.010 & 0.338 & 0.500 & 0.250 \end{bmatrix}.$$

Given the above capacity multiplier, and a raw capacity of 1000 computing units, the tiered capacity is calculated as

$$\begin{aligned}&\text{Tiered capacity}\\&= 1000 \times \begin{bmatrix}0.515 & 0.220 & 0.010 & 0.338 & 0.500 & 0.250\end{bmatrix}\\&= \begin{bmatrix}515 & 220 & 10 & 338 & 500 & 250\end{bmatrix}.\end{aligned}$$

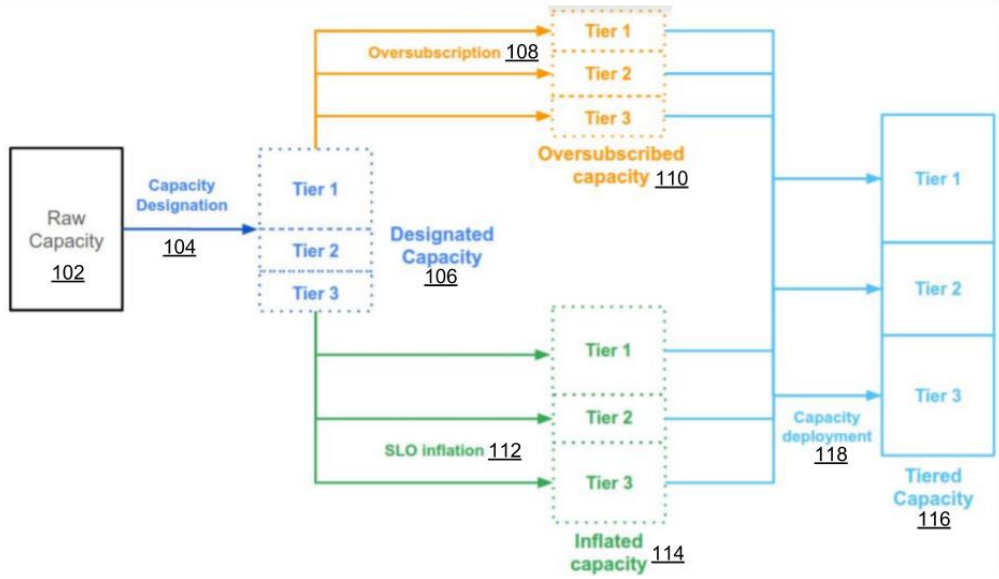

Fig. 1: Capacity pipeline

The process of transforming raw capacity to tiered capacity, e.g., via equations (1) and (2), is referred to as the capacity pipeline, illustrated in Fig. 1. A certain raw capacity (102) is converted to a designated capacity (106) via a capacity designating policy decision (104), e.g., taken by a cluster administrator. The designated capacity is expanded to an oversubscribed capacity (110) via oversubscription (108), e.g., a process based on load analysis. The designated capacity is expanded to an inflated capacity (114) via SLO inflation (112). The designated capacity, the oversubscription fraction, and the SLO inflation are combined to produce tiered capacity (116). The tiered capacity is used in admission control during deployment of capacity (118).

The techniques thus far have addressed the problem of deriving tiered capacity from raw capacity. This enables computation of the tiered capacity that can be provided on a specific cluster given a total amount of raw capacity, designated capacity, oversubscription fraction, and SLO inflation. However, cluster administrators also need the minimum raw capacity given tiered demand, in the presence of known SLO inflation and oversubscription fractions. Raw capacity is obtained from tiered demand by traversing the capacity pipeline in reverse. In mathematical form, the raw capacity is obtained by inverting equations (1) and (2), in order to get the equations below:

$$\text{Designated capacity} = \text{Tiered demand} \times (\text{SLO inflation} + \text{Oversubscription fraction})^{-1} \quad (3)$$

$$\text{Raw capacity} = \text{Designated capacity} \times \mathbf{1} \quad (4)$$

In equations (3) and (4), the notation $^{-1}$ refers to matrix inverse, multiplicative ($\times$) and additive (+) operators are matrix operations, and **1** refers to $[1, 1, \ldots, 1]^T$, a column vector of ones with the same number of elements as that of designated capacity. Thus, in effect, equation (4) computes raw capacity by summing the entries of designated capacity.

The solution of equations (3) and (4) results in the minimum amount of capacity required to support load over tiers covered by multiple-availability SLOs under the assumption that the load at each tier consumes at least the bonus capacity generated from the SLO inflation and oversubscription fractions. This assumption is met by the cluster administrator adjusting these fractions according to the load, up to an upper boundary.

CONCLUSION

This disclosure provides techniques that enable elastic computing service providers to optimize allocation of computing resources across multiple tiers of service level objectives, while allowing for periods of oversubscription to be met by cross-tier movement of resources. Further, the techniques enable providers of computing services to plan their computing investments given rates of oversubscription and anticipated demand across tiers. Service providers frequently plan investments based on peak, rather than average, demand and may therefore be left with surplus capacity during periods of below-peak demand. While such surplus can be resold at lower levels of service guarantees, such reselling increases demand (oversubscription). By accounting for oversubscriptions and inflations of service level objectives across tiers, the techniques of this disclosure enable optimal multi-tier resource allocation.