

APPENDIX A
PROOF OF THEOREM 1

The Stage III resource allocation problem is a linear programming problem with the objective of minimizing operational cost $C(\mathbf{x})$ 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:

$$\begin{aligned} x_n^{TL*} &= g_n^T - x_n^{TH*} \\ x_n^{IH*} &= g_n^I - x_n^{IL*} \end{aligned}$$

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

To solve the Problem (8), due to the $P_n^T(g_n^T)$ is only determined by g_n^T and $P_n^I(g_n^I)$ is only determined by g_n^I . We can solve the problem by optimize g_n^T and g_n^I , respectively.

A. For the training service

Client n solves:

$$\max_{g_n^T \geq 0} P_n^T(g_n^T) = \gamma_n^T \cdot \log(1 + g_n^T / \theta_n) - p^T \cdot g_n^T \quad (25)$$

Taking the first-order derivative:

$$\frac{\partial P_n^T}{\partial g_n^T} = \gamma_n^T \cdot \frac{1/\theta_n}{1 + g_n^T / \theta_n} - p^T = \frac{\gamma_n^T}{\theta_n + g_n^T} - p^T \quad (26)$$

Setting the derivative to zero:

$$\frac{\gamma_n^T}{\theta_n + g_n^T} = p^T \quad (27)$$

Solving for g_n^T :

$$\theta_n + g_n^T = \frac{\gamma_n^T}{p^T} \quad (28)$$

$$g_n^T = \frac{\gamma_n^T}{p^T} - \theta_n = \theta_n \left(\frac{\gamma_n^T}{p^T \theta_n} - 1 \right) \quad (29)$$

Considering $g_n^T \geq 0$, we require:

$$\frac{\gamma_n^T}{p^T} - \theta_n \geq 0 \Rightarrow p^T \leq \frac{\gamma_n^T}{\theta_n} \quad (30)$$

Therefore, the optimal training demand is:

$$g_n^*(p^T) = \theta_n \left[\frac{\gamma_n^T}{p^T \theta_n} - 1 \right]_+ = \mathbf{1}(p^T \leq \frac{\gamma_n^T}{\theta_n}) \theta_n \left(\frac{\gamma_n^T}{p^T \theta_n} - 1 \right) \quad (31)$$

B. For the inference service

Client n solves:

$$\max_{g_n^I \geq 0} P_n^I(g_n^I) = -\gamma_n^I \cdot \frac{1}{\left(\frac{g_n^I}{d_n} - \lambda_n \right)^2} - p^I \cdot g_n^I \quad (32)$$

Taking the first-order derivative:

$$\frac{\partial P_n^I}{\partial g_n^I} = \gamma_n^I \cdot \frac{2}{d_n \left(\frac{g_n^I}{d_n} - \lambda_n \right)^3} - p^I \quad (33)$$

Setting the derivative to zero:

$$\gamma_n^I \cdot \frac{2}{d_n \left(\frac{g_n^I}{d_n} - \lambda_n \right)^3} = p^I \quad (34)$$

Solving for g_n^I :

$$\left(\frac{g_n^I}{d_n} - \lambda_n \right)^3 = \frac{2\gamma_n^I}{d_n p^I} \quad (35)$$

$$\frac{g_n^I}{d_n} - \lambda_n = \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \quad (36)$$

$$g_n^{I(1)} = d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \right) \quad (37)$$

Back to (33), we identify the positive and negative of it. We first give the conclusion: The (33) has a singularity $g_n^{I(2)} = d_n \lambda_n$, where $\frac{\partial P_n^I}{\partial g_n^I}(g_n^{I(2)}) = +\infty$ and $\frac{\partial P_n^I}{\partial g_n^I}(g_n^{I(2)}) = -\infty$. The $\frac{\partial P_n^I}{\partial g_n^I}$ is negative within $[0, g_n^{I(2)}]$, positive within $(g_n^{I(2)}, g_n^{I(1)})$ and negative within $(g_n^{I(1)}, +\infty)$. Thus, we have $P_n^I(g_n^I)$ is decreasing within $[0, g_n^{I(2)}]$, increasing within $(g_n^{I(2)}, g_n^{I(1)})$ and decreasing within $(g_n^{I(1)}, +\infty)$. So the maximum of $\frac{\partial P_n^I}{\partial g_n^I}$ may be $g_n^I = 0$ or $g_n^I = g_n^{I(1)}$. This is proved in Appendix C.

We define a threshold price $p_n^{I\dagger}$ to identify which one is the maximum point: when $p^I \leq p_n^{I\dagger}$, the maximum point is $g_n^{I*} = g_n^{I(1)}$, otherwise, the maximum point is $g_n^{I*} = 0$.

Substitute the two points into $P_n^I(g_n^I)$ and the threshold price $p_n^{I\dagger}$ is obtained by solving:

$$P_n^I(0) = P_n^I(g_n^{I(1)}(p_n^{I\dagger})) \quad (38)$$

$$p_n^{I\dagger} d_n \lambda_n - \frac{3}{\sqrt[3]{4}} (p_n^{I\dagger} d_n)^{2/3} (\gamma_n^I)^{1/3} = \frac{-\gamma_n^I}{\lambda_n^2} \quad (39)$$

Simplifying:

$$\frac{\gamma_n^I}{\lambda_n^2} = \frac{3}{2^{2/3}} (\gamma_n^I)^{1/3} (d_n p^I)^{2/3} + p^I d_n \lambda_n \quad (40)$$

Solving this equation yields:

$$p_n^{I\dagger} = \frac{\gamma_n^I}{4d_n\lambda_n^3} \quad (41)$$

Thus, when $p^I \leq p_n^{I\dagger}$, the optimal point is $g_n^{I*}(p^I) = g_n^{I(1)} = d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \right)$, otherwise $g_n^{I*}(p^I) = 0$. We can reexpress it as

$$\begin{aligned} g_n^{I*}(p^I) &= \begin{cases} d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \right), & \text{if } p^I \leq p_n^{I\dagger}, \\ 0, & \text{otherwise,} \end{cases} \\ &= \mathbf{1}(p^I \leq p_n^{I\dagger}) d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \right). \end{aligned} \quad (42)$$

APPENDIX C

PROOF OF THE MONOTONICITY OF $\frac{\partial P_n^I}{\partial g_n^I}$

For ease of description, we denote $x = \frac{g_n^I}{d_n} - \lambda_n$, then the derivative can be written as:

$$\frac{\partial P_n^I}{\partial g_n^I} = \gamma_n^I \cdot \frac{2}{d_n x^3} - p^I \quad (43)$$

where x varies with g_n^I . The sign of the derivative depends on the sign and magnitude of x .

- 1) When $x < 0$, i.e., $g_n^I < d_n \lambda_n$, there is $x^3 < 0$, and $\frac{2}{d_n x^3} < 0$, thus $\gamma_n^I \cdot \frac{2}{d_n x^3} - p^I < 0$. The original function P_n^I decreases as g_n^I increases.
- 2) When $x > 0$, i.e., $g_n^I > d_n \lambda_n$, there is $x^3 > 0$, so $\frac{2}{d_n x^3} > 0$, thus $\gamma_n^I \cdot \frac{2}{d_n x^3} > 0$. The derivative $\frac{\partial P_n^I}{\partial g_n^I} = \gamma_n^I \cdot \frac{2}{d_n x^3} - p^I$, and its sign depends on the comparison between $\gamma_n^I \cdot \frac{2}{d_n x^3}$ and p^I . Define the critical point $x_c = \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3}$, corresponding to

$$g_c = d_n \lambda_n + d_n x_c = d_n \lambda_n + d_n \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3}. \quad (44)$$

- When $x < x_c$, i.e., $g_n^I < g_c$, there is $\gamma_n^I \cdot \frac{2}{d_n x^3} > p^I$, so the derivative $\frac{\partial P_n^I}{\partial g_n^I} > 0$. The original function P_n^I increases as g_n^I increases.
 - When $x > x_c$, i.e., $g_n^I > g_c$, there is $\gamma_n^I \cdot \frac{2}{d_n x^3} < p^I$, so the derivative $\frac{\partial P_n^I}{\partial g_n^I} < 0$. The original function P_n^I decreases as g_n^I increases.
- 3) When $x = 0$, i.e., $g_n^I = d_n \lambda_n$, the denominator in the derivative expression becomes zero, so the derivative is undefined.

In one word, the $\frac{\partial P_n^I}{\partial g_n^I}$ is negative within $[0, g_n^{I(2)}]$, positive within $(g_n^{I(2)}, g_n^{I(1)})$ and negative within $(g_n^{I(1)}, +\infty)$. Thus, we have $P_n^I(g_n^I)$ is decreasing within $[0, g_n^{I(2)}]$, increasing within $(g_n^{I(2)}, g_n^{I(1)})$ and decreasing within $(g_n^{I(1)}, +\infty)$. So the maximum of $\frac{\partial P_n^I}{\partial g_n^I}$ may be $g_n^I = 0$ or $g_n^I = g_n^{I(1)}$.

APPENDIX D PROOF OF EQUATION (17)

The cost function $C(\mathbf{x}^*(\mathbf{g}^*(\mathbf{p})))$ 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:

$$C(\mathbf{x}^*(\mathbf{g}^*(\mathbf{p}))) = c^{TH} D^T(p^T) + c^{IL} D^I(p^I) \quad (45)$$

- 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:

$$C(\mathbf{x}^*(\mathbf{g}^*(\mathbf{p}))) = c^{TH} G^H + c^{TL} (D^T(p^T) - G^H) + c^{IL} D^I(p^I) \quad (46)$$

- 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:

$$C(\mathbf{x}^*(\mathbf{g}^*(\mathbf{p}))) = c^{TH} D^T(p^T) + c^{IL} G^L + c^{IH} (D^I(p^I) - G^L) \quad (47)$$

- 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 E 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:

$$p^T \leq \frac{\gamma_n^T}{\theta_n} \quad (48)$$

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

$$\mathcal{C}_{p,q}^T = \left\{ n \in \mathcal{N} : \frac{\gamma_n^T}{\theta_n} \geq a_p^T \right\} \quad (49)$$

For inference services, client n has positive demand when:

$$p^I \leq p_n^{I\dagger} \quad (50)$$

The active inference clients are:

$$\mathcal{C}_{p,q}^I = \left\{ n \in \mathcal{N} : p_n^{I\dagger} \geq a_q^I \right\} \quad (51)$$

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

APPENDIX F 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:

$$\sum_{n \in \mathcal{C}_{p,q}^T} \theta_n \left(\frac{\gamma_n^T}{p^T \theta_n} - 1 \right) = G^H \quad (52)$$

$$\sum_{n \in \mathcal{C}_{p,q}^T} \left(\frac{\gamma_n^T}{p^T} - \theta_n \right) = G^H \quad (53)$$

$$\frac{1}{p^T} \sum_{n \in \mathcal{C}_{p,q}^T} \gamma_n^T - \sum_{n \in \mathcal{C}_{p,q}^T} \theta_n = G^H \quad (54)$$

$$\frac{1}{p^T} \sum_{n \in \mathcal{C}_{p,q}^T} \gamma_n^T = G^H + \sum_{n \in \mathcal{C}_{p,q}^T} \theta_n \quad (55)$$

$$a_p'^T = \frac{\sum_{n \in \mathcal{C}_{p,q}^T} \gamma_n^T}{G^H + \sum_{n \in \mathcal{C}_{p,q}^T} \theta_n} \quad (56)$$

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

$$\sum_{n \in \mathcal{C}_{p,q}^I} d_n \left(\lambda_n + \left(\frac{2\gamma_n^I}{d_n p^I} \right)^{1/3} \right) = G^L \quad (57)$$

$$\sum_{n \in \mathcal{C}_{p,q}^I} d_n \lambda_n + \sum_{n \in \mathcal{C}_{p,q}^I} d_n^{2/3} (2\gamma_n^I)^{1/3} p^{-1/3} = G^L \quad (58)$$

$$p^{-1/3} = \frac{G^L - \sum_{n \in \mathcal{C}_{p,q}^I} d_n \lambda_n}{\sum_{n \in \mathcal{C}_{p,q}^I} d_n^{2/3} (2\gamma_n^I)^{1/3}} \quad (59)$$

$$a_q'^I = \left(\frac{\sum_{n \in \mathcal{C}_{p,q}^I} d_n^{2/3} (2\gamma_n^I)^{1/3}}{G^L - \sum_{n \in \mathcal{C}_{p,q}^I} d_n \lambda_n} \right)^3 \quad (60)$$

APPENDIX G

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:

$$C(\mathbf{x}^*(\mathbf{g}^*(\mathbf{p}))) = A_{p,q} D_{p,q}^T(p^T) + B_{p,q} D_{p,q}^I(p^I) + E_{p,q}, \quad (61)$$

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:

$$R(\mathbf{p}, \mathbf{g}^*(\mathbf{p})) = p^T D_{p,q}^T(p^T) + p^I D_{p,q}^I(p^I) \quad (62)$$

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

$$A_{p,q} = c^{TH}$$

$$B_{p,q} = c^{IL}$$

$$E_{p,q} = 0$$

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

$$A_{p,q} = c^{TL}$$

$$B_{p,q} = c^{IL}$$

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

Case 3: $D^T(p^T) \leq G^H$ 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:

$$R(\mathbf{p}, \mathbf{g}^*(\mathbf{p})) = p^T D_{p,q}^T(p^T) + p^I D_{p,q}^I(p^I) \quad (63)$$

Therefore, the profit function becomes:

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

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

APPENDIX H PROOF OF THEOREM 3

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

A. For p^I :

The derivative with respect to p^I is:

$$\frac{\partial F_{p,q}}{\partial p^I} = D_{p,q}^I(p^I) + p^I \frac{\partial D_{p,q}^I}{\partial p^I} - B_{p,q} \frac{\partial D_{p,q}^I}{\partial p^I} \quad (64)$$

Since $D_{p,q}^I(p^I)$ is decreasing with 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:

$$p_{p,q}^{I*} = a_{q+1}^I \quad (65)$$

B. For p^T :

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

$$\frac{\partial F_{p,q}}{\partial p^T} = D_{p,q}^T(p^T) + (p^T - A_{p,q}) \frac{\partial D_{p,q}^T}{\partial p^T} = 0 \quad (66)$$

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.

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

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

The objective function with respect to p^T is:

$$F_{p,q}(p^T) = p^T D_{p,q}^T(p^T) - A_{p,q} D_{p,q}^T(p^T) \quad (67)$$

where

$$\begin{aligned} D_{p,q}^T(p^T) &= \sum_{n \in C_{p,q}^T} \theta_n \left(\frac{\gamma_n^T}{p^T \theta_n} - 1 \right) \\ &= \frac{1}{p^T} \sum_{n \in C_{p,q}^T} \gamma_n^T - \sum_{n \in C_{p,q}^T} \theta_n \end{aligned} \quad (68)$$

Let $\Gamma = \sum_{n \in C_{p,q}^T} \gamma_n^T$ and $\Theta = \sum_{n \in C_{p,q}^T} \theta_n$, then:

$$D_{p,q}^T(p^T) = \frac{\Gamma}{p^T} - \Theta \quad (69)$$

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

$$\begin{aligned} F_{p,q}^T(p^T) &= p^T \left(\frac{\Gamma}{p^T} - \Theta \right) - A_{p,q} \left(\frac{\Gamma}{p^T} - \Theta \right) \\ &= \Gamma - \Theta p^T - \frac{A_{p,q} \Gamma}{p^T} + A_{p,q} \Theta \end{aligned} \quad (70)$$

Taking the derivative with respect to p^T :

$$\frac{\partial F_{p,q}^T}{\partial p^T} = -\Theta + \frac{A_{p,q} \Gamma}{(p^T)^2} \quad (71)$$

Setting the derivative to zero to find the critical point:

$$-\Theta + \frac{A_{p,q} \Gamma}{(p^T)^2} = 0 \quad (72)$$

$$\frac{A_{p,q} \Gamma}{(p^T)^2} = \Theta \quad (73)$$

$$p^T = \sqrt{\frac{A_{p,q} \Gamma}{\Theta}} \quad (74)$$

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

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

$$\frac{\partial^2 F_{p,q}^T}{\partial (p^T)^2} = -\frac{2A_{p,q} \Gamma}{(p^T)^3} < 0 \quad (\text{since } A_{p,q} > 0, \Gamma > 0, p^T > 0) \quad (75)$$

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:

$$\sum_{n=1}^N (g_n^T(\mathbf{p}) + g_n^I(\mathbf{p})) \leq G \quad (76)$$

Substituting the demand functions:

$$D_{p,q}^T(p^T) + D_{p,q}^I(p^I) \leq G \quad (77)$$

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

$$D_{p,q}^T(p^T) + D_{p,q}^I(a_{q+1}^I) \leq G \quad (78)$$

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

$$\left(\frac{\Gamma}{p^T} - \Theta \right) + D_{p,q}^I(a_{q+1}^I) \leq G \quad (79)$$

Solving for p^T :

$$\frac{\Gamma}{p^T} \leq G + \Theta - D_{p,q}^I(a_{q+1}^I) \quad (80)$$

$$p^T \geq \frac{\Gamma}{G + \Theta - D_{p,q}^I(a_{q+1}^I)} \quad (81)$$

Thus, the constraint boundary is:

$$b_p^T = \frac{\sum_{n \in C_{p,q}^T} \gamma_n^T}{G + \sum_{n \in C_{p,q}^T} \theta_n - D_{p,q}^I(a_{q+1}^I)} \quad (82)$$

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

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:

$$p_{p,q}^{T*} = \max(a_p^T, b_p^T) \quad (83)$$

- 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:

$$p_{p,q}^{T*} = \max(p'^T, b_p^T) \quad (84)$$

- 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:

$$p_{p,q}^{T*} = a_{p+1}^T \quad (85)$$

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.