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 training service, client n solves:

$$\max_{g_n^T \geq 0} P_n^T(g_n^T) = \gamma_n^T \cdot \log\left(1 + g_n^T/\theta_n\right) - p^T \cdot g_n^T$$

Taking the first-order derivative:

$$\frac{\partial P_n^T}{\partial g_n^T} = \gamma_n^T \cdot \frac{1/\theta_n}{1 + g_n^T/\theta_n} - p^T = \frac{\gamma_n^T}{\theta_n + g_n^T} - p^T$$

Setting the derivative to zero:

$$\frac{\gamma_n^T}{\theta_n + g_n^T} = p^T$$

Solving for g_n^T:

$$\theta_n + g_n^T = \frac{\gamma_n^T}{p^T}$$

$$g_n^T = \frac{\gamma_n^T}{p^T} - \theta_n$$

For non-negativity, we require:

$$\frac{\gamma_n^T}{p^T} - \theta_n \geq 0 \quad \Rightarrow \quad p^T \leq \frac{\gamma_n^T}{\theta_n}$$

Therefore, the optimal training demand is:

$$g_n^{T*}(p^T) = \theta_n \left[\frac{\gamma_n^T}{p^T \theta_n} - 1\right]_+ = \mathbf{1}(p^T \leq \frac{\gamma_n^T}{\theta_n})\theta_n \left(\frac{\gamma_n^T}{p^T \theta_n} - 1\right)$$

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.

Determination of $p_n^{I\dagger}$:

The threshold price $p_n^{I\dagger}$ is the price at which the client becomes indifferent between using the service or not, i.e., when the maximum payoff equals zero:

$$P_n^I(g_n^{I*}) = 0$$

Substituting the optimal demand:

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

From the first-order condition, we have:

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

Substituting this into the indifference condition:

$$-\gamma_n^I \cdot \frac{1}{\left(\left(\frac{2\gamma_n^I}{d_n p^I}\right)^{1/3}\right)^2} - p^I \cdot d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I}\right)^{1/3}\right) = 0$$

Simplifying:

$$-\gamma_n^I \cdot \frac{1}{\left(\frac{2\gamma_n^I}{d_n p^I}\right)^{2/3}} - p^I d_n \lambda_n - p^I d_n \left(\frac{2\gamma_n^I}{d_n p^I}\right)^{1/3} = 0$$

$$-\gamma_n^I \cdot \left(\frac{d_n p^I}{2\gamma_n^I}\right)^{2/3} - p^I d_n \lambda_n - (2\gamma_n^I d_n^2 p^I)^{1/3} = 0$$

Let $X = (p^I d_n)^{1/3}$, then:

$$-\gamma_n^I \cdot \left(\frac{X^3}{2\gamma_n^I}\right)^{2/3} - X^3 \lambda_n - (2\gamma_n^I d_n^2)^{1/3} X = 0$$

$$-\gamma_n^I \cdot \frac{X^2}{(2\gamma_n^I)^{2/3}} - X^3 \lambda_n - (2\gamma_n^I d_n^2)^{1/3} X = 0$$

Multiply through by $(2\gamma_n^I)^{2/3}$:

$$-\gamma_n^I X^2 - X^3 \lambda_n (2\gamma_n^I)^{2/3} - (2\gamma_n^I d_n^2)^{1/3} X (2\gamma_n^I)^{2/3} = 0$$

$$-\gamma_n^I X^2 - \lambda_n (2\gamma_n^I)^{2/3} X^3 - 2\gamma_n^I d_n^{2/3} X = 0$$

Divide by $-X$ (note $X > 0$):

$$\gamma_n^I X + \lambda_n (2\gamma_n^I)^{2/3} X^2 + 2\gamma_n^I d_n^{2/3} = 0$$

This is a quadratic in X:

$$\lambda_n (2\gamma_n^I)^{2/3} X^2 + \gamma_n^I X + 2\gamma_n^I d_n^{2/3} = 0$$

Solving this quadratic equation gives the closed-form expression for X, and thus for $p_n^{I\dagger} = \frac{X^3}{d_n}$.

The exact closed-form solution is:

$$p_n^{I\dagger} = \frac{1}{d_n} \left(\frac{-\gamma_n^I + \sqrt{(\gamma_n^I)^2 - 8\lambda_n (2\gamma_n^I)^{5/3} d_n^{2/3}}}{2\lambda_n (2\gamma_n^I)^{2/3}}\right)^3$$

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:** $D^T(p^T) \le G^H$ and $D^I(p^I) \le G^L$
   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 $D^I(p^I) \le G^L$
   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:** $D^T(p^T) \le G^H$ 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:

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

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

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

For inference services, client n has positive demand when:

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

The active inference clients are:

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

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:

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

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

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

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

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

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

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

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

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

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

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:

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

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:

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

The specific values of $A_{p,q}$, $B_{p,q}$, and $E_{p,q}$ depend on which of the four cases in Equation (17) applies in subregion $Q'_{p,q}$:

Case 1: $D^T(p^T) \le G^H$ and $D^I(p^I) \le G^L$

$$A_{p,q} = c^{TH}$$
$$B_{p,q} = c^{IL}$$
$$E_{p,q} = 0$$

Case 2: $D^T(p^T) > G^H$ and $D^I(p^I) \le G^L$

$$A_{p,q} = c^{TL}$$
$$B_{p,q} = c^{IL}$$
$$E_{p,q} = (c^{TH} - c^{TL})G^H$$

Case 3: $D^T(p^T) \le G^H$ and $D^I(p^I) > G^L$

$$A_{p,q} = c^{TH}$$
$$B_{p,q} = c^{IH}$$
$$E_{p,q} = (c^{IL} - c^{IH})G^L$$

Case 4: $D^T(p^T) > G^H$ and $D^I(p^I) > G^L$

$$A_{p,q} = +\infty$$
$$B_{p,q} = +\infty$$
$$E_{p,q} = +\infty$$

The revenue function remains:

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

Therefore, the profit function becomes:

$$\begin{aligned} F_{p,q}(\boldsymbol{p}) &= R(\boldsymbol{p}, \boldsymbol{g}^*(\boldsymbol{p})) - C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) \\ &= p^T D^T_{p,q}(p^T) + p^I D^I_{p,q}(p^I) \\ &\quad - A_{p,q}D^T_{p,q}(p^T) - B_{p,q}D^I_{p,q}(p^I) - E_{p,q} \end{aligned}$$

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:

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

Since $D^I_{p,q}(p^I)$ 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:

$$p^{I*}_{p,q} = a^I_{q+1}$$

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

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

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

The three cases in the theorem cover all possible scenarios:

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

The constraint boundary b^T_p ensures the solution satisfies the total resource constraint.

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

1. Derivation of p'^T (critical point for p^T):

The objective function with respect to p^T is:

$$F_{p,q}(p^T) = p^T D^T_{p,q}(p^T) - A_{p,q}D^T_{p,q}(p^T) + \text{terms independent of } p^T$$

Where $D^T_{p,q}(p^T) = \sum_{n\in\mathcal{C}^T_{p,q}} \theta_n\left(\frac{\gamma^T_n}{p^T\theta_n} - 1\right) = \frac{1}{p^T}\sum_{n\in\mathcal{C}^T_{p,q}} \gamma^T_n - \sum_{n\in\mathcal{C}^T_{p,q}} \theta_n$

Let $\Gamma = \sum_{n\in\mathcal{C}^T_{p,q}} \gamma^T_n$ and $\Theta = \sum_{n\in\mathcal{C}^T_{p,q}} \theta_n$, then:

$$D^T_{p,q}(p^T) = \frac{\Gamma}{p^T} - \Theta$$

The p^T-dependent part of the objective function is:

$$F^T_{p,q}(p^T) = p^T\left(\frac{\Gamma}{p^T} - \Theta\right) - A_{p,q}\left(\frac{\Gamma}{p^T} - \Theta\right) = \Gamma - \Theta p^T - \frac{A_{p,q}\Gamma}{p^T} + A_{p,q}\Theta$$

Taking the derivative with respect to p^T:

$$\frac{\partial F^T_{p,q}}{\partial p^T} = -\Theta + \frac{A_{p,q}\Gamma}{p^{T2}}$$

Setting the derivative to zero to find the critical point:

$$-\Theta + \frac{A_{p,q}\Gamma}{p^{T2}} = 0$$

$$\frac{A_{p,q}\Gamma}{p^{T2}} = \Theta$$

$$p^T = \sqrt{\frac{A_{p,q}\Gamma}{\Theta}}$$

This gives us the critical point $p'^T = \sqrt{\frac{A_{p,q}\sum_{n\in\mathcal{C}_{p,q}^T}\gamma_n^T}{\sum_{n\in\mathcal{C}_{p,q}^T}\theta_n}}$.

To confirm this is a maximum, we check the second derivative:

$$\frac{\partial^2 F_{p,q}^T}{\partial(p^T)^2} = -\frac{2A_{p,q}\Gamma}{p^{T3}} < 0 \quad (\text{since } A_{p,q} > 0, \Gamma > 0, p^T > 0)$$

Thus, p'^T is indeed a maximum point for the p^T-dependent part of the objective function.

2. Derivation of b_p^T (constraint boundary for p^T):

The resource constraint from Equation (7b) is:

$$\sum_{n=1}^{N}(g_n^T(\boldsymbol{p}) + g_n^I(\boldsymbol{p})) \leq G$$

Substituting the demand functions:

$$D_{p,q}^T(p^T) + D_{p,q}^I(p^I) \leq G$$

We already have $p_{p,q}^{I*} = a_{q+1}^I$ from the monotonicity analysis. Substituting this:

$$D_{p,q}^T(p^T) + D_{p,q}^I(a_{q+1}^I) \leq G$$

Using the expression for $D_{p,q}^T(p^T)$:

$$\left(\frac{\Gamma}{p^T} - \Theta\right) + D_{p,q}^I(a_{q+1}^I) \leq G$$

Solving for p^T:

$$\frac{\Gamma}{p^T} \leq G + \Theta - D_{p,q}^I(a_{q+1}^I)$$

$$p^T \geq \frac{\Gamma}{G + \Theta - D_{p,q}^I(a_{q+1}^I)}$$

Thus, the constraint boundary is:

$$b_p^T = \frac{\sum_{n\in\mathcal{C}_{p,q}^T}\gamma_n^T}{G + \sum_{n\in\mathcal{C}_{p,q}^T}\theta_n - D_{p,q}^I(a_{q+1}^I)}$$

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

3. Optimal solution selection:

Based on the position of p'^T relative to the subregion boundaries $[a_p^T, a_{p+1}^T)$ and the constraint boundary b_p^T, we have three cases:

- **Case 1:** $p'^T < a_p^T$
  The critical point is below the subregion, so the function is decreasing in the subregion. The optimal is at the lower boundary, but must satisfy the constraint:
  $$p_{p,q}^{T*} = \max(a_p^T, b_p^T)$$
- **Case 2:** $a_p^T \leq p'^T < a_{p+1}^T$
  The critical point is within the subregion. The optimal is the critical point, but must satisfy the constraint:
  $$p_{p,q}^{T*} = \max(p'^T, b_p^T)$$
- **Case 3:** $a_{p+1}^T \leq p'^T$
  The critical point is above the subregion, so the function is increasing in the subregion. The optimal is at the upper boundary:
  $$p_{p,q}^{T*} = a_{p+1}^T$$
  Note: In this case, we don't need to consider b_p^T explicitly because we've already verified the feasibility of the subregion, which ensures that a_{p+1}^T satisfies the constraint.

This completes the proof of Theorem 3.