APPENDIX A
PROOF OF THEOREM 1

The Stage III resource allocation problem is a linear programming problem with the objective of minimizing operational cost $C(\boldsymbol{x})$ subject to constraints (3), (4), (9b), and (9c).

Given the cost structure where $c^{TH} < c^{TL}$ and $c^{IH} > c^{IL}$, the optimal allocation strategy is to:

1) Assign as much training demand as possible to high-performance GPUs (cheaper for training)
2) Assign as much inference demand as possible to low-performance GPUs (cheaper for inference)

The optimal allocation follows directly from this cost-minimization principle:

- When $\sum_{n=1}^N g_n^T \leq G^H$, all training demand can be allocated to high-performance GPUs: $x_n^{TH*} = g_n^T$
- When $\sum_{n=1}^N g_n^T > G^H$, high-performance GPUs are fully utilized and training demand is proportionally allocated: $x_n^{TH*} = g_n^T \frac{G^H}{\sum_{n=1}^N g_n^T}$
- When $\sum_{n=1}^N g_n^I \leq G^L$, all inference demand can be allocated to low-performance GPUs: $x_n^{IL*} = g_n^I$
- When $\sum_{n=1}^N g_n^I > G^L$, low-performance GPUs are fully utilized and inference demand is proportionally allocated: $x_n^{IL*} = g_n^I \frac{G^L}{\sum_{n=1}^N g_n^I}$

The remaining allocations are determined by the demand constraints:

$$x_n^{TL*} = g_n^T - x_n^{TH*}$$
$$x_n^{IH*} = g_n^I - x_n^{IL*}$$

If $\sum_{n=1}^N g_n^T + \sum_{n=1}^N g_n^I > G^H + G^L = G$, the problem is infeasible due to insufficient GPU capacity.

APPENDIX B
PROOF OF THEOREM 2

For the inference service, client n solves:

$$\max_{g_n^I \geq 0} P_n^I(g_n^I) = -\gamma_n^I \cdot \frac{1}{\left(\frac{g_n^I}{d_n} - \lambda_n\right)^2} - p^I \cdot g_n^I$$

Taking the first-order derivative:

$$\frac{\partial P_n^I}{\partial g_n^I} = \gamma_n^I \cdot \frac{2}{d_n \left(\frac{g_n^I}{d_n} - \lambda_n\right)^3} - p^I$$

Setting the derivative to zero:

$$\gamma_n^I \cdot \frac{2}{d_n \left(\frac{g_n^I}{d_n} - \lambda_n\right)^3} = p^I$$

Solving for g_n^I:

$$\left(\frac{g_n^I}{d_n} - \lambda_n\right)^3 = \frac{2\gamma_n^I}{d_n p^I}$$

$$\frac{g_n^I}{d_n} - \lambda_n = \left(\frac{2\gamma_n^I}{d_n p^I}\right)^{1/3}$$

$$g_n^I = d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I}\right)^{1/3}\right)$$

The threshold price $p_n^{I\dagger}$ is obtained by solving:

$$-\gamma_n^I \cdot \frac{1}{\left(\frac{g_n^I}{d_n} - \lambda_n\right)^2} - p^I \cdot g_n^I = 0$$

which gives the condition for positive demand.

APPENDIX C
PROOF OF EQUATION (17)

The cost function $C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p})))$ is derived by substituting the optimal Stage III allocation into the cost function and considering all possible cases of resource utilization:

1) **Case 1:** $D^T(p^T) \leq G^H$ and $D^I(p^I) \leq G^L$
   All training on high-performance GPUs, all inference on low-performance GPUs:
   $$C = c^{TH} D^T(p^T) + c^{IL} D^I(p^I)$$
2) **Case 2:** $D^T(p^T) > G^H$ and $D^I(p^I) \leq G^L$
   High-performance GPUs fully utilized for training, excess training on low-performance GPUs, all inference on low-performance GPUs:
   $$C = c^{TH} G^H + c^{TL}(D^T(p^T) - G^H) + c^{IL} D^I(p^I)$$
3) **Case 3:** $D^T(p^T) \leq G^H$ and $D^I(p^I) > G^L$
   All training on high-performance GPUs, low-performance GPUs fully utilized for inference, excess inference on high-performance GPUs:
   $$C = c^{TH} D^T(p^T) + c^{IL} G^L + c^{IH}(D^I(p^I) - G^L)$$
4) **Case 4:** $D^T(p^T) > G^H$ and $D^I(p^I) > G^L$
   Infeasible due to insufficient resources.

The indicator functions in Equation (17) precisely capture these four cases.

APPENDIX D
DERIVATION OF $\mathcal{C}_{p,q}^T$ AND $\mathcal{C}_{p,q}^I$

In subregion $Q_{p,q}$ with $p^T \in [a_p^T, a_{p+1}^T)$ and $p^I \in [a_q^I, a_{q+1}^I)$, the active client sets are determined by the price thresholds:

For training services, client n has positive demand when:

$$p^T \leq \frac{\gamma_n^T}{\theta_n}$$

Since $p^T \in [a_p^T, a_{p+1}^T)$, the active training clients are:

$$\mathcal{C}_{p,q}^T = \left\{ n \in \mathcal{N} : \frac{\gamma_n^T}{\theta_n} \geq a_{p+1}^T \right\}$$

For inference services, client n has positive demand when:

$$p^I \leq p_n^{I\dagger}$$

The active inference clients are:

$$\mathcal{C}_{p,q}^I = \left\{ n \in \mathcal{N} : p_n^{I\dagger} \geq a_{q+1}^I \right\}$$

These sets remain constant within each subregion $Q_{p,q}$.

Appendix E
Proof of Equation (21)

The additional critical points ${a'}_p^T$ and ${a'}_q^I$ are derived from the resource capacity constraints:

For training services, the critical point where $D^T(p^T) = G^H$ is found by solving:

$$\sum_{n \in \mathcal{C}_{p,q}^T} \theta_n \left(\frac{\gamma_n^T}{p^T \theta_n} - 1 \right) = G^H$$

$$\sum_{n \in \mathcal{C}_{p,q}^T} \left(\frac{\gamma_n^T}{p^T} - \theta_n \right) = G^H$$

$$\frac{1}{p^T} \sum_{n \in \mathcal{C}_{p,q}^T} \gamma_n^T - \sum_{n \in \mathcal{C}_{p,q}^T} \theta_n = G^H$$

$$\frac{1}{p^T} \sum_{n \in \mathcal{C}_{p,q}^T} \gamma_n^T = G^H + \sum_{n \in \mathcal{C}_{p,q}^T} \theta_n$$

$${a'}_p^T = \frac{\sum\limits_{n \in \mathcal{C}_{p,q}^T} \gamma_n^T}{G^H + \sum\limits_{n \in \mathcal{C}_{p,q}^T} \theta_n}$$

For inference services, the critical point where $D^I(p^I) = G^L$ is found by solving:

$$\sum_{n \in \mathcal{C}_{p,q}^I} d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \right) = G^L$$

$$\sum_{n \in \mathcal{C}_{p,q}^I} d_n \lambda_n + \sum_{n \in \mathcal{C}_{p,q}^I} d_n^{2/3} (2\gamma_n^I)^{1/3} p^{-1/3} = G^L$$

$$p^{-1/3} = \frac{G^L - \sum_{n \in \mathcal{C}_{p,q}^I} d_n \lambda_n}{\sum_{n \in \mathcal{C}_{p,q}^I} d_n^{2/3} (2\gamma_n^I)^{1/3}}$$

$${a'}_q^I = \left(\frac{\sum\limits_{n \in \mathcal{C}_{p,q}^I} d_n^{2/3} (2\gamma_n^I)^{1/3}}{G^L - \sum\limits_{n \in \mathcal{C}_{p,q}^I} d_n \lambda_n} \right)^3$$

Appendix F
Proof of Equation (22)

In each subregion $Q'_{p,q}$, the indicator functions in Equation (17) become constants, allowing us to express the cost function as a linear combination of the demand functions:

$$C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) = A_{p,q} D_{p,q}^T(p^T) + B_{p,q} D_{p,q}^I(p^I) + E_{p,q}$$

where the coefficients $A_{p,q}$, $B_{p,q}$, and $E_{p,q}$ depend on which case of the piecewise cost function applies in subregion $Q'_{p,q}$.

The revenue function remains:

$$R(\boldsymbol{p}, \boldsymbol{g}^*(\boldsymbol{p})) = p^T D_{p,q}^T(p^T) + p^I D_{p,q}^I(p^I)$$

Therefore, the profit function becomes:

$$F_{p,q}(\boldsymbol{p}) = p^T D_{p,q}^T(p^T) + p^I D_{p,q}^I(p^I) - A_{p,q} D_{p,q}^T(p^T) - B_{p,q} D_{p,q}^I(p^I) - E_{p,q}$$

This reformulation eliminates the indicator functions and provides a continuous objective function within each subregion.

Appendix G
Proof of Theorem 3

Within each subregion $Q'_{p,q}$, we analyze the monotonicity properties of the objective function $F_{p,q}(\boldsymbol{p})$:

For p^I: The derivative with respect to p^I is:

$$\frac{\partial F_{p,q}}{\partial p^I} = D_{p,q}^I(p^I) + p^I \frac{\partial D_{p,q}^I}{\partial p^I} - B_{p,q} \frac{\partial D_{p,q}^I}{\partial p^I}$$

Since $D_{p,q}^I(p^I)$ is decreasing in p^I and the marginal cost $B_{p,q}$ is constant, the function is increasing in p^I within the subregion. Therefore, the optimal p^I is at the upper boundary:

$$p_{p,q}^{I*} = a_{q+1}^I$$

For p^T: The optimal p^T is found by analyzing the first-order condition:

$$\frac{\partial F_{p,q}}{\partial p^T} = D_{p,q}^T(p^T) + (p^T - A_{p,q}) \frac{\partial D_{p,q}^T}{\partial p^T} = 0$$

Solving this equation gives the critical point p'^T. The optimal solution depends on the position of p'^T relative to the subregion boundaries $[a_p^T, a_{p+1}^T]$ and the resource constraint boundary b_p^T.

The three cases in the theorem cover all possible scenarios:

1) If $p'^T < a_p^T$, the function is decreasing in the subregion, so the optimal is at the lower boundary
2) If $p'^T \in [a_p^T, a_{p+1}^T)$, the critical point is feasible and optimal
3) If $p'^T \geq a_{p+1}^T$, the function is increasing in the subregion, so the optimal is at the upper boundary

The constraint boundary b_p^T ensures the solution satisfies the total resource constraint.