Appendix:Optimal Pricing and Resource Configuration for GPU-accelerated Services in Computing Markets

Zihe Wang[1], Shijia Huang[1], Qian Ma[1], and Xu Chen[2]
[1]*School of Intelligent Systems Engineering, Sun Yat-sen University*
[2]*School of Computer Science and Engineering, Sun Yat-sen University*

APPENDIX A
PROOF OF LEMMA 1

For each service $i \in \mathcal{S}$, type-i users make VCRB decisions to minimize their total cost $\Gamma_i(x_i;\phi^G)$. A type-i user's total cost on the CPU VCRB is

$$\Gamma_i(x_i=0;\phi_i^G) = p^C + \frac{l_i}{f^C}, \tag{A.1}$$

and that on the GPU VCRB is

$$\Gamma_i(x_i=1;\phi^G) = p^G + \frac{l_i}{f^G\theta_i}. \tag{A.2}$$

Comparing above two equations, we obtain that,

1) if $-(p^G-p^C)f^C+\lambda l_i<0$, $\Gamma_i(x_i=0;\phi_i^G)<\Gamma_i(x_i=1;\phi_i^G)$ always holds, hence the type-i user will choose the CPU VCRB.
2) if $-(p^G-p^C)f^C+\lambda l_i\geq 0$,
   a) if $\theta_i < \frac{\lambda l_i f^C}{[-(p^G-p^C)f^C+\lambda l_i]f^G}$, the type-$i$ user will choose the CPU VCRB since $\Gamma_i(x_i=0;\phi_i^G) < \Gamma_i(x_i=1;\phi_i^G)$,
   b) if if $\theta_i \geq \frac{\lambda l_i f^C}{[-(p^G-p^C)f^C+\lambda l_i]f^G}$, the type-$i$ user will choose the GPU VCRB since $\Gamma_i(x_i=0;\phi_i^G) \geq \Gamma_i(x_i=1;\phi_i^G)$.

We observe that the type-i user chooses the GPU VCRB only in case $(1b)$. Hence by substituting the conditions of case $(1b)$ into the definition of the discriminant function in (13), we obtain that a type-i user will choose the GPU VCRB if $\Delta(\phi^G;\mathcal{T}_i)$ satisfies

$$0 \leq \Delta(\phi^G;\mathcal{T}_i) \leq \theta_i. \tag{A.3}$$

This completes our proof.

APPENDIX B
PROOF OF THEOREM 1

We decompose the multiple GPU VCRB design problem $(\mathbf{P2})$ into S subproblems. For each $i\in\mathcal{S}$, subproblem i aims to maximize SP's profit from service i which is formulated as

$$(\mathbf{P2}'): \quad \max_{\phi_i^G} \quad H_i(\phi_i^G) \tag{B.1}$$
$$\text{s.t.} \quad p_i^G \geq p^C, \tag{B.2}$$
$$f_i^G \geq 0. \tag{B.3}$$

The objective function of the subproblem $(\mathbf{P2}')$ is discontinuous due to users' rational VCRB decisions, hence we solve the subproblem $(\mathbf{P2}')$ by dividing it into two cases, i.e., type-i users choose the GPU VCRB and the CPU VCRB, respectively. Our analysis goes as follows,

1) type-i users choose the CPU VCRB: in this case where $\Delta(\phi_i^G,\mathcal{T}_i) \in (-\infty,0)\cup(\theta_i,\infty)$, no GPU VCRB is demanded by type-i users, thus no GPU VCRB is deployed. Hence the optimal GPU VCRB design is $\phi_{i,1}^{G,*} = (p_{i,1}^{G,*}, f_{i,1}^{G,*}) = (0,0)$ and the profit of SP is $H_{i,1} = H_i(\phi_{i,1}^{G,*}) = N\omega_i p^C$.
2) type-i users choose the GPU VCRB: in this case where $\Delta(\phi_i^G,\mathcal{T}_i)\in[0,\theta_i]$, we rewrite the VCRB design problem for service i as follows,

$$\max_{\phi_i^G} \ H_i(\phi_i^G) = N\omega_i(p_i^G - p^D f_i^G) \tag{B.4}$$
$$\text{s.t.} \ \ p_i^G \geq p^C, \tag{B.5}$$
$$f_i^G \geq 0, \tag{B.6}$$
$$\Delta(\phi_i^G,\mathcal{T}_i)\in[0,\theta_i]. \tag{B.7}$$

Constraint (B.7) ensures type-i users' total costs on the GPU VCRB are lower than that on the CPU VCRB.

To solve the above problem, we first ignore constraints (B.5) and (B.6), and then rewrite constraint (B.7) as $[-(p_i^G-p^C)f^C+\lambda l_i]\theta_i f_i^G - \lambda f^C l_i \geq 0$, then we solve this VCRB design problem with Lagrange multiplier method.

Assume that $[-(p_i^G-p^C)f^C+\lambda l_i]\theta_i f_i^G - \lambda f^C l_i = a (a\geq 0)$, we define the Lagrange function as

$$\begin{aligned} L(\phi_i^G,\alpha) &= N\omega_i(p_i^G - p^D f_i^G) \\ &+ \alpha\left\{[-(p_i^G-p^C)f^C+\lambda l_i]\theta_i f_i^G - \lambda f^C l_i - a\right\}, \end{aligned} \tag{B.8}$$

where α is the Lagrange multiplier. Based on the Lagrange function (B.8), we have following partial differential equations,

$$\frac{\partial L}{\partial p_i^G} = N\omega_i - \alpha\theta_i f^C f_i^G = 0, \tag{B.9}$$
$$\frac{\partial L}{\partial f_i^G} = -N\omega_i p^D + [-(p_i^G-p^C)f^C+\lambda l_i]\alpha\theta_i = 0, \tag{B.10}$$

$$\frac{\partial L}{\partial \alpha} = [-(p_i^G - p^C)f^C + \lambda l_i]\theta_i f_i^G - \lambda f^C l_i - a = 0. \tag{B.11}$$

Solving above equations (B.9) - (B.11) gives us the optimal VCRB design under the condition $[-(p_i^G - p^C)f^C + \lambda l_i]\theta_i f_i^G - \lambda f^C l_i = a$ as follows,

$$p_i^{G,*,a} = -\sqrt{\frac{p^D(\lambda f^C l_i + a)}{\theta_i f^C}} + \frac{\lambda l_i}{f^C} + p^C, \tag{B.12}$$

$$f_i^{G,*,a} = \sqrt{\frac{\lambda f^C l_i + a}{\theta_i f^C p^D}}. \tag{B.13}$$

Then we substitute Eq. (B.12) and Eq. (B.13) into Eq. (B.4) to calculate the maximum profit under the condition $[-(p_i^G - p^C)f^C + \lambda l_i]\theta_i f_i^G - \lambda f^C l_i = a$ as

$$\begin{aligned} H_i^{*,a}(a) &= H_i(\phi_i^{G,*,a}) \\ &= N\left[p^C + \frac{\lambda l_i}{f^C} - 2\sqrt{\frac{p^D(\lambda f^C l_i + a)}{\theta_i f^C}}\right]. \end{aligned} \tag{B.14}$$

Eq. (B.14) shows that $H_i^{G,*,a}(a)$ increases as a decreases, which indicates that $H_i^{G,*,a}(a)$ reaches the maximum value when $a = 0$. Finally, we obtain the optimal solution to the problem (B.4) as follows,

$$p_{i,2}^{G,*} = -\sqrt{\frac{\lambda p^D l_i}{\theta_i}} + \frac{\lambda l_i}{f^C} + p^C, \tag{B.15}$$

$$f_{i,2}^{G,*} = \sqrt{\frac{\lambda l_i}{\theta_i p^D}}, \tag{B.16}$$

and the maximum profit is

$$H_{i,2}^* = H_i^{*,a}(0) = N\omega_i\left[p^C + \frac{\lambda l_i}{f^C} - 2\sqrt{\frac{\lambda p^D l_i}{\theta_i}}\right]. \tag{B.17}$$

Comparing SP's profit in above two cases $H_{i,1} = H_i(\phi_{i,1}^{G,*}) = N\omega_i p^C$ and $H_{i,2}^* = H_i(\phi_{i,2}^{G,*}) = N\omega_i\left[p^C + \frac{\lambda l_i}{f^C} - 2\sqrt{\frac{\lambda p^D l_i}{\theta_i}}\right]$, we conclude as follows. For each $i \in \mathcal{S}$, if type-i users with $\mathcal{T}_i = \{l_i, \theta_i\}$ satisfy,

1) $\frac{\lambda l_i}{f^C} - 2\sqrt{\frac{\lambda p^D l_i}{\theta_i}} < 0$, SP will not design a detailed GPU VCRB for service i, referred by $\phi_i^{G,*} = (p_{i,1}^{G,*}, f_{i,1}^{G,*}) = (0, 0)$, since SP will achieve a higher profit $H_{i,1}^* = N\omega_i p^C$ if type-i users choose the CPU VCRB,
2) $\frac{\lambda l_i}{f^C} - 2\sqrt{\frac{\lambda p^D l_i}{\theta_i}} \geq 0$, SP will design a detailed GPU VCRB $\phi_i^{G,*} = (p_{i,2}^{G,*}, f_{i,2}^{G,*}) = (-\sqrt{\frac{\lambda p^D l_i}{\theta_i}} + \frac{\lambda l_i}{f^C} + p^C, \sqrt{\frac{\lambda l_i}{\theta_i p^D}})$ for service i to achieve a higher profit $H_{i,2}^* = N\left[p^C + \frac{\lambda l_i}{f^C} - 2\sqrt{\frac{p^D(\lambda f^C l_i + a)}{\theta_i f^C}}\right]$.

This completes our proof.

Appendix C
Proof of Lemma 2

Given any service set $\mathcal{S} = \{1, 2, \ldots, S\}$ and SP's optimal uniform GPU VCRB design $\phi^{G,uni,*} = \{p^{G,uni,*}, f^{G,uni,*}\}$, we learn that $H^* \geq Np^C$ since Np^C is SP's original profit where it does not launch any GPU VCRB. Consider that a new service $k = s+1$ with a negative potential value is introduced into the service set $\mathcal{S}$. We denote the new service set as $\mathcal{S}'$.

1) If type-k users choose the CPU VCRB: Observe that under the optimal GPU VCRB design $\phi^{G,uni,*}$ for $\mathcal{S}$, service k will choose the CPU VCRB. (We argue this by contradiction. If type-k users choose the GPU VCRB $\phi^{G,uni,*}$, SP would receive a non-negative profit from service k which contradicts to the assumption that service k's potential value is negative.) And for other services, we have learned that $\phi^{G,uni,*}$ is the optimal design that maximizes SP's profit. Hence, the optimal GPU VCRB design $\phi^{G,uni,*}$ for $\mathcal{S}$ is still the optimal design for $\mathcal{S}'$.
2) If service k chooses GPU VCRB: Assume that ϕ_k^G is any GPU VCRB that will be chosen by type-k users. Since service k's potential value is negative, any user that chooses GPU VCRB ϕ_k^G will bring SP a profit less than p^C. Therefore, SP's profit under ϕ_k^G will not reach Np^C and hence less than the profit in the first case.

From above two cases, we learn that the optimal uniform GPU VCRB design for the new service set $\mathcal{S}'$ is the same as the optimal uniform design for the original service set $\mathcal{S}$. Hence we conclude that introducing a service with a negative potential value will not change SP's optimal uniform VCRB design. Similarly, excluding a service with a negative potential value in the original service set will also not change the optimal uniform GPU VCRB design. Finally, we summarize that in the uniform GPU VCRB design problem, services with negative potential values do not affect SP's optimal uniform VCRB design.

This completes our proof.

Appendix D
Proof of Lemma 3

Given two services i and j, if there exists a balanced GPU VCRB design between service i and j, there must exist some GPU VCRB design ϕ_{ij}^G that makes both $\Delta(\phi_{ij}^G, \mathcal{T}_i) = 0$ and $\Delta(\phi_{ij}^G, \mathcal{T}_j) = 0$ true. Combining above two equations, we obtain the balanced design as follows,

$$p_{ij}^G = \frac{\lambda}{f^C}\frac{\theta_j - \theta_i}{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}} + p^C, \tag{D.1}$$

$$f_{ij}^G = f^C\frac{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}}{\theta_i\theta_j(\frac{1}{l_j} - \frac{1}{l_i})}. \tag{D.2}$$

Let $z = \frac{\lambda}{f^C}\frac{\theta_j - \theta_i}{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}}$, hence $z \geq 0$ is equivalent to $p_{ij}^G \geq p^C$. If there exists a balanced GPU VCRB design between service i and j, both z and f_{ij}^G are positive. Then we discuss the existence of the balanced design by z and f_{ij}^G in the following four cases:

1) If $(l_i - l_j)(\theta_i - \theta_j) > 0$: Without loss of generality, we assume that $(l_i - l_j) < 0$ and $(\theta_i - \theta_j) < 0$. We have $z f_{ij}^G = \lambda \frac{\theta_j - \theta_i}{\theta_i \theta_j (\frac{1}{l_j} - \frac{1}{l_i})} < 0$, which indicates that either z or f_{ij}^G is negative. Hence there exists no balanced design between service i and j.
2) If $l_i - l_j = 0$ and $\theta_i - \theta_j \neq 0$: In this case, $f_{ij}^G \to \infty$, hence the balanced design between service i and service j does not exist.
3) If $l_i - l_j \neq 0$ and $\theta_i - \theta_j = 0$: In this case, we have $z = 0$ and $f_{ij}^G = \frac{f^C}{\theta_i}$, which is a feasible design. Hence in this case, GPU VCRB design $\phi_{ij}^G = (p^C, \frac{f^C}{\theta_i})$ is the balanced design between service i and j but is not profitable since $p_{ij}^G = p^C$.
4) If $(l_i - l_j)(\theta_i - \theta_j) < 0$: In this case, we verify that $z \geq 0$ and $f_{ij}^G \geq 0$. Thus, GPU VCRB design $\phi_{ij}^G = (p_{ij}^G, f_{ij}^G) = (\frac{\lambda}{f^C} \frac{\theta_j - \theta_i}{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}} + p^C, f^C \frac{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}}{\theta_i \theta_j (\frac{1}{l_j} - \frac{1}{l_i})})$ is the balanced design between service i and j.

Then we give the expressions of service i and j's balanced profit. Substituting Eq. (D.1) and Eq. (D.2) into SP's profit gives us the expression of balanced profit as

$$H_{ij} = N \left[p^C + \frac{\lambda}{f^C} \frac{\theta_j - \theta_i}{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}} - p^D f^C \frac{\frac{\theta_j}{l_j} - \frac{\theta_i}{l_i}}{\theta_i \theta_j (\frac{1}{l_j} - \frac{1}{l_i})} \right]. \tag{D.3}$$

This completes our proof.

APPENDIX E
PROOF OF THEOREM 2

In this section, we derive the optimal uniform GPU VCRB design for two services. We analyze the problem in two cases, i.e., $(l_i - l_j)(\theta_i - \theta_j) \geq 0$ and $(l_i - l_j)(\theta_i - \theta_j) < 0$. In the following analysis, without loss of generality, we assume that service j has a higher potential value, i.e., $q_i \leq q_j$.

A. case with $(l_i - l_j)(\theta_i - \theta_j) \geq 0$

In this case, we observe the relationship between two services' properties $\mathcal{T}_i$ and $\mathcal{T}_j$ as shown in the following lemma.

Lemma 5. *In the case where $(\theta_i - \theta_j)(l_i - l_j) \geq 0$, if the service with higher (lower) potential value chooses CPU (GPU) VCRB, the other will also choose CPU (GPU) VCRB.*

Proof. Given service i and service j that satisfy $(\theta_i - \theta_j)(l_i - l_j) \geq 0$, since $(\theta_i - \theta_j)(l_i - l_j) \geq 0$ and $q_i \leq q_j$, we can infer that $\theta_i - \theta_j \leq 0$ and $l_i - l_j \leq 0$. Given the GPU VCRB ϕ^G, if service j chooses CPU VCRB, it refers that $\Delta(\phi^G; \mathcal{T}_j) \notin [0, \theta_i]$. Then we have

1) if $\Delta(\phi^G; \mathcal{T}_j) < 0$, $\Delta(\phi^G; \mathcal{T}_i) \leq \Delta(\phi^G; \mathcal{T}_j) < 0$, hence type-$i$ users will also choose the CPU VCRB,
2) if $\Delta(\phi^G; \mathcal{T}_j) > \theta_j$, then $[-(p^G - p^C)f^C + \lambda l_i] < 0$, and hence $\Delta(\phi^G; \mathcal{T}_i) > \theta_i$. Type-$i$ users will also choose the CPU VCRB.

From above two cases we learn that if service j choose the CPU VCRB, service i will also choose the CPU VCRB.

If service i choose the GPU VCRB, we have that $\Delta(\phi^G; \mathcal{T}_i) \in [0, \theta_i]$, then

1) we have $[-(p^G - p^C)f^C + \lambda l_i] < 0$, hence $\Delta(\phi^G; \mathcal{T}_j) \leq \theta_j$,
2) we have

$$\begin{aligned}
\Delta(\phi^G; \mathcal{T}_j) =& \theta_j - \frac{\lambda l_j f^C}{[-(p^G - p^C)f^C + \lambda l_j] f^G}, \\
\geq& \theta_j - \frac{\lambda l_i f^C}{[-(p^G - p^C)f^C + \lambda l_i] f^G}, \\
\geq& \theta_i - \frac{\lambda l_i f^C}{[-(p^G - p^C)f^C + \lambda l_i] f^G}, \\
=& \Delta(\phi^G; \mathcal{T}_i) \geq 0.
\end{aligned} \tag{E.1}$$

Hence we prove that if service i choose the GPU VCRB, service j will also choose the GPU VCRB.

This completes our proof. □

Lemma 5 eliminates the case where the service with a higher potential value chooses the CPU VCRB and the other chooses the GPU VCRB, and simplifies the constraints by demonstrating the dominance relation between the two constraints. Next, we discuss the optimal uniform design in the rest three cases.

1) Both type-i and type-j users choose the CPU VCRB: in this case where $\Delta(\phi^G; \mathcal{T}_i) \notin [0, \theta_i]$ and $\Delta(\phi^G; \mathcal{T}_j) \notin [0, \theta_j]$, no GPU VCRB is demanded, thus no GPU resource is deployed. SP's profit is $H(0, 0) = N p^C$.
2) type-i users choose the CPU VCRB and type-j users choose the GPU VCRB: in this case where $\Delta(\phi^G; \mathcal{T}_i) \notin [0, \theta_i]$ and $\Delta(\phi^G; \mathcal{T}_j) \in [0, \theta_j]$, the constraint $\Delta(\phi^G; \mathcal{T}_i) \notin [0, \theta_i]$ is inactive, hence we formulate the uniform VCRB design problem as follows,

$$\max_{\phi^{G,uni}} \; H(\phi^{G,uni}) = N w_i p^C + N w_j (p^{G,uni} - p^D f^{G,uni}) \tag{E.2}$$

$$s.t. \;\; \Delta(\phi^G; \mathcal{T}_j) \in [0, \theta_j]. \tag{E.3}$$

Similar to $(\mathbf{P2})$, we derive the optimal solution and the corresponding profit as follows,

$$\phi^{G,uni,*} = \phi_j^{G,*}, \tag{E.4}$$

$$H(\phi_j^{G,*}) = N p^C + N w_j q_j. \tag{E.5}$$

3) Both service i and service j choose the GPU VCRB: in this case where $\Delta(\phi^G; \mathcal{T}_i) \in [0, \theta_i]$ and $\Delta(\phi^G; \mathcal{T}_j) \in [0, \theta_j]$, constraint $\Delta(\phi^G; \mathcal{T}_j) \in [0, \theta_j]$ can be eliminated according to lemma 5, hence we formulate the uniform VCRB design problem as follows,

$$\max_{\phi^{G,uni}} \; H(\phi^{G,uni}) = N(p^{G,uni} - p^D f^{G,uni}) \tag{E.6}$$

$$s.t. \;\; \Delta(\phi^G; \mathcal{T}_i) \in [0, \theta_i]. \tag{E.7}$$

Similar to $(\mathbf{P2})$, we derive the optimal solution and the corresponding profit as follows,

$$\phi^{G,uni,*} = \phi_i^{G,*}, \tag{E.8}$$

$$H(\phi_i^{G,*}) = N p^C + N q_i. \tag{E.9}$$

summarizing above three cases gives us the optimal uniform design as follows,

$$\phi^{G,uni,*} = \begin{cases} \phi_i^{G,*}, & if \quad q_i \geq \omega_j q_j, \\ \phi_j^{G,*}, & if \quad q_i < \omega_j q_j. \end{cases} \tag{E.10}$$

Recall the assumption that both services have nonnegative potential values, hence $\phi^G = (0,0)$ cannot be the optimal design.

B. case with $(l_i - l_j)(\theta_i - \theta_j) < 0$

In this case, the solution process could be more complex. Given two services i and j with $\mathcal{T}_i$ and T_j, respectively, we introduce two variables as follows,

$$a = [-(p^G - p^C)f^C + \lambda l_i]f^G \Delta(\phi^G, \mathcal{T}_i), \tag{E.11}$$

$$b = [-(p^G - p^C)f^C + \lambda l_j]f^G \Delta(\phi^G, \mathcal{T}_j). \tag{E.12}$$

And we give p^G and f^G's expressions with respect to a and b as follows,

$$p^G(a,b) = p^C + \lambda \frac{l_i \frac{f^C l_j + b}{\theta_j} - l_j \frac{f^C l_i + a}{\theta_i}}{[\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}]f^C}, \tag{E.13}$$

$$f^G(a,b) = \frac{\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}}{l_j - l_i}. \tag{E.14}$$

In the following, we discuss the relationship between a, b and users' VCRB decisions.

1) If $\Delta(\phi^G, \mathcal{T}_i) < 0$, we derive that $[-(p^G - p^C)f^C + \lambda l_i]f^G > 0$, hence $a < 0$ and type-i users choose the CPU VCRB according to Lemma 1.
2) If $0 \leq \Delta(\phi^G, \mathcal{T}_i) \leq \theta_i$, we derive that $[-(p^G - p^C)f^C + \lambda l_i]f^G > 0$, hence $a \geq 0$ and type-i users choose the GPU VCRB according to Lemma 1.
3) If $\Delta(\phi^G, \mathcal{T}_i) > \theta_i$, we derive that $[-(p^G - p^C)f^C + \lambda l_i]f^G < 0$, hence $a < 0$ and type-i users choose the CPU VCRB according to Lemma 1.

Concluding above three cases, we learn that when $a < 0$, type-i users will choose the CPU VCRB and when $a \geq 0$, type-i users will choose the GPU VCRB. Similarly, when $b < 0$, type-j users will choose the CPU VCRB and when $b \geq 0$, type-j users will choose the GPU VCRB.

Let $H'(a,b)$ denote the profit with respect to a and b. We analyze the optimal solution to $(\mathbf{P3'})$ following four cases below.

1) If both service i and j choose the CPU VCRB, i.e.,$a < 0$ and $b < 0$: In this case, no GPU VCRB is demanded and thus no GPU resource is deployed. SP's profit is $H'((0,0)) = Np^C$.
2) If both service i and j choose the GPU VCRB, i.e.,$a \geq 0$ and $b \geq 0$: In this case, we rewrite the objective function with respect to a and b as

$$H'(a,b) = Np^C + N[\lambda \frac{l_i \frac{f^C l_j + b}{\theta_j} - l_j \frac{f^C l_i + a}{\theta_i}}{[\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}]f^C} - p^D \frac{\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}}{l_j - l_i}]. \tag{E.15}$$

Hence we re-formulate the problem as follows,

$$\max_{(a,b)} \ H'(a,b) \tag{E.16}$$

$$s.t. \ \ a \geq 0, \tag{E.17}$$

$$b \geq 0. \tag{E.18}$$

We calculate the partial derivatives of $H'(a,b)$ with respect to a and b as follows,

$$\frac{\partial H'(a,b)}{\partial a} = -N\lambda \frac{(l_j - l_i)(f^C l_j + b)}{\left(\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}\right)^2 f^C \theta_i \theta_j} + \frac{Np^D}{\theta_i(l_j - l_i)}, \tag{E.19}$$

$$\frac{\partial H'(a,b)}{\partial b} = -N\lambda \frac{(l_j - l_i)(f^C l_i + a)}{\left(\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}\right)^2 f^C \theta_i \theta_j} - \frac{Np^D}{\theta_j(l_j - l_i)}. \tag{E.20}$$

It is easy to see that $\frac{\partial H^G(a,b)}{\partial a}$ decreases as a increases in the interval $(0, \frac{f^C l_j + b}{\theta_j}\theta_i - f^C l_i)$ and $\frac{\partial H^G(a,b)}{\partial b}$ decreases as a increases in the interval $(\frac{f^C l_i + a}{\theta_i}\theta_j - f^C l_j, +\infty)$. By setting $\frac{\partial H^G(a,b)}{\partial a} = 0$ and $\frac{\partial H^G(a,b)}{\partial b} = 0$, we obtain stationary points as follows,

$$a_0(b) = -\theta_i(l_j - l_i)\sqrt{\frac{\lambda(f^C l_j + b)}{\theta_j p^D f^C}} + \frac{f^C l_j + b}{\theta_j}\theta_i - f^C l_i, \tag{E.21}$$

$$b_0(a) = \theta_j(l_j - l_i)\sqrt{\frac{\lambda(f^C l_i + a)}{\theta_i p^D f^C}} + \frac{f^C l_i + a}{\theta_i}\theta_j - f^C l_j. \tag{E.22}$$

We observe that the optimal solution (a^*, b^*) subjects to $a^* b^* = 0$. (Otherwise, it indicates that the corresponding solution $(p^{G,*}, f^{G,*})$ doesn't lie on the boundary of the feasible region. Since the feasible region with respect to p^G and f^G is a plain, above solution cannot be the optimal solution, which leads to the contradiction.)

a) If $a^* = 0$: $\frac{\partial H^G(a,b)}{\partial b} \downarrow$ in $(\frac{f^C l_i}{\theta_i}\theta_j - f^C l_j, +\infty)$, that is $\frac{\partial H^G(a,b)}{\partial b} \downarrow$ in $(0, +\infty)$ since $\frac{f^C l_i}{\theta_i}\theta_j - f^C l_j < 0 \leq b$. Substituting $a = a* = 0$ into eq. (E.22), we have the stationary point as follows,

$$b_0(0) = \theta_j(l_j - l_i)\sqrt{\frac{\lambda(f^C l_i)}{\theta_i p^D f^C}} + \frac{f^C l_i}{\theta_i}\theta_j - f^C l_j. \tag{E.23}$$

i) If $b_0(0) < 0$, that is $(l_j - l_i)\sqrt{\frac{\lambda l_i}{\theta_i p^D}} + \frac{f^C l_i}{\theta_i} - \frac{f^C l_j}{\theta_j} < 0$, we have

$$(a^*, b^*) = (0,0), \tag{E.24}$$

$$H'(0,0) = N\left[p^C + \lambda \frac{l_i \frac{l_j}{\theta_j} - l_j \frac{l_i}{\theta_i}}{[\frac{l_j}{\theta_j} - \frac{l_i}{\theta_i}]f^C} - p^D \frac{\frac{f^C l_j}{\theta_j} - \frac{f^C l_i}{\theta_i}}{l_j - l_i}\right]. \tag{E.25}$$

ii) If $b_0(0) \geq 0$, that is $(l_j - l_i)\sqrt{\frac{\lambda l_i}{\theta_i p^D}} + \frac{f^C l_i}{\theta_i} - \frac{f^C l_j}{\theta_j} \geq 0$, we have

$$(a^*, b^*) = (0, b_0(0)), \tag{E.26}$$

$$H'(0, b_0(0)) = Np^C + Nq_i. \tag{E.27}$$

b) If $b^* = 0$: $\frac{\partial H^G(a,b)}{\partial a} \downarrow$ in $(0, \frac{f^C l_j}{\theta_j}\theta_i - f^C l_i)$. Substituting $b = b* = 0$ into eq. (E.21), we have the stationary point $a_0(0) = -\theta_i(l_j - l_i)\sqrt{\frac{\lambda(f^C l_j)}{\theta_j p^D f^C}} + \frac{f^C l_j}{\theta_j}\theta_i - f^C l_i$.

i) If $a_0(0) < 0$, that is $-(l_j - l_i)\sqrt{\frac{\lambda l_j}{\theta_j p^D}} + \frac{f^C l_j}{\theta_j} - \frac{f^C l_i}{\theta_i} < 0$, we have

$$(a^*, b^*) = (0, 0), \tag{E.28}$$

$$H'(0,0) = N\left[p^C + \lambda \frac{l_i \frac{l_j}{\theta_j} - l_j \frac{l_i}{\theta_i}}{[\frac{l_j}{\theta_j} - \frac{l_i}{\theta_i}]f^C} - p^D \frac{\frac{f^C l_j}{\theta_j} - \frac{f^C l_i}{\theta_i}}{l_j - l_i}\right]. \tag{E.29}$$

ii) If $a_0(0) \geq 0$, that is $-(l_j - l_i)\sqrt{\frac{\lambda l_j}{\theta_j p^D}} + \frac{f^C l_j}{\theta_j} - \frac{f^C l_i}{\theta_i} \geq 0$, we have

$$(a^*, b^*) = (a_0(0), 0), \tag{E.30}$$

$$H'(a_0(0), 0) = Np^C + Nq_j. \tag{E.31}$$

3) If type-i users choose the CPU VCRB and type-j users choose the GPU VCRB, i.e.,$a < 0$ and $b \geq 0$: In this case, we rewrite the objective function as

$$H'(a,b) = Np^C + N\omega_j[\lambda \frac{l_i \frac{f^C l_j + b}{\theta_j} - l_j \frac{f^C l_i + a}{\theta_i}}{[\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}]f^C} - p^D \frac{\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}}{l_j - l_i}]. \tag{E.32}$$

Hence we re-formulate the problem as follows,

$$\max_{(a,b)} \; H'(a,b) \tag{E.33}$$

$$s.t. \quad a < 0, \tag{E.34}$$

$$b \geq 0. \tag{E.35}$$

Similar to case (2), we have $b^* = 0$ and we conclude as follows,

a) If $a_0(0) < 0$, we have

$$(a^*, b^*) \to (0^-, 0), \tag{E.36}$$

$$H'(0^-, 0) < Np^C + N\omega_j\left[\lambda \frac{l_i \frac{l_j}{\theta_j} - l_j \frac{l_i}{\theta_i}}{(\frac{l_j}{\theta_j} - \frac{l_i}{\theta_i})f^C} - p^D \frac{\frac{f^C l_j}{\theta_j} - \frac{f^C l_i}{\theta_i}}{l_j - l_i}\right]. \tag{E.37}$$

b) If $a_0(0) \geq 0$, we have

$$(a^*, b^*) = (a_0(0), 0), \tag{E.38}$$

$$H'(a_0(0), 0) = Np^C + N\omega_j q_j. \tag{E.39}$$

4) If type-i users choose the GPU VCRB and type-j users choose the CPU VCRB, i.e.,$a \geq 0$ and $b < 0$: In this case, we rewrite the objective function as

$$H'(a,b) = Np^C + N\omega_i[\lambda \frac{l_i \frac{f^C l_j + b}{\theta_j} - l_j \frac{f^C l_i + a}{\theta_i}}{[\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}]f^C} - p^D \frac{\frac{f^C l_j + b}{\theta_j} - \frac{f^C l_i + a}{\theta_i}}{l_j - l_i}]. \tag{E.40}$$

Hence we re-formulate the problem as follows,

$$\max_{(a,b)} \; H'(a,b) \tag{E.41}$$

$$s.t. \quad a \geq 0, \tag{E.42}$$

$$b < 0. \tag{E.43}$$

Similar to case (2), we have $a^* = 0$ and we conclude as follows,

a) If $b_0(0) < 0$, we have

$$(a^*, b^*) \to (0, 0^-), \tag{E.44}$$

$$H'(0, 0^-) < Np^C + N\omega_i\left[\lambda \frac{l_i \frac{l_j}{\theta_j} - l_j \frac{l_i}{\theta_i}}{(\frac{l_j}{\theta_j} - \frac{l_i}{\theta_i})f^C} - p^D \frac{\frac{f^C l_j}{\theta_j} - \frac{f^C l_i}{\theta_i}}{l_j - l_i}\right]. \tag{E.45}$$

b) If $b_0(0) \geq 0$, we have

$$(a^*, b^*) = (0, b_0(0)), \tag{E.46}$$

$$H'(0, b_0(0)) = Np^C + N\omega_i q_i. \tag{E.47}$$

Notice that when $a = 0$ and $b = 0$, it indicates that $\Delta(\phi^G, \mathcal{T}_i) = 0$ and $\Delta(\phi^G, \mathcal{T}_j) = 0$, respectively. And $(a, b) = (0, b_0(0))$ and $(a, b) = (a_0(0), 0)$ correspond to service i's and service j's optimal design, respectively. Besides, we find that if $a_0(0) \geq 0$ and $b_0(0) \geq 0$, then $\Delta(\phi_j^{G,*}, \mathcal{T}_i) \in [0, \theta_i]$ and $\Delta(\phi_i^{G,*}, \mathcal{T}_j) \in [0, \theta_j]$, respectively. Combining above results, we summarize the optimal solution to $(\mathbf{P3'})$ as follows,

1) if $\Delta(\phi_j^{G,*}; \mathcal{T}_i) \in [0, \theta_i]$ and $\Delta(\phi_i^{G,*}; \mathcal{T}_j) \notin [0, \theta_j]$,

$$\phi^{G,uni,*} = \begin{cases} \phi_i^{G,*}, & if \quad \omega_i q_i > q_j, \\ \phi_j^{G,*}, & if \quad \omega_i q_i \leq q_j, \end{cases} \tag{E.48}$$

2) if $\Delta(\phi_j^{G,*}; \mathcal{T}_i) \notin [0, \theta_i]$ and $\Delta(\phi_i^{G,*}; \mathcal{T}_j) \in [0, \theta_j]$,

$$\phi^{G,uni,*} = \begin{cases} \phi_i^{G,*}, & if \quad q_i \geq \omega_j q_j, \\ \phi_j^{G,*}, & if \quad q_i < \omega_j q_j, \end{cases} \tag{E.49}$$

3) if $\Delta(\phi_j^{G,*}; \mathcal{T}_i) \notin [0, \theta_i]$ and $\Delta(\phi_i^{G,*}; \mathcal{T}_j) \notin [0, \theta_j]$,

$$\phi^{G,uni,*} = \begin{cases} \phi_i^{G,*}, & if \quad \omega_i q_i = \max\{\omega_i q_i, \omega_j q_j, q_{ij}\}, \\ \phi_j^{G,*}, & if \quad \omega_j q_j = \max\{\omega_i q_i, \omega_j q_j, q_{ij}\}, \\ \phi_{ij}^{G}, & if \quad q_{ij} = \max\{\omega_i q_i, \omega_j q_j, q_{ij}\}. \end{cases} \tag{E.50}$$

This completes our proof.

APPENDIX F
PROOF OF LEMMA 4

As shown in Theorem 2, the optimal uniform GPU VCRB design for two services is among service i's optimal design, service j's optimal design and their balanced design. Given the uniform GPU VCRB design $\phi^{G,uni}$, we prove Lemma 4 in the following four cases.

1) If there does not exists an $i \in \mathcal{S}$ satisfies $\Delta(\phi^{G,uni}, \mathcal{T}_i) = 0$: SP could increase the service price or decrease the computation capacity to achieve a higher profit without changing users' VCRB decisions. Hence $\phi^{G,uni}$ is not the optimal design.
2) If there is only one service i that satisfies $\Delta(\phi^{G,uni}, \mathcal{T}_i) = 0$: We observe that without changing users' VCRB decisions, the discriminant function $\Delta(\phi^G, \mathcal{T}_i)$ is differentiable with respect to p^G and f^G and $\frac{\partial \Delta}{\partial p^G} < 0$ and $\frac{\partial \Delta}{\partial f^G} > 0$. Let $\boldsymbol{t}$ be the tangent vector of $\Delta(\phi^G, \mathcal{T}_i) = 0$ at $(p^{G,uni}, f^{G,uni})$ as follows,
$$\boldsymbol{t} = (-\frac{\partial \Delta}{\partial f^G}, \frac{\partial \Delta}{\partial p^G})|_{(p^G, f^G)=(p^{G,uni}, f^{G,uni})} \tag{F.1}$$
Hence the uniform design $\phi^{G,uni}$ is a local optimal solution iff $\nabla_{\boldsymbol{t}} H(\phi^G)|_{\phi^G=\phi^{G,uni}} = 0$. Then we prove that $\nabla_{\boldsymbol{t}} H(\phi^G)|_{\phi^G=\phi^{G,uni}} = 0$ holds iff $\phi^{G,uni} = \phi_i^{G,*}$. According to the definition of service i's optimal design $\phi_i^{G,*}$, we have that $\nabla_{\boldsymbol{t}} H_i(\phi^G)|_{\phi^G=\phi_i^{G,*}} = 0$. Let $\mathcal{S}_1 = \{i : i \in \mathcal{S}, x_i = 0\}$ denote the service set that chooses the CPU VCRB, and $\mathcal{S}_2 = \{i : i \in \mathcal{S}, x_i = 1\}$ denote the service set that chooses the GPU VCRB given $\phi^{G,uni}$. Since $\Delta(\phi^{G,uni}, \mathcal{T}_j) \neq 0$ for each $j \in \mathcal{S}, j \neq i$ as given in the conditions, we have that
$$\begin{aligned}\nabla_{\boldsymbol{t}} H_j(\phi^G)|_{\phi^G=\phi^{G,uni}} &= \nabla_{\boldsymbol{t}} H_j(\phi^G)|_{\phi^G=\phi_i^{G,*}} \\ &= \begin{cases} 0, & i \in \mathcal{S}_1, \\ \nabla_{\boldsymbol{t}} H_i(\phi^G)|_{\phi^G=\phi_i^{G,*}} = 0, & i \in \mathcal{S}_2. \end{cases}\end{aligned} \tag{F.2}$$
Hence we have that
$$\nabla_{\boldsymbol{t}} H(\phi^G)|_{\phi^G=\phi_i^{G,*}} = \sum_{i \in \mathcal{S}} \nabla_{\boldsymbol{t}} H_i(\phi^G)|_{\phi^G=\phi_i^{G,*}} = 0. \tag{F.3}$$
Similarly we know that when $\phi^{G,uni} \neq \phi_i^{G,*}$, $\nabla_{\boldsymbol{t}} H(\phi^G)|_{\phi^G=\phi^{G,uni}} \neq 0$. Then we have prove that in this case, the optimal uniform GPU VCRB design could only be service i's optimal design $\phi_i^{G,*}$.
3) If there are only two services $i, j \in \mathcal{S}, i \neq j$ satisfies $\Delta(\phi^{G,uni}, \mathcal{T}_i) = 0$ and $\Delta(\phi^{G,uni}, \mathcal{T}_j) = 0$: From the definition of the balanced design, we learn that in this case, the uniform design is service i and service j's balanced design, i.e., $\phi^{G,uni} = \phi_{ij}^G$. And the balanced design is included in the candidate range.
4) If there are more than two services $i \in \mathcal{S}'$ ($\mathcal{S}'$ is a subset of $\mathcal{S}$) satisfies $\Delta(\phi^{G,uni}, \mathcal{T}_i) = 0$: Since there are mere two variables, i.e., p^G and f^G, two services uniquely determine a design. Hence in this case, the balanced design of any two services $i, j \in \mathcal{S}'$ is the same design and is the uniform design $\phi^{G,uni}$.

In conclusion, the optimal uniform GPU VCRB design is either one service's optimal design or two services' balanced design.

This completes our proof.

APPENDIX G
PROOF OF PROPOSITION 1

In Theorem 2, we have proved the optimality of the solution. Then we analyze the complexity of Algorithm 1.

In the best case, all services have negative potential values, and the algorithm terminates after calculating each service's potential value and constructing the service set $\mathcal{S}_1$, whose time complexity is $\Omega(S)$. Hence the algorithm's time complexity in the best case is $\Omega(S)$.

In the worst case, all services have non-negative potential values, and the algorithm takes $O(S)$ time to construct the service set $\mathcal{S}_1$, $O(S)$ time to search the optimal design of each service, and $O(S^2)$ time to search the balanced designs. Hence the overall time complexity is $O(S^2)$.

This completes our proof.