Towards Truthful Resource Reservation in Cloud Computing

Daniel Funke, Fabian Brosig and Michael Faber
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: daniel.funke@ieee.org; [fabian.brosig; michael.faber]@kit.edu

***Abstract*—Prudent capacity planning to meet their clients future computational needs is one of the major issues cloud computing providers face today. By offering resource reservations in advance, providers gain insight into the projected demand of their customers and can act accordingly. However, customers need to be given an incentive, e.g. discounts granted, to commit early to a provider and to honestly, i.e. *truthfully*, reserve their predicted future resource requirements. Customers may reserve capacity deviating from their truly predicted demand, in order to exploit the mechanism for their own benefit, thereby causing futile costs for the provider.**

In this paper we prove, using a game theoretic approach, that i) truthful reservation is the best, i.e. *dominant*, strategy for customers if they are capable to make precise forecasts of their demands and that ii) deviations from truthtelling can be profitable for customers if their demand forecasts are uncertain.

I. Introduction

Cloud computing grew immensely in recent years [4, 8]. One of the challenges Cloud Providers (CPs) face today are the hardly predictable computational demands of their Cloud Consumers (CCs). Therefore, choosing the appropriate capacity expansion policy to meet consumers' expectations is a major issue for all providers. In order to get a more precise understanding of their clients' needs, CPs can offer capacity reservation contracts [4, 8]. The benefits of such contracts are manifold. The providers gain insight into the potential computational requirements of the near future and can adjust computing capacity accordingly. The customers can rely on the reserved computation capacity to be available.

The reservation mechanism should balance the interests of CCs and CPs. On the one hand, reservation fees need to be expensive enough to avoid exploitation by the consumers. An over-reservation may let the CP increase capacity in false confidence that it will actually be consumed and generate revenue. On the other hand, the CCs need to be given incentives to commit to a CP and reserve capacity in advance, for instance by deducting reservation fees from the total consumption costs of the client. Moreover, CPs may grant additional discounts on reserved capacity, which however enables further exploitation possibilities for the clients.

Ideally, all parties trust each other to be honest and truthful. The CCs require precise and untampered performance data from their CP to derive accurate estimates of their computational demands. They also rely on the provider to scale capacity as needed to fully meet all reservations and not to *overbook* resources. The providers, on the other hand, have to face substantial investments to increase their resources based on the customers' reservation. They therefore require truthful resource reservations to cover those expenses. Clients should not *overstate* their true demands and speculate that an abundance of resources will lead to deteriorating prices later.

We consider resource reservation as a game between CP and CC. Both parties seek to play a strategy in that game, which maximizes their respective utility, that is maximal performed computations for minimal costs. Choosing the amount of capacity to reserve, given a demand forecast, constitutes the strategic decision of the CC. In a truthful mechanism, CCs draw maximal utility from the induced game by reserving their truly predicted demand. This furthermore avoids futile costs for the CP and hence benefits all parties involved. We therefore pose the research question:

Is there a truthful mechanism *for advance resource reservation in cloud computing, i.e. where reserving the truly predicted demand is the* dominant *(best) strategy for the client?*

We approach this question from both the CPs' and the CCs' side. First, we examine current research in the field of capacity reservation contracts [13] and deliberate its applicability to the cloud computing market. Second, we review literature on procurement strategies involving long term contracts [10] and scrutinize its suitability for cloud computing. We then design a truthful mechanisms for capacity reservation.

The contributions of this paper are:

i) review of literature on capacity reservation contracts as well as long-term procurement strategies and application of those works to cloud computing,
ii) extension of Inderfurth and Kelle's work [10] to deductible reservation costs as considered in [13],
iii) proof of truthful reservation behavior with exact demand forecasting capability,
iv) proof of profitable deviation from truth-telling in case of stochastic forecasting.

The rest of this paper is organized as follows. Section II gives a brief overview of relevant game-theoretic background. Sections III and IV approach capacity reservation from the CP's and the CC's side, respectively. We then proof the existence of a dominant truthful strategy in deterministic forecasting models and its nonexistence in stochastic ones in Section V. The paper is concluded in Section VI with a summary of our findings and an outlook on future work.

978-1-936968-63-3 © 2012 ICST

DOI 10.4108/icst.valuetools.2012.250173

II. Game-theoretic Background

This section provides basic game-theoretic definitions and theorems we make use of in the following sections.

In general we consider a game of n players $i \in \mathbf{I}$, each possessing a strategy set Σ_i. All players choose one strategy σ_i to play in the game. We denote $\sigma = (\sigma_1, \ldots, \sigma_n) \in \mathbf{\Sigma} = \mathbf{\Sigma}_1 \times \cdots \times \mathbf{\Sigma}_n$ the strategy configuration chosen by the players. When we examine the options for one player i, we often assume that the part of the strategy configuration *not* depending on i, namely $(\sigma_1, \ldots \sigma_{i-1}, \sigma_{i+1}, \ldots \sigma_n) =: \sigma_{-i}$, remains unchanged. Therefore we conveniently denote a strategy configuration σ as (σ_{-i}, σ_i). We similarly apply the $\cdot_{-i}$ operator to all vectors and sets we encounter. The goal of every player is to maximize the utility $u_i(\sigma_{-i}, \sigma_i)$ they draw from the game.

Definition 1 (Dominant Strategy [2]) *A strategy $\sigma_i^* \in \mathbf{\Sigma}_i$ is a dominant strategy for player i iff for all other strategies $\sigma_i \in \mathbf{\Sigma}_i \setminus \{\sigma_i^*\}$ and for all $\sigma_{-i} \in \mathbf{\Sigma}_{-i}$*

$$u_i(\sigma_{-i}, \sigma_i^*) \geq u_i(\sigma_{-i}, \sigma_i)$$

with at least one equality being strict.

Playing the dominant strategy is always a prudent choice. As we assume all players to be *rational*, deviating from σ_i^* is not an option. Unfortunately dominant strategies must not always exist, we therefore have to resort to the following definition of a stable strategy configuration, whose existence is guaranteed if mixed strategies are allowed. A mixed strategy $\mathbf{s}_i$ is a probability distribution over the *finite* set of pure strategies $\mathbf{\Sigma}_i$, i.e. $\mathbf{s}_i = (s_1, \ldots, s_{|\mathbf{\Sigma}_i|})$ with $\sum_{1 \leq k \leq |\mathbf{\Sigma}_i|} s_k = 1$. Hence, $u_i(\mathbf{s}_{-i}, \mathbf{s}_i)$ gives the expected utility.

Definition 2 (Nash Equilibrium [2]) *A strategy configuration σ^* is a Nash equilibrium iff for all players i for all their strategies $\sigma_i \in \mathbf{\Sigma}_i \setminus \{\sigma_i^*\}$*

$$u_i(\sigma_{-i}^*, \sigma_i^*) \geq u_i(\sigma_{-i}^*, \sigma_i).$$

A *dominant* strategy yields optimal utility regardless of the strategy choice of the other players. A *Nash* equilibrium strategy yields only optimal utility if all other players continue to play their equilibrium strategy, i.e. player i cannot increase their utility by being the sole player to deviate from σ^*.

The objective of our work is to design a *truthful mechanism* for reserving resources in cloud computing. We therefore define both termini technici.

Definition 3 (Mechanism [19]) *Given a set of alternatives $\mathbf{A}$ and for each player $i \in \mathbf{I}$, $|\mathbf{I}| = n$, a valuation function $v_i : \mathbf{A} \to \mathbb{R}$. A mechanism is a social choice function $f : \mathbb{R}^n \to \mathbf{A}$ and a vector of payment functions $p_1, \ldots, p_n$, where player i pays $p_i : \mathbb{R}^n \to \mathbb{R}$.*

The strategy set $\mathbf{\Sigma}_i$ of each player is therefore $\mathbb{R}$, the valuation they report to the mechanism. Specifically, their valuation $v_i(a)$ for all alternatives $a \in \mathbf{A}$ is in $\mathbf{\Sigma}_i$. A widely used mechanism is an auction. The set of alternatives $\mathbf{A}$ equals $\mathbf{I}$; $x \in \mathbf{A}$ denotes that player $x \in \mathbf{I}$ won the auctioned item. For all i in $\mathbf{I}$, $v_i(i)$ equals the player's valuation of the good and $v_i(j)$ is 0, $\forall j \in \mathbf{I} \setminus \{i\}$. Each player i reports a bid b_i to the mechanism. The social choice function $f = \operatorname{argmax}_{i \in \mathbf{I}} b_i$ allocates the item to the player with the highest bid, say $w \in \mathbf{I}$. For all $i \in \mathbf{I} \setminus \{w\}$, $p_i = 0$, whereas the payment of w is determined by some previously announced rule, for instance $p_w = b_w$ or $p_w = \max_{i \in \mathbf{I} \setminus \{w\}} b_i$, i.e. the second highest bid in a Vickrey auction. The players' utility is determined by $u_i = v_i(f(\mathbf{b})) - p_i$.

Definition 4 (Truthfulness [19]) *A mechanism $(f, p_1, \ldots, p_n)$ is truthful or incentive compatible iff $\forall i \in \mathbf{I}$ and $\forall a \in \mathbf{A}$ there exists* no *$x_i \in \mathbb{R}$ such that*

$$v_i(f(x_i, \mathbf{v}_{-i}(a))) - p_i(x_i, \mathbf{v}_{-i}(a)) > v_i(f(v_i(a), \mathbf{v}_{-i}(a))) - p_i(v_i(a), \mathbf{v}_{-i}(a)).$$

Truthfulness gives each player an incentive to provide their true valuation $v_i(a)$ rather than a lied one x_i to the mechanism, since truth-telling yields the highest possible utility. For instance, a second price auction is an incentive compatible mechanism [19].

A mechanism could be an arbitrary complex procedure, however the following result proves that we can restrict ourselves to examine *direct* mechanisms. In a *direct* mechanism each player secretly submits their valuation v_i in a single shot game, as opposed to some scheme requiring repeated actions by the players.

Theorem 1 (Revelation Principle [19, 9]) *If there is an* arbitrary *mechanism that implements f in dominant strategies, then there exists an equivalent* direct *and* truthful *mechanism for f. The payments obtained in the truthful scheme are identical to those of the original scheme at equilibrium.*

The concepts introduced in Definitions 2 - 4 and Theorem 1 form the basis for our search of a truthful resource reservation scheme. In order to follow the discussion of the paper by Li et al. [13] in Section III, we need to present the definition of a particular game studied in game theory.

Definition 5 (Stackelberg Game [2, 21]) *A Stackelberg game is a sequential game with one* leader *and one or more* follower(s)*. The leader must know* ex ante *that the follower observes the leader's actions and has no means to commit to an action before the leader's move. Consequently the leader, assuming* rationality *on the follower's side, can anticipate the best response of the follower to the leader's action. By backward induction the leader calculates their best action based on that knowledge.*

When both players play their best response to the other players' actions, a Nash equilibrium is induced in the Stackelberg game. The follower may threaten to deviate from the equilibrium strategy. However, this would not only hurt the leader but the follower as well. Hence, deviation is irrational and any threat by the follower can be deemed non-credible by the leader.

III. Cloud Provider View

In this section we examine choices to be made by the CP. Two major issues to be addressed by every CP are pricing and expansion policies. Before we study the work of Li et al. [13] on sharing the risk of capacity expansion between provider and customer through costly reservations, we will first briefly turn to the question whether the CP achieves optimal profits by providing plentiful resources or by keeping capacity intentionally scarce.

This issue is addressed by Harris et al. [9] who examine different pricing schemes for situations of abundant and scarce resources on the provider's side. They come to the conclusion that if demand exceeds the (exogenously) given capacity the provider achieves optimal profits by offering a priority pricing scheme, which entitles customers who are willing to pay more for the same good with a higher probability of being serviced. They furthermore show, that the same expected (optimal) profit can be achieved by conducting a modified Vickrey auction for the good. This positive result gives credence to the already widely used auctions in cloud computing [14]. If the (exogenously) given capacity meets or exceeds the expected demand, all customers will choose the lowest offered price. Consequently, Harris et al. [9] show, that the provider achieves indeed optimal expected profit by offering a single price identical with the lowest priority price in the scarce resource scenario. On the same note, Menache et al. [17] show that a single price suffices to induce a socially optimal resource allocation in cloud computing. Harris et al. [9] then turn to the question whether it is profitable to keep resources intentionally scarce if capacity can be chosen endogenously. They proof the contrary as the gain in expected income from the priority pricing scheme is outweighed by the loss due to unmet demand. For details of the proofs please refer to their paper [9].

Even given the optimality of offering plentiful capacity, a CP might be reluctant to proactively expand their capacity considering uncertainty in the actual demand on the cloud computing market. Since a capacity deficit would not merely diminish the provider's revenue but also the CCs' utility, sharing the financial risk of capacity expansion between the parties seems appropriate. Li et al. [13] propose the usage of deductible reservation contracts to share the risk. The clients reserve a certain amount of capacity upfront, paying some reservation price per unit. This payment is fully deductible when they utilize the resources after demand realization. Unused reserved resources however, are not refundable. The authors further propose that the provider announces an *excess* capacity level prior to the reservations that is provided in addition to the reserved amount. Li et al. [13] derive optimal strategies for both parties that depend on the relation of reservation price, unit price and the clients' valuation for the performed computations.

A. The Model

Li et al. [13] consider one seller S and one buyer B in a single period setting. The seller's decision variable is the amount of excess capacity E they are willing to offer regardless of the amount reserved by the client. The buyer on the other hand faces a stochastic demand D, adhering to some probability density function f_D. Considering the uncertain demand, B has to commit to a reservation amount R. The exact sequence of events is as follows:

1) S announces an excess capacity level E. The initial capacity is assumed to be zero.
2) Based on E and f_D, the buyer reserves R units of capacity, paying a fee of $r \cdot R$.
3) The seller expands the capacity level to $C = R + E$, incurring marginal cost c.
4) The demand D is realized and $\min(D, C)$ units are procured at a price of p per unit. The buyer has a valuation v for each consumed unit; unmet demand will be lost.
5) The amount of $r \cdot \min(D, R)$ is deducted from the buyer's purchasing cost by S. In case of excessive reservation, $r \cdot \max(R - D, 0)$ is kept by the seller.
6) Unused capacity is salvaged by S with a salvaging value of s per unit.

Li et al. [13] model the decision process as Stackelberg game, see Definition 5. The seller, as the leader, can anticipate the decision making process of the buyer, the follower in the game. Hence, S can optimize their announced excess capacity level E, knowing that B will choose the best response to E. By limiting the model to a single period setting, Li et al. [13] avoid the issue of repeated games, where deviating from the equilibrium strategy might be profitable. The authors propose the reasonable assumption that $v > p > c > s$. They assume the reservation price r to be exogenously given, with $r < p$.

Discussion: The model is apt for cloud computing markets. The exact computational needs are certainly hard to predict by the CCs, therefore modeling a stochastic demand is appropriate. Harris et al. [9] showed that providing more capacity than demand, $C > D$, cannot be optimal. However, they were considering a deterministic demand. By supplying an excess capacity, the CP shares the risk of demand uncertainty with the consumers, who otherwise would have to bear it themselves. If demand realizes below expectation the CP suffers from unused excess capacity *and* the customers from over-reservation. If CPs are not willing to commit to excess capacity, customers would be reluctant to reserve capacity upfront but rather resort to spot market procurement, i.e. buying computing power just in time. Salvaging can be applied twofold in cloud computing, either by a) selling the unnecessary hardware or b) offering the available capacity on the spot market. In both cases the assumption $p > c > s$ is appropriate: a) The hardware was just purchased by the provider at a rate of c per unit. Even though the hardware was not used to provide computational power to the clients, it has to be considered *used*, as it was installed in the CP's data center. Used equipment will generally be sold with a discount, thus $c > s$. We exclude the case of a scarce hardware market. b) Offering unused capacity on the spot market would suggest $s = p > c$. But for resources to be available for the spot market, the actual demand D has to be smaller than the projected demand $D + E$. Hence there is a

surplus of resources at price p and S has to lower prices in order to generate more demand. The assumption that prices have to be lowered below marginal cost is a limitation of the model of Li et al. [13]. ◊

B. Optimal Strategies

In this section we derive optimal strategies in the aforementioned reservation game for both parties. Firstly, we formalize the decision problem of the client and derive its optimal solution. Suppose the seller announced excess capacity level E. The client is seeking the optimal reservation amount R, such that the total capacity $C = E+R$ maximizes their utility.

$$\begin{aligned}\Pi_B &= \max_{R\geq 0}\{E[\overbrace{(v-p)\cdot\min(D,C)}^{\text{client's utility}} - \overbrace{r\cdot R + r\cdot\min(D,R)}^{\text{effective reservation costs}}]\} \qquad (1)\\ &= \max_{R\geq 0}\{(v-p)\cdot(C-E[C-D]) - r\cdot E[R-D]\}\\ &= \max_{R\geq 0}\left\{(v-p)\cdot\left(C-\int_0^C (C-x)f_D(x)dx\right)\right.\\ &\quad \left. -r\cdot\int_0^R (R-x)f_D(x)dx\right\}. \qquad (2)\end{aligned}$$

LEMMA 1 ([13]) *Given an excess capacity level E, there exists a unique solution $\hat{R}$ to the buyer's decision problem Π_B given by*

$$r\cdot\overbrace{F_D(\hat{R})}^{Pr[D\leq\hat{R}]} = (v-p)\cdot\overbrace{(1-F_D(\hat{R}+E))}^{Pr[D>\hat{R}+E]}.$$

Li et al. [13] emphasize that the optimal reservation amount $\hat{R}$ is monotonically decreasing in E. The more excess capacity the seller is willing to offer, the more risk of uncertain demand they have to bear. The authors proof both claims by inspecting the first and second order derivatives of (2) and Lemma 1 [13].

Using the implicit function $\hat{R}(E)$ given by Lemma 1, the decision problem of the seller can now be formalized in order to derive the optimal amount of available capacity $C = E + \hat{R}(E)$.

$$\begin{aligned}\Pi_S &= \max_{E\geq 0}\{E[\overbrace{p\cdot\min(D,C)}^{\text{revenue}} + \overbrace{s\cdot\max(0,C-D)}^{\text{salvaging}}\\ &\quad + \overbrace{r\cdot\hat{R}(E) - r\cdot\min(D,\hat{R}(E))}^{\text{gain from reservation}} - \overbrace{c\cdot C}^{\text{expansion costs}}]\} \qquad (3)\\ &= \max_{E\geq 0}\{\overbrace{(p-c)\cdot C}^{\text{opt. revenue}} - \overbrace{(p-s)\cdot E[C-D]}^{\text{loss due to overcapacity}} + \overbrace{r\cdot E[\hat{R}(E)-D]}^{\text{gain from reservation}}\}\\ &= \max_{E\geq 0}\left\{(p-c)\cdot C - (p-s)\cdot\int_0^C (C-x)f_D(x)dx\right.\\ &\quad \left. +r\cdot\int_0^{\hat{R}(E)} (\hat{R}(E)-x)f_D(x)dx\right\}. \qquad (4)\end{aligned}$$

Henceforth, Li et al. [13] assume f_D to be the uniform distribution on the interval $[0,\lambda]$ to derive explicit expressions in the further analysis. They admit that this constitutes a simplification of reality but claim that it is sufficient to gain insight into the main features of capacity reservation contracts and that similar expressions can be derived for other distributions.

With the uniformity assumption the implicit function $\hat{R}(E)$ resolves to

$$\begin{aligned}r\cdot F_D(\hat{R}) &= (v-p)\cdot(1-F_D(\hat{R}+E))\\ r\cdot\frac{\hat{R}}{\lambda} &= (v-p)\cdot\left(1-\frac{\hat{R}+E}{\lambda}\right)\\ \hat{R} &= \frac{(v-p)(\lambda-E)}{v-p+r}. \qquad (5)\end{aligned}$$

With the uniformity assumption and limiting E to $[0,\lambda]$, the decision problem (4) can be simplified to

$$\Pi_S = \max_{0\leq E\leq\lambda}\left\{(p-c)\cdot C - (p-s)\cdot\frac{C^2}{2\lambda} + r\cdot\frac{\hat{R}^2}{2\lambda}\right\}. \qquad (6)$$

Discussion: Offering more excess capacity as there could possibly be demand is certainly no prudent choice; the upper bound λ is therefore valid in this model. Limiting E to values ≥ 0 in fact contradicts the common practice of *overbooking* in the current cloud computing market [7]. Nevertheless, as Harris et al. [9] proved, this practice yields inferior expected profits and can thus be safely excluded from an optimal strategy. ◊

In order to simplify the further discussion, we introduce the following parameters [13]:

$$\alpha_1 = \frac{v}{2v-s} \qquad \alpha_2 = \frac{c-s}{2v-s}$$
$$\theta_1 = \frac{v-p}{p-s} \qquad \theta_2 = \frac{v-p+c-s}{p-c}$$

Observing that Π_S is convex for $0\leq p\leq\alpha_1 v$ and $r\in[0,p]$ as well as for $p>\alpha_1 p$ and $r\in[0,\theta_1(v-p)]$ and concave otherwise [13, Lemma 2], leads to the following theorem for the optimal reservation and excess capacity amounts:

THEOREM 2 (OPTIMAL RESERVATION AND EXCESS CAPACITY AMOUNTS [13])
A) If $0\leq p\leq(\alpha_1+\alpha_2)v$ then $\hat{E}=0$ and $\hat{R}=\lambda\cdot\frac{(v-p)}{v-p+r}$.
B) If $p>(\alpha_1+\alpha_2)v$ then

$$\hat{E} = \begin{cases}0 & r\in[0,\theta_2(v-p)]\\ \lambda\cdot\frac{(p-c)[r-\theta_2(v-p)]}{(p-s)[r-\theta_1(v-p)]} & r\in[\theta_2(v-p),p]\end{cases}$$

and

$$\hat{R} = \begin{cases}\lambda\cdot\frac{(v-p)}{v-p+r} & r\in[0,\theta_2(v-p)]\\ \lambda\cdot\frac{(v-p)\cdot(c-s)}{r\cdot(p-s)-(v-p)^2} & r\in[\theta_2(v-p),p]\end{cases}$$

For an elaborated proof of the theorem refer to [13]. We will merely discuss the underlying intuition. If the price p of the good does not exceed a $(\alpha_1+\alpha_2)v$ portion of the buyer's valuation, the penalty due to lost utility for the buyer in case of capacity shortages outweighs the risk of reserving too much. Therefore the seller can be certain that the reserved capacity suffices to meet projected demand; most likely even with some

safety margin. Thus, offering an excess capacity level $\hat{E} > 0$ is redundant. The same strategy for the seller applies if the purchase price p exceeds $(\alpha_1 + \alpha)v$ but the reservation price r is below $(v - p)\theta_2$. In this case the capacity shortage penalty for the buyer is not as severe as in the former mentioned setting, but the reservation price is sufficiently low in order for the client to be willing to undergo more financial risk to ensure ample resources. Only if the reservation price exceeds $(v - p)\theta_2$, the seller may anticipate the buyer's reluctance to reserve plentiful capacity and engage in a proactive expansion policy with $\hat{E} > 0$.

C. Choosing the Optimal Reservation Price

Hitherto the discussion assumed an exogenously given reservation price r. We will now examine the effects of the reservation price on the parties' optimal strategies as to give some insight into the seller's decision problem of choosing the optimal reservation price [13].

Firstly, we consider the case of $p < (\alpha_1 + \alpha_2)v$. Following Theorem 2, $\hat{E} = 0$; it can be easily seen that $\lim_{r \to 0} \hat{R}(r) = \lambda$ and $\hat{R}(r)$ is decreasing in r. Li et al. [13] introduce a further parameter

$$\theta_3 = \frac{v - p + 2c - 2s}{v + p - 2c}.$$

If the unit price p is below $\theta_3(v - p)$ then the seller's expected profit $\Pi_S(r)$ is increasing for all $r \in [0, p]$. For a unit price exceeding the threshold the optimal reservation price is

$$\hat{r} = \theta_3(v - p). \tag{7}$$

Secondly, we examine purchasing prices exceeding $(\alpha_1 + \alpha_2)v$. When r is chosen to be below $\theta_2(v - p)$, $\hat{E}$ equals 0 and the same argument as above holds, including the optimal reservation price of (7).[1] However, if r exceeds $\theta_2(v - p)$ the buyer's reservation is still decreasing in r but the seller's excess capacity $\hat{E}$ increases at a higher rate in r. Hence, $\hat{E} + \hat{R}$ increases in r as S has to ensure sufficient levels of capacity to meet future demand. Nevertheless, this reduces their expected profit, since the seller shares more risk of the uncertainty in the demand distribution. This makes an appealing case *against* high reservation prices.

Discussion: The results of [13] are very promising for developing a truthful resource reservation scheme in cloud computing. Given the assumptions of the paper, it would be optimal for a cloud provider to choose a reservation price of $\theta_3(v - p)$ (7), thus rendering the excess capacity level $\hat{E} = 0$ optimal. Hence, all risk of uncertain future demand is borne by the client. However, choosing $\hat{R}$ according to Theorem 2 yields optimal expected profit for the client as well, therefore it would be prudent to act according to the strategy set forth in Section III-B.

We will now dissect the assumptions of the paper by Li et al. [13]. The basic decision problems Π_B and Π_S are stated for general demand distributions by Li et al. [13], yet their detailed analysis presumes a uniform distribution in some interval $[0, \lambda]$. The authors' claim, that a similar analysis can be performed for other distributions as well, needs to be verified for probability distributions more suited for demand modeling, e.g. Gaussian, Poisson, Gamma or Pareto distribution [1, 6, 10].

A more fundamental problem arises from the assumption of a Stackelberg game. It presumes complete visibility by the CP into the consumer's decision making process. More specifically, the provider requires the customer's valuation v to calculate $\theta_{\{1,2,3\}}$, $\alpha_{\{1,2\}}$ and corresponding thresholds for p and r. This visibility is generally not given and customers might be reluctant to share this information [12, 20], especially if one CP serves clients competing in the same real-world market. Therefore reporting a valuation to the CP constitutes the strategic decision that all clients are facing. Given that for a reported valuation v values for p, $\hat{r}$, $\hat{E}$ and $\hat{R}$ may be optimally derived, it remains to be examined whether reporting the truthful valuation to the CP is the dominant strategy. Reporting an inferior valuation $v' < v$ for instance, could lead to more excess capacity E on the seller's side, thus shifting risk away from the buyer. We will revisit this issue in Section V.

Even if reporting the true valuation, from a game-theoretic viewpoint, is the dominant strategy, a CC might choose not to report it to the mechanism, because they don't want to reveal the value to the CP [12, 20]. However, this can be avoided by utilizing Secure Computation (SC) [18]. SC allows two or more parties to jointly compute a function $y = f(x_{CC}, x_{CP})$ on their combined inputs, while assuring that each party only learns the result y but *not* the input of the other party. Therefore the optimal parameters of the expansion strategy can be computed with privacy of the inputs protected by SC and correctness of the provided values guaranteed by the dominant strategy. ◊

IV. Cloud Consumer View

We now turn to decisions to be made by the CC. In general, two options present itself to a company to obtain resources: long term contracts and short term spot market procurement. Inderfurth and Kelle [10] prove that the mixture of both promises great cost savings potential compared with single sourced approaches. As computing power can be readily acquired from a CP by the clients [4], they might be reluctant to engage in a long term commitment. However, long term contracts promise guaranteed resources and stable prices, whereas spot market prices are rather volatile [16]. Inderfurth and Kelle [10] consider capacity reservation contracts as *real options*. The buyer acquires the right to purchase a certain quantity of goods for a specified price in the future but is not required to exercise this right. The mixed strategy proposed by the authors allows the client to leverage low spot market price, while still benefiting from the security provided by the reservation contract.

[1] Note that $\theta_3 < \theta_2$.

A. The Model

A buyer B wishes to combine long term capacity reservation contracts and spot market procurement to reduce their total purchasing costs. Inderfurth and Kelle [10] examine a *combined capacity reservation - base stock* policy. The buyer sets a long term capacity reservation level of R units. In each period this reservation incurs $r \cdot R$ reservation costs and entitles B to purchase up to R units at price p_c per unit. The client is further characterized by an inventory holding cost h per unit, an inventory level I_t at the beginning of each period and a valuation of the good v per unit.[2] Furthermore, B experiences a stochastic demand $\mathcal{D}$, adhering to some probability density function $f_{\mathcal{D}}$, and stochastic spot market prices $\mathcal{P}$ with density function $f_{\mathcal{P}}$. In each period t, the spot market price $\mathcal{P}$ is realized, denoted as $p_{s,t}$. B then needs to decide how many goods are to be ordered from the long term source $O_c \leq R$ and how many are to be acquired on the spot market O_s. Thereafter the demand $\mathcal{D}$ is realized, denoted by D_t, and the total period costs are computed. Since Inderfurth and Kelle [10] consider a base stock policy, a base stock level S has to be determined, up to which stock is replenished in every period. The authors consider a backorder situation, therefore total order quantity of period t, $O_{c,t} + O_{s,t}$, equals demand of period $t-1$ D_{t-1}.

Discussion: At first glance, the model of Inderfurth and Kelle [10] does not seem to be particularly appropriate for the cloud computing market. Nevertheless, studying their results still reveals some insight into mixed procurement strategies that can be adopted by CCs. The authors assume a storable good which does not apply to computing resources. However, as we will see in the next section, the optimal reservation level is independent of the storability of the good. Furthermore, even though computing power is not backorderable, the results of Inderfurth and Kelle [10] may still apply as they rely on the expectation and variability of demand and spot market price which are equal in all periods.[3] ◊

B. The Mixed Procurement Strategy

The strategy proposed by Inderfurth and Kelle [10] is straightforward. If the observed spot market price $p_{s,t}$ is below the contracted price p_c solely the spot market is used for procurement. When $p_{s,t} > p_c$ up to R units are purchased from the long term supplier and if further goods are required to replenish up to level S the spot market is used.

$$O_{c,t} = \begin{cases} 0 \\ \min(S - I_t, R) \end{cases}$$

$$O_{s,t} = \begin{cases} S - I_t & \text{if } p_{s,t} < p_c \\ \max(S - I_t - R, 0) & \text{if } p_{s,t} \geq p_c. \end{cases}$$

[2]For Inderfurth and Kelle [10], v constitutes the shortage cost per unit. We presume that the penalty for one unit of lost good equals the utility from one unit of consumed good, therefore justifying the assumption of shortage cost and valuation being identical.

[3]$\mathcal{D}$ and $\mathcal{P}$ are assumed to be independent and identically distributed (i.i.d.).

The total amount purchased in period t equals the demand D_{t-1} in the backorder situation. Prior to realization of demand it can therefore be considered a random variable; the spot market price likewise. To derive the expected purchase costs we consider $p_{s,t} < p_c$ and $p_{s,t} \geq p_c$ separately. In order to ease the proof of the subsequent Corollary 1, we introduce the parameter ϱ_c, denoting the tipping point between spot market and long term source procurement. For now, $\varrho_c = p_c$.
$p_{s,t} < \varrho_c$:

$$\begin{aligned} P_t^{<}(R) &= p_c \cdot E[\overbrace{O_{c,t}}^{0} | p_{s,t} < \varrho_c] \\ &\quad + E[\mathcal{P} | p_{s,t} < \varrho_c] \cdot E[\overbrace{O_{s,t}}^{E[\mathcal{D}]} | p_{s,t} < \varrho_c] \\ &= \int_0^{\varrho_c} x f_{\mathcal{P}}(x) dx \cdot \int_0^{\infty} x f_{\mathcal{D}}(x) dx. \end{aligned}$$

$p_{s,t} \geq \varrho_c$:

$$\begin{aligned} P_t^{\geq}(R) &= p_c \cdot E[\overbrace{O_{c,t}}^{E[\mathcal{D}|D_t \leq R] + R \cdot Pr[D_t > R]} | p_{s,t} \geq \varrho_c] + E[\mathcal{P} | p_{s,t} \geq \varrho_c] \\ &\quad \cdot E[\overbrace{O_{s,t}}^{E[\mathcal{D}-R|D_t \geq R]} | p_{s,t} \geq \varrho_c] \\ &= p_c(1 - F_{\mathcal{P}}(\varrho_c)) \cdot \left(\int_0^R x f_{\mathcal{D}}(x) dx + R(1 - F_{\mathcal{D}}(R)) \right) \\ &\quad + \int_{\varrho_c}^{\infty} x f_{\mathcal{P}}(x) dx \cdot \int_R^{\infty} (x - R) f_{\mathcal{D}}(x) dx. \end{aligned}$$

The expected purchasing costs $P_t(R) = P_t^{<}(R) + P_t^{\geq}(R)$ therefore merely depend on R. The expected inventory holding and shortage costs solely depend on S and are given by

$$L_t(S) = h \cdot \overbrace{\int_0^S (S - x) f_{\mathcal{D}}(x) dx}^{E[S - \mathcal{D} | D_t \leq S]} + v \cdot \overbrace{\int_S^{\infty} (x - S) f_{\mathcal{D}}(x) dx}^{E[\mathcal{D} - S | D_t > S]}.$$

Thus, the total costs of period t are

$$TC_t(R, S) = P_t(R) + L_t(S) + rR. \tag{8}$$

As the effects of the policy parameters R and S are separated, optimal parameters $\hat{R}$ and $\hat{S}$ can be obtained by setting the respective partial derivative of $TC_t(R, S)$ to zero, yielding the following theorem, proved in [10, Appendix A].

THEOREM 3 (OPTIMAL PARAMETERS FOR MIXED PROCUREMENT [10]) *Optimal reservation level $\hat{R}$ and base stock level $\hat{S}$ for the combined capacity reservation - base stock policy in a backorder situation are given by*

$$\hat{R} = F_{\mathcal{D}}^{-1} \left(\frac{\delta - r}{\delta} \right)$$

and

$$\hat{S} = F_{\mathcal{D}}^{-1} \left(\frac{v}{v + h} \right).$$

$F_{\mathcal{D}}^{-1}$ is the inverse of the cumulative distribution function of the demand and the conditional expected gain *δ, denoting the expected profit of having fixed price p_c when $p_{s,t} > \varrho_c$, with $\varrho_c = p_c$, is given by*

$$\delta = E[\mathcal{P} - p_c | p_{s,t} > \varrho_c] = \int_{\varrho_c}^{\infty} (x - p_c) f_{\mathcal{P}}(x) dx.$$

Discussion: The optimal reservation amount $\hat{R}$ depends only on the distribution of the demand, the reservation price and the expected advantage of the fixed price over the spot market price. It can hence also be applied if the storability of the considered good is *not* given, as in the case of cloud computing. Obviously, the optimal storage level $\hat{S}$ can be disregarded in that case. There are two approaches to address the non-storability of the good formally:

$h = 0$: The inventory holding cost of a non-storable good could be considered zero. Thus, $\frac{v}{v+h} = 1$ and $\hat{S}$ equals $\inf_{x \in \mathbb{R}}\{F_{\mathcal{D}} = 1\}$, either the upper bound of the projected demand or infinity.

$h \to \infty$: As inventory holding costs approach infinity, the term $\frac{v}{v+h} \to 0$. Consequently, $\hat{S}$ is zero.

The second approach is more suitable for our needs. The experienced demand is met by the buyer "on credit", resulting in negative stock, which is then replenished in the next period. Keeping in mind that the result by Inderfurth and Kelle [10] depends on the expectation of demand, this nearly yields the just in time procurement required in cloud computing.

Additionally observe that the authors assume *non*-deductible reservation costs in their model. In the following, we extend the model by introducing deductible reservation costs, as considered in Section III, and derive the optimal reservation amount in that scenario. Let $p_c = p'_c + r$. With deductible reservation costs r, it is beneficial for the buyer to turn to the spot market instead of the long term source for procurement iff $p_{s,t} + r < p'_c$. Therefore let the tipping point $\varrho_c = p'_c - r = p_c - 2r$. Furthermore, assume that orders in excess of R units are served at the spot market price $p_{s,t}$ and not p_c, even if fulfilled by the long term supplier. With these assumptions we modify (8) to

$$\begin{aligned} TC'_t(R,S) &= P_t(R) + L_t(S) \\ &\quad + \begin{cases} rR & \text{if } p_{s,t} < \varrho_c \\ r \cdot E[R - \mathcal{D} | D_t \leq R] & \text{if } p_{s,t} \geq \varrho_c \end{cases} \\ &= \cdots + rRF_{\mathcal{P}}(\varrho_c) \\ &\quad + r(1 - F_{\mathcal{P}}(\varrho_c)) \cdot \int_0^R (R - x) f_{\mathcal{D}}(x) dx. \end{aligned} \tag{9}$$

COROLLARY 1 (OPTIMAL PARAMETERS FOR MIXED PROCUREMENT WITH DEDUCTIBLE RESERVATION COSTS) *If reservation costs can be deducted from the payment owed to the long term supplier the optimal reservation amount is given by*

$$\hat{R} = F_{\mathcal{D}}^{-1}\left(\frac{\delta - rF_{\mathcal{P}}(\varrho_c)}{\delta + r(1 - F_{\mathcal{P}}(\varrho_c))}\right),$$

with $\varrho_c = p_c - 2r$. The optimal base stock level $\hat{S}$ and conditional expected gain δ remain as given by Theorem 3.

Proof: By taking the partial derivative of $TC'_t(R,S)$ with respect to R we obtain

$$\begin{aligned} \frac{\partial TC'_t(R,S)}{\partial R} &= (1 - F_{\mathcal{D}}(R)) \cdot \left(p_c(1 - F_{\mathcal{P}}(\varrho_c)) - \int_{\varrho_c}^{\infty} x f_{\mathcal{P}}(x) dx\right) \\ &\quad + rF_{\mathcal{P}}(\varrho_c) + r(1 - F_{\mathcal{P}}(\varrho_c))F_{\mathcal{D}}(R). \end{aligned}$$

Setting $\frac{\partial TC'_t(R,S)}{\partial R}$ to zero and rearranging yields

$$F_{\mathcal{D}}(R) = \frac{\int_{\varrho_c}^{\infty} x f_{\mathcal{P}}(x) dx - p_c(1 - \int_0^{\varrho_c} f_{\mathcal{P}}(x) dx) - rF_{\mathcal{P}}(\varrho_c)}{\int_{\varrho_c}^{\infty} x f_{\mathcal{P}}(x) dx - p_c(1 - \int_0^{\varrho_c} f_{\mathcal{P}}(x) dx) + r(1 - F_{\mathcal{P}}(\varrho_c))}.$$

Observing that

$$\begin{aligned} &\int_{\varrho_c}^{\infty} x f_{\mathcal{P}}(x) dx - p_c(1 - \int_0^{\varrho_c} f_{\mathcal{P}}(x) dx) \\ &= \int_{\varrho_c}^{\infty} x f_{\mathcal{P}}(x) dx - p_c \int_{\varrho_c}^{\infty} f_{\mathcal{P}}(x) dx \end{aligned}$$

equals δ and taking the inverse $F_{\mathcal{D}}^{-1}(\cdot)$ proves the corollary. ■

$F_{\mathcal{P}}(\varrho_c)$ can be considered the probability of net reservation costs, i.e. ordering fewer products from the long term source than reserved, while still having to pay reservation fees for them. If reservation costs are *not* deductible these fees always have to be paid, thus $F_{\mathcal{P}}(\varrho_c)$ can be set to 1 and Corollary 1 equals Theorem 3.

Given these two considerations, studying the further results of Inderfurth and Kelle [10] seems justified, as the gained insight can be adapted to cloud computing markets with little effort. ◊

C. Single Sourced Approaches

Inderfurth and Kelle [10] now compare the mixed procurement strategy to strategies comprising only one source of purchasing. Firstly, they examine a purely spot market-based procurement strategy. The total expected purchasing costs are given by

$$P_{\text{spot}} = E[\mathcal{P}] \cdot E[\mathcal{D}].$$

The inventory and shortage costs remain the same, thus the optimal base stock level $\hat{S}$ is still given by Theorem 3. The cost difference CD_{spot} therefore depends only on purchasing costs

$$CD_{\text{spot}}(R,s) = P_{\text{spot}} - (P(R,S) + rR). \tag{10}$$

Again, the authors assume non-deductible reservation prices, but the following observations still apply if (10) is modified similar to (9) in order to represent deductible reservation costs. The advantage of combined sourcing over spot market procurement *increases* with spot market price variability and *decreases* with demand variability, reservation price and fixed price. Inderfurth and Kelle [10] further note that even in a *superior spot market price situation*, i.e. $E[\mathcal{P}] < P(R,S) + rR$, combined sourcing is beneficial if the spot market price variability $\frac{\sigma_{\mathcal{P}}}{\mu_{\mathcal{P}}}$ is sufficiently high to yield $r < \delta$, consequently $\hat{R} > 0$ in Theorem 3. The authors conduct extensive numerical analysis to substantiate their observations [10, Table 2,3].

Analyzing strategies based purely on long-term contracts proves more difficult. The stock of the buyer cannot always be replenished up to level S, namely if $D_t > R$. Therefore, base stock level S and capacity reservation level R depend on each other and simple closed form formulae cannot be found. Inderfurth and Kelle [10, Appendix C] provide a steady-state distribution of the expected inventory level and derive respective costs from it. However, their main observations again rely on numerical analysis [10, Table 4]. First they note that optimal reservation and base stock levels are slightly lower for combined sourcing strategies than for long-term-based sourcing. This effect increases with higher demand and spot market price variability thus leveraging the flexibility provided by the spot market option of combined sourcing. In general the advantage of combined sourcing *increases* with higher spot market price variability and *decreases* with reservation and fixed price. The effect of demand variability depends on the shape of the distribution and cannot be cast into a simple observation. In general, the combined sourcing strategy proves to be a prudent choice even in the case of a *inferior spot market situation*, i.e. $E[\mathcal{P}] > P(R,S) + rR$, due to the prevention of stock shortages and profits from occasional low spot market prices.

Discussion: The deliberations of Inderfurth and Kelle [10] provide compelling arguments for CCs to engage in long term reservation contracts. The positive effects of cost savings and guaranteed availability of resources should diminish the reluctance to commit to one CP. This is especially true if we lift the assumption of plentiful resources on the spot market. The authors note furthermore that capacity reservation contracts can be applied for the procurement of utilities such as electricity [11, 22]. As cloud computing aims at providing computing as utility [4], the arguments are hence very appropriate. Prior game-theoretic works on the field of utility pricing is less applicable to cloud computing. For instance Littlechild [15] considers fixed charges for telephone networks and assumes the problem to be a cooperative game. The players share the total costs of providing telephone services among each other proportionate to the benefit they draw from having access to the network. CCs competing for computational resources in the cloud computing market are oftentimes competitors in other markets as well, thus rendering the cooperative game assumption dubious. ◊

V. Reservation Mechanisms

As seen in the previous section, reserving capacity in advance yields benefits over sole spot market procurement. Hitherto, we considered an *exogenously* determined demand, described by some probability distribution. We will now study the case when demand is determined *endogenously* by some forecasting method. The CC require precise information about their current resource consumption in order to predict future computational demands. We assume that this information is provided by the CP accurately and *truthfully*. We consider perfect and stochastic forecasting abilities of the CC and examine reservation strategies in both cases.

A. The Model

We consider one CP P offering computing services to one buyer B. P offers spot market computing services for some price p per unit of computation. The customer however may reserve a specified quantity of computation r, being charged some price $(c_r \cdot p) \cdot r$ for the reservation, which entitles B to perform up to r units of computation for a discounted price $(c_d \cdot p)$ per unit. The buyer has a computational demand of d_t units in period t. We furthermore assume $c_r + c_d \leq 1$. Therefore, given reservation r and actual consumption $s_t = \min(d_t, C_t)$, B has to pay

$$p_t(r, s_t) = (c_r p) \cdot r + (c_d p) \cdot \min(r, s_t) + p \cdot \max(0, s_t - r)$$

The utility of the consumer is defined as

$$u_B(r, d_t) = \alpha \min(d_t, C_t) - p_t(r, s_t).$$

Factor α determines the value of the computation to the consumer, we may safely assume $\alpha \geq p$ otherwise *not* performing the computations would always yield more utility. The total amount of computational resources offered by the CP in a given period is denoted by C_t. If the available resources are insufficient to meet all computational needs opportunity costs arise in form of lost utility.

The utility of the provider is defined as

$$u_P(r, d_t) = p_t(r, s_t) - \gamma C_t.$$

To provide the offered capacity, γC_t monetary units are required. Again, we may safely assume $\gamma \leq p$ otherwise it would advantageous for the CP to *not* provide any computing power. Furthermore, the opportunity costs for not providing enough computational power are given implicitly again, following an analog argument as in the buyer's case.

Discussion: Restricting the model to only one provider and consumer is a strong but common [10, 13] assumption of the model. However, assuming that both parties are endowed with an optimal strategy, there is no reason for any number of clients (or providers) to deviate from it. Hence, a group of CCs all acting according to the same strategy can be seen as one homogeneous client in strategic considerations. ◊

B. Perfect Forecasting

The CPs desire to learn the true computational needs of their clients in advance, i.e. $r = d_t$, for accurate capacity planning. They thus wish to set p, c_r and c_d in a manner promoting truth-telling as dominant strategy for all clients. We first present a result under strong (unrealistic) assumptions:

Lemma 2 *Given the ability of perfect demand forecasting, i.e.* $d_t = \mathrm{pred}(d_{t-1})$*, reserving the true computational need in advance,* $r = \mathrm{pred}(d_{t-1})$*, is the dominant strategy for all fixed* p *and* $c_r + c_d \leq 1$.

Proof: We firstly consider the payment of the buyer in the case of always sufficient resources, i.e. even if $r < d_t$

the actual consumption $s_t = d_t$. B is not worse off reserving $r = d_t$ instead of using the spot market

$$p(d_t, d_t) = \overbrace{(c_r + c_d)}^{\leq 1} pd_t \leq pd_t = p(0, d_t). \tag{11}$$

Reserving more capacity than anticipated, $r = d_t + \delta$, yields a payment of

$$p(d_t + \delta, d_t) = (c_r + c_d)pd_t + c_r p\delta > p(d_t, d_t); \tag{12}$$

reserving less capacity, $r = d_t - \delta$ results in a payment of

$$\begin{aligned} p(d_t - \delta, d_t) &= (c_r + c_d)p(d_t - \delta) + p\delta \\ &= (c_r + c_d)pd_t + p\delta \overbrace{(1 - \underbrace{(c_r + c_d)}_{\leq 1})}^{\geq 0} \geq p(d_t, d_t). \end{aligned} \tag{13}$$

Due to (12) being strict, reserving $r = d_t$ is the dominant strategy of B. We now consider the capacity offered by the provider. Knowing that $r = d_t = \text{pred}(d_{t-1})$ is the dominant strategy of B, P will offer $C_t = r$ units of capacity, as this yields optimal earnings (refer to Harris et al. [9]). Hence, if $r = d_t - \delta$ the buyer faces opportunity costs of $(\alpha - p)\delta > 0$. Thus, the benefit in (11) and (13) is strict when considering opportunity costs. ■

Note that $c_r + c_d = 1$ constitutes the case of deductible reservation costs as studied in previous sections.

Discussion: Considering the multitude of CPs in the market, a shortage of resources on the spot market appears unlikely. Hence, if no discount for advanced resource reservation is granted, i.e. $c_r + c_d = 1$, the CC are indifferent to reserving resources or spot market procurement in (11). Considering the general reluctance to early commitment, B would probably chose *not* to reserve any resources in advance. However, currently migration between different cloud providers is often cumbersome [23]. Thus, if the capacity of the CP of choice is exhausted there are no more *usable* resources on the spot market. Therefore, the CC have a strong interest in a plentiful capacity level of their CP. ◊

C. Stochastic Forecasting

We now consider a demand forecast $\mathcal{D}_t$ described by some probability density function $f_{\mathcal{D},t}$. Given the precise measurement of resource consumption d_{t-1} in period $t-1$, prediction function $\text{pred}(d_{t-1})$ conditions the general demand distribution $f_{\mathcal{D}}$ on $\mathcal{D}_{t-1} = d_{t-1}$ yielding $f_{\mathcal{D},t} = f_{\mathcal{D}}(x|\mathcal{D}_{t-1} = d_{t-1})$.

We will demonstrate that untruthful behavior of the CC can be beneficial for them by exaggerating their predicted demand. For the following lemma we assume a rational CP, thus $C_t = r$, as well as $f_{\mathcal{D},t}$ being an unbiased estimator, i.e. $E[\mathcal{D}_t|d_t] = d_t$.

LEMMA 3 *Given a stochastic demand prediction $\mathcal{D}_t$, adhering to probability density function $f_{\mathcal{D},t} = \text{pred}(d_{t-1})$, the buyer can beneficially deviate from truth-telling, $r = E[\mathcal{D}_t]$, by exaggerating their resource requirements, i.e. $r = E[\mathcal{D}_t] + \delta$.*

Proof: We proof the benefit of deviating from truth-telling, $r = \tilde{D} = E[\mathcal{D}_t]$, by analyzing the expected expenditures for reserving an extra δ units of capacity and possible savings obtained from it. Depending on the realization of $\mathcal{D}_t$, three cases need to be distinguished:

i) if $\mathcal{D}_t < \tilde{D}$ reserving $\tilde{D} + \delta$ yields no benefit but incurs extra costs of $c_r p\delta$;
ii) on the contrary, if $\mathcal{D}_t > \tilde{D} + \delta$ the full benefit of exaggerating the expected demand by δ is experienced, leading to cost savings of $(1 - (c_r + c_d))p\delta$;
iii) if $\mathcal{D}_t$ is in $[\tilde{D}, \tilde{D} + \delta]$ extra expenditures and cost savings break even at

$$c_r p\delta = (1 - c_d)p(\mathcal{D}'_t - \tilde{D}) \Longrightarrow \mathcal{D}'_t = \tilde{D} + \frac{c_r}{1 - c_d}\delta.$$

For deviating from $r = \tilde{D}$ to be beneficial, the expected savings must outweigh the expected additional expenditures

$$\begin{aligned} &c_r p\delta \cdot Pr[\mathcal{D}_t \leq \tilde{D}] + E[c_r p\delta - (1 - c_d)p(\mathcal{D}_t - \tilde{D})|\tilde{D} \leq \mathcal{D}_t \leq \mathcal{D}'_t] \\ &\leq \\ &E[(1 - c_d)p(\mathcal{D}_t - \tilde{D}) - c_r p\delta|\mathcal{D}'_t \leq \mathcal{D}_t \leq \tilde{D} + \delta] \\ &+(1 - (c_r + c_d))p\delta \cdot Pr[\mathcal{D}_t \geq \tilde{D} + \delta]. \end{aligned}$$

Algebraic manipulations reveal that if

$$\delta \leq \frac{(1 - c_d) \overbrace{\int_{\tilde{D}}^{\tilde{D}+\delta} (x - \tilde{D}) f_{\mathcal{D},t} dx}^{E[\mathcal{D}_t - \tilde{D}|\tilde{D} \leq \mathcal{D}_t \leq \tilde{D} + \delta]}}{c_r \underbrace{F_{\mathcal{D},t}(\tilde{D})}_{Pr[\mathcal{D}_t \leq \tilde{D}]} - (1 - (c_r + c_d)) \underbrace{(1 - F_{\mathcal{D},t}(\tilde{D} + \delta))}_{Pr[\mathcal{D}_t \geq \tilde{D} + \delta]}} \tag{14}$$

it is beneficial to reserve $r = E[\mathcal{D}_t] + \delta$ units of capacity instead of truthfully reserving $E[\mathcal{D}_t]$ units. ■

As can be seen from (14), a higher discount granted by the CP on resources reserved in advance, $c_r + c_d \ll 1$, allows for higher profitable deviations from the truly expected demand. Consequently, the CP incurs higher costs for providing the reserved, but partially unused, capacity which diminishes their utility. Nevertheless, the CP can promote truthful behavior, by requiring a larger portion of the price at reservation time, i.e. higher c_r values.

When (14) is applied to a demand uniformly distributed in $[0, \lambda]$, as studied in section III-B, the condition for δ simplifies to

$$\delta \leq \frac{2(1 - (c_r + c_d))\lambda - 2(1 - c_d)\tilde{D}}{1 - (2c_r + c_d)}.$$

Discussion: Lemma 3 demonstrates that designing a resource reservation scheme promoting truthful behavior on the clients' side becomes more challenging when using models closer to reality. Even though demand forecasting has been extensively studied by many researchers, using manifold settings, models and techniques [e.g. 3, 5], in practice, *perfect* forecasting is not possible. ◊

VI. Conclusion

Designing a capacity reservation mechanism in cloud computing that entails truthfulness of the clients is a complex task. Customers need to be compelled to reserve resources in advance, e.g. by granting discounts, while simultaneously being hindered from exploiting the system to their advantage.

In the first part of our paper we showed that it is indeed beneficial for customers and providers alike to participate in reservation mechanisms. To that end, we examined economic literature and applied it to the cloud computing market. We established that cloud providers achieve optimal expected earnings by providing sufficient capacity to meet future demand, instead of keeping resources scarce to artificially create higher prices [9]. Even confronted with this, a CP might still be reluctant to expand capacity as required, as demand uncertainty could render costly investments futile. Through capacity reservation contracts with deductible reservation costs, CP and CC share the risks of capacity expansion, with the higher portion of the risk being borne by the party which also profits the most from abundant resources [13]. Nevertheless, the cloud computing clients might be reluctant to commit early to a CP and rather procure their computing power on the spot market, leaving the providers with the entire expansion risk. However, a diverse strategy, utilizing long-term contracts as well as spot market procurement, proves to be beneficial to the clients [10]. Long-term purchasing, with fixed prices and guaranteed capacity, is augmented by short-term spot market procurement, exploiting low prices as they occur.

In the second part of our paper, we examined how clients might exploit a resource reservation mechanism. Firstly, we considered capacity reservation under perfect deterministic demand forecasting. Lemma 2 establishes truthful resource reservation as dominant strategy in that case. Secondly, we treated uncertain stochastic demand forecasting. In that case, the client can profitably deviate from truth-telling by exaggerating their predicted demand, as shown in Lemma 3. Our result provides a basic intuition for the CPs on how their pricing model affects truthful behavior. Furthermore, we extended the work of Inderfurth and Kelle [10], who consider a scenario with non-deductible reservation costs and provide a closed-form formula for the optimal capacity reservation amount in that model. Corollary 1 gives a closed-form formula for that quantity in a deductible reservation costs model.

As part of our future work, we will examine how a client can influence the expansion policy [13] of their CP by providing untrue valuations in order to bear less risk of the costly capacity expansion. We furthermore strive to derive closed-form formulae for all of the CP's parameters, so that they are enabled to set forth a pricing and reservation scheme with truth-telling as dominant strategy.

Further aspects that we will consider are: i) effects of unused capacity on future price development, ii) resource reservation as repeated game including punishments for deviating from the (truthful) equilibrium strategy, iii) malicious CPs, including overbooking of resources and providing tampered resource utilization data, therefore affecting the CC's demand forecasting.

References

[1] K. S. Azoury, "Bayes solution to dynamic inventory models under unknown demand distribution," *Management Science*, vol. 31, no. 9, pp. 1150–1160, 1985.

[2] E. N. Barron, *Game Theory - An Introduction.* Wiley, 2008.

[3] D. Bunn, "Forecasting loads and prices in competitive power markets," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 163–169, 2000.

[4] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities," in *HPCC 2008.* IEEE, Sep. 2008, pp. 5–13.

[5] E. Caron, F. Desprez, and A. Muresan, "Forecasting for Grid and Cloud Computing On-Demand Resources Based on Pattern Matching," in *CloudCom 2010.* IEEE, Nov. 2010, pp. 456–463.

[6] S. A. Conrad, "Sales data and the estimation of demand," *Operational Research Quarterly*, vol. 27, no. 1, pp. 123–127, 1976.

[7] D. Durkee, "Why cloud computing will never be free," *Communications of the ACM*, vol. 53, no. 5, p. 62, May 2010.

[8] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared," in *GCE 2008.* IEEE, Nov. 2008, pp. 1–10.

[9] M. Harris, A. Raviv, and B. M. Harris, "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *The American Economic Review*, vol. 71, no. 3, pp. 347–365, 1981.

[10] K. Inderfurth and P. Kelle, "Capacity reservation under spot market price uncertainty," *International Journal of Production Economics*, vol. 133, no. 1, pp. 272–279, Sep. 2011.

[11] P. R. Kleindorfer and D. Wu, "Integrating long- and short-term contracting via business-to-business exchanges for capital-intensive industries," *Management Science*, vol. 49, no. 11, pp. 1597–1615, 2003.

[12] H. L. Lee and S. Whang, "Information sharing in a supply chain," Graduate School of Business Stanford University, Tech. Rep. 1549, 2000.

[13] J. Li, M. Xu, and R. Yang, "Optimal capacity expansion policy with a deductible reservation contract," *Journal of Service Science and Management*, vol. 04, no. 01, pp. 27–34, 2011.

[14] W.-Y. Lin, G.-Y. Lin, and H.-Y. Wei, "Dynamic auction mechanism for cloud resource allocation," in *CCGrid 2010.* IEEE, 2010, pp. 591–592.

[15] S. Littlechild, "A Game-Theoretic Approach to Public Utility Pricing," *Economic Inquiry*, vol. 8, no. 2, pp. 162–166, Jun. 1970.

[16] M. Mattess, C. Vecchiola, and R. Buyya, "Managing peak loads by leasing cloud infrastructure services from a spot market," in *HPCC 2010.* IEEE, Sep 2010, pp. 180–188.

[17] I. Menache, A. Ozdaglar, and N. Shimkin, "Socially optimal pricing of cloud computing resources," in *ValueTools 2011.* ACM, 2011.

[18] S. Micali and P. Rogaway, "Secure computation," in *CRYPTO 1991*, ser. LNCS. Springer, 1992, vol. 576, pp. 392–404.

[19] N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, Eds., *Algorithmic Game Theory.* Cambridge University Press, 2007.

[20] T. M. Simatuapang and R. Sridharan, "The collaborative supply chain," *The International Journal of Logistics Management*, vol. 13, no. 2, pp. 15–30, 2002.

[21] H. von Stackelberg, *Market Structure and Equilibrium.* Springer, 1934.

[22] C.-K. Woo, I. Horowitz, A. Olson, B. Horii, and C. Baskette, "Efficient frontiers for electricity procurement by an ldc with multiple purchase options," *Omega*, vol. 34, no. 1, pp. 70–80, Jan 2006.

[23] L. Youseff, M. Butrico, and D. Da Silva, "Toward a unified ontology of cloud computing," in *GCE 2008.* IEEE, Nov 2008, pp. 1–10.