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

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

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

Taking the first-order derivative,

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

Setting the derivative to zero,

$$\frac{\gamma_n^T}{\theta_n + g_n^T} = p^T. \tag{27}$$

Solving for g_n^T,

$$\theta_n + g_n^T = \frac{\gamma_n^T}{p^T}, \tag{28}$$

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

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

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

Therefore, the optimal training demand is

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

B. For the inference service

Client n solves

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

Taking the first-order derivative,

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

Setting the derivative to zero,

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

Solving for g_n^I,

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

and we denote the solution g_n^I of (35) by $g_n^{I(1)}$ and

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

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

$$P_n^I(0) = P_n^I(g_n^{I(1)}(p^{I\dagger})), \tag{37}$$

which is

$$p_n^{I\dagger} d_n\lambda_n - \frac{3}{\sqrt[3]{4}}(p_n^{I\dagger} d_n)^{2/3}(\gamma_n^I)^{1/3} = \frac{-\gamma_n^I}{\lambda_n^2}. \tag{38}$$

Continue simplifying (38) as

$$\frac{\gamma_n^I}{\lambda_n^2} = \frac{3}{2^{2/3}}(\gamma_n^I)^{1/3}(d_n p^I)^{2/3} + p^I d_n\lambda_n. \tag{39}$$

Solving this equation and we have

$$p_n^{I\dagger} = \frac{\gamma_n^I}{4d_n\lambda_n^3}. \tag{40}$$

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

$$\begin{aligned} g_n^{I*}(p^I) &= \begin{cases} d_n\left(\lambda_n + \left(\frac{2\gamma_n^I}{d_np^I}\right)^{\frac{1}{3}}\right), & \text{if } p^I \leq p_n^{I\dagger}, \\ 0, & \text{otherwise}, \end{cases} \\ &= \mathbf{1}(p^I \leq p_n^{I\dagger})d_n\left(\lambda_n + \left(\frac{2\gamma_n^I}{d_np^I}\right)^{\frac{1}{3}}\right). \end{aligned} \tag{41}$$

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

$$\frac{\partial P_n^I}{\partial g_n^I} = \gamma_n^I \cdot \frac{2}{d_nx^3} - p^I. \tag{42}$$

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 $\frac{2}{d_nx^3} < 0$, thus $\gamma_n^I \cdot \frac{2}{d_nx^3} - p^I < 0$. 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 $\frac{2}{d_nx^3} > 0$, thus $\gamma_n^I \cdot \frac{2}{d_nx^3} > 0$. The derivative $\frac{\partial P_n^I}{\partial g_n^I} = \gamma_n^I \cdot \frac{2}{d_nx^3} - p^I$, and its sign depends on the comparison between $\gamma_n^I \cdot \frac{2}{d_nx^3}$ and p^I. Define the critical point $x_c = \left(\frac{2\gamma_n^I}{d_np^I}\right)^{1/3}$, corresponding to

$$g_c = d_n\lambda_n + d_nx_c = d_n\lambda_n + d_n\left(\frac{2\gamma_n^I}{d_np^I}\right)^{1/3}. \tag{43}$$

   - When $x < x_c$, i.e., $g_n^I < g_c$, there is $\gamma_n^I \cdot \frac{2}{d_nx^3} > p^I$, 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 $\gamma_n^I \cdot \frac{2}{d_nx^3} < p^I$, 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 $C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p})))$ 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 and

$$C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) = c^{TH}D^T(p^T) + c^{IL}D^I(p^I). \tag{44}$$

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 and

$$C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) = c^{TH}G^H + c^{TL}(D^T(p^T) - G^H) + c^{IL}D^I(p^I). \tag{45}$$

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 and

$$C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) = c^{TH}D^T(p^T) + c^{IL}G^L + c^{IH}(D^I(p^I) - G^L). \tag{46}$$

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

$$p^T \leq \frac{\gamma_n^T}{\theta_n}. \tag{47}$$

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

$$\mathcal{C}_{p,q}^T = \left\{n \in \mathcal{N} : \frac{\gamma_n^T}{\theta_n} \geq a_p^T\right\}. \tag{48}$$

For inference services, client n has positive demand when

$$p^I \leq p_n^{I\dagger}. \tag{49}$$

The active inference clients are

$$\mathcal{C}_{p,q}^I = \left\{n \in \mathcal{N} : p_n^{I\dagger} \geq a_q^I\right\}. \tag{50}$$

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:

$$\sum_{n\in\mathcal{C}^T_{p,q}} \theta_n \left(\frac{\gamma^T_n}{p^T\theta_n} - 1\right) = G^H. \tag{51}$$

And we denote the solution p^T of (51) by ${a'}^T_p$ and

$${a'}^T_p = \frac{\sum\limits_{n\in\mathcal{C}^T_{p,q}} \gamma^T_n}{G^H + \sum\limits_{n\in\mathcal{C}^T_{p,q}} \theta_n}. \tag{52}$$

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

$$\sum_{n\in\mathcal{C}^I_{p,q}} d_n \left(\lambda_n + \left(\frac{2\gamma^I_n}{d_n p^I}\right)^{1/3}\right) = G^L \tag{53}$$

And we denote the solution p^I of (53) by ${a'}^I_q$ and

$${a'}^I_q = \left(\frac{\sum\limits_{n\in\mathcal{C}^I_{p,q}} d_n^{2/3}(2\gamma^I_n)^{1/3}}{G^L - \sum\limits_{n\in\mathcal{C}^I_{p,q}} d_n\lambda_n}\right)^3 \tag{54}$$

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

$$C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) = A_{p,q}D^T_{p,q}(p^T) + B_{p,q}D^I_{p,q}(p^I) + E_{p,q}, \tag{55}$$

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

$$R(\boldsymbol{p}, \boldsymbol{g}^*(\boldsymbol{p})) = p^T D^T_{p,q}(p^T) + p^I D^I_{p,q}(p^I). \tag{56}$$

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 $Q'_{p,q}$[3]:

Case 1: $D^T_{p,q}(p^T) \le G^H$ and $D^I_{p,q}(p^I) \le G^L$.

$$\begin{aligned} C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) &= c^{TH}D^T_{p,q}(p^T) + c^{IL}D^I_{p,q}(p^I), \\ A_{p,q} &= c^{TH}, \\ B_{p,q} &= c^{IL}, \\ E_{p,q} &= 0. \end{aligned}$$

Case 2: $D^T_{p,q}(p^T) > G^H$ and $D^I_{p,q}(p^I) \le G^L$.

$$\begin{aligned} C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) &= c^{TH}G^H + c^{TL}(D^T_{p,q}(p^T) - G^H) + c^{IL}D^I_{p,q}(p^I) \\ &= c^{TL}D^T_{p,q}(p^T) + c^{IL}D^I_{p,q}(p^I) + (c^{TH} - c^{TL})G^H, \\ A_{p,q} &= c^{TL}, \\ B_{p,q} &= c^{IL}, \\ E_{p,q} &= (c^{TH} - c^{TL})G^H. \end{aligned}$$

[3]Here, the Case 2 - 3 do not do not imply that the entire subregion satisfies the constraints (7b), as $D^T_{p,q}(p^T) + D^I_{p,q}(p^I) > G$ may still happen.

Case 3: $D^T_{p,q}(p^T) \le G^H$ and $D^I_{p,q}(p^I) > G^L$.

$$\begin{aligned} C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) &= c^{TH}G^H + c^{TL}(D^T_{p,q}(p^T) - G^H) + c^{IL}D^I_{p,q}(p^I) \\ &= c^{TH}D^T_{p,q}(p^T) + c^{IH}D^I_{p,q}(p^I) + (c^{IL} - c^{IH})G^L, \\ A_{p,q} &= c^{TH}, \\ B_{p,q} &= c^{IH}, \\ E_{p,q} &= (c^{IL} - c^{IH})G^L. \end{aligned}$$

Case 4: $D^T_{p,q}(p^T) > G^H$ and $D^I_{p,q}(p^I) > G^L$.

$$\begin{aligned} C(\boldsymbol{x}^*(\boldsymbol{g}^*(\boldsymbol{p}))) &= +\infty, \\ A_{p,q} &= +\infty, \\ B_{p,q} &= +\infty, \\ E_{p,q} &= +\infty. \end{aligned}$$

The revenue function remains

$$R(\boldsymbol{p}, \boldsymbol{g}^*(\boldsymbol{p})) = p^T D^T_{p,q}(p^T) + p^I D^I_{p,q}(p^I). \tag{57}$$

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 H
Proof of Theorem 3

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

A. For p^I:

The p^I related part of the objective function is

$$F_{p,q}(p^I) = p^I D^I_{p,q}(p^I) - B_{p,q}D^I_{p,q}(p^I), \tag{58}$$

where $D^I_{p,q}(p^I) = \sum_{n\in\mathcal{C}^I_{p,q}} d_n \left(\lambda_n + \left(\frac{2\gamma^I_n}{d_n p^I}\right)^{1/3}\right)$.

Let $\Lambda = \sum_{n\in\mathcal{C}^I_{p,q}} d_n\lambda_n$ and $\Upsilon = \sum_{n\in\mathcal{C}^I_{p,q}} d_n^{2/3}(2\gamma^I_n)^{1/3}$, then

$$D^I_{p,q}(p^I) = \Lambda + \Upsilon {p^I}^{-1/3}. \tag{59}$$

Substituting to the (58), we have

$$\begin{aligned} F^I_{p,q}(p^I) &= p^I(\Lambda + \Upsilon {p^I}^{-1/3}) - B_{p,q}(\Lambda + \Upsilon {p^I}^{-1/3}) \\ &= \Lambda p^I + \Upsilon {p^I}^{2/3} - B_{p,q}\Lambda - B_{p,q}\Upsilon {p^I}^{-1/3}. \end{aligned} \tag{60}$$

Taking the derivative with respect to p^I,

$$\frac{\partial F^I_{p,q}}{\partial p^I} = \Lambda + \frac{2}{3}\Upsilon {p^I}^{-1/3} + \frac{1}{3}B_{p,q}\Upsilon {p^I}^{-4/3}. \tag{61}$$

Now we analyze each term

- $\Lambda > 0$ (sum of positive terms).
- $\frac{2}{3}\Upsilon {p^I}^{-1/3} > 0$ (since $\Upsilon > 0$ and $p^I > 0$).
- $\frac{1}{3}B_{p,q}\Upsilon {p^I}^{-4/3} > 0$ (since $B_{p,q} > 0$).

All three terms are strictly positive for $p^I > 0$. Therefore,

$$\frac{\partial F^I_{p,q}}{\partial p^I} > 0 \quad \text{for all } p^I > 0. \tag{62}$$

This proves that $F_{p,q}(\boldsymbol{p})$ is strictly increasing with respect to p^I within each subregion $Q'_{p,q}$. Therefore, the optimal p^I is at the upper boundary of the subregion

$$p^{I*}_{p,q} = a^I_{q+1}. \tag{63}$$

B. For p^T:

The p^T related part of the objective function is

$$F_{p,q}(p^T) = p^T D^T_{p,q}(p^T) - A_{p,q} D^T_{p,q}(p^T) \tag{64}$$

where

$$\begin{aligned} D^T_{p,q}(p^T) &= \sum_{n\in\mathcal{C}^T_{p,q}} \theta_n \left(\frac{\gamma^T_n}{p^T\theta_n} - 1\right) \\ &= \frac{1}{p^T}\sum_{n\in\mathcal{C}^T_{p,q}} \gamma^T_n - \sum_{n\in\mathcal{C}^T_{p,q}} \theta_n. \end{aligned} \tag{65}$$

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

$$D^T_{p,q}(p^T) = \frac{\Gamma}{p^T} - \Theta. \tag{66}$$

Substituting to the (64), we have

$$\begin{aligned} F^T_{p,q}(p^T) &= p^T\left(\frac{\Gamma}{p^T} - \Theta\right) - A_{p,q}\left(\frac{\Gamma}{p^T} - \Theta\right) \\ &= \Gamma - \Theta p^T - \frac{A_{p,q}\Gamma}{p^T} + A_{p,q}\Theta. \end{aligned} \tag{67}$$

Taking the derivative with respect to p^T,

$$\frac{\partial F^T_{p,q}}{\partial p^T} = -\Theta + \frac{A_{p,q}\Gamma}{(p^T)^2}. \tag{68}$$

The $F^T_{p,q}(p^T)$ is not monotonic with p^T.

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

Setting the derivative to zero to find the critical point,

$$-\Theta + \frac{A_{p,q}\Gamma}{(p^T)^2} = 0. \tag{69}$$

And we have

$$p^T = \sqrt{\frac{A_{p,q}\Gamma}{\Theta}}. \tag{70}$$

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

Next, we check the second derivative

$$\frac{\partial^2 F^T_{p,q}}{\partial (p^T)^2} = -\frac{2A_{p,q}\Gamma}{(p^T)^3} < 0 \quad (\text{since } A_{p,q} > 0, \Gamma > 0, p^T > 0). \tag{71}$$

Thus, p'^T is indeed a maximum point of the objective function. Solving this equation gives the critical point p'^T. The optimal solution depends on the position of p'^T relative to the subregion boundaries $[a^T_p, a^T_{p+1}]$.

The three cases in the theorem cover all possible scenarios:

1) If $p'^T \le a^T_p$, the function is decreasing in the subregion, so the optimal is at the lower boundary a^T_p.
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 a^T_{p+1}.

After that, we identify the optimal p^T in each subregion. However, we still not consider the constraint (7b). For a subregion $Q'_{p,q}$, there may be some points satisfied the constraint but the others not. Thus, we need check the impact of (7b) to the closed-form solution.

Since the optimal p^I is the upper bound of p^I in the subregion, considering the $D^I_{p,q}(p^I)$ is deceasing to p^I, if there exists feasible points in this subregion, the upper bound is also the best choice to satisfy the constraint. Thus, the solution of p^I is not affected by the constraint (7b). Next, we will identify the constraint's impact to the p^T.

*2) **Derivation of b^T_p (constraint boundary for p^T):***

The resource constraint from (7b) is

$$\sum_{n=1}^{N} (g^T_n(\boldsymbol{p}) + g^I_n(\boldsymbol{p})) \le G. \tag{72}$$

Substituting the $D^T_{p,q}(p^T)$ and $D^I_{p,q}(p^I)$ to it, we have

$$D^T_{p,q}(p^T) + D^I_{p,q}(p^I) \le G. \tag{73}$$

We already have $p^{I*}_{p,q} = a^I_{q+1}$ from the monotonicity analysis. Substituting and we have

$$D^T_{p,q}(p^T) + D^I_{p,q}(a^I_{q+1}) \le G. \tag{74}$$

Using the expression for $D^T_{p,q}(p^T)$,

$$\left(\frac{\Gamma}{p^T} - \Theta\right) + D^I_{p,q}(a^I_{q+1}) \le G. \tag{75}$$

Solving for p^T, we have

$$p^T \ge \frac{\Gamma}{G + \Theta - D^I_{p,q}(a^I_{q+1})}. \tag{76}$$

Thus, the constraint boundary is

$$b^T_p = \frac{\sum_{n\in\mathcal{C}^T_{p,q}} \gamma^T_n}{G + \sum_{n\in\mathcal{C}^T_{p,q}} \theta_n - D^I_{p,q}(a^I_{q+1})}, \tag{77}$$

where $D^I_{p,q}(a^I_{q+1}) = \sum_{n\in\mathcal{C}^I_{p,q}} d_n \left(\lambda_n + \left(\frac{2\gamma^I_n}{d_n a^I_{q+1}}\right)^{1/3}\right)$.

In a word, to satisfy (7b), the solution p^T in each subregion must be no less than b^T_p.

*3) **Optimal solution selection:***

Based on the position of p'^T relative to the subregion boundaries $[a^T_p, a^T_{p+1})$ and the constraint boundary b^T_p, we rewrite three cases:

- Case 1: $p'^T < a^T_p$.
  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

$$p^{T*}_{p,q} = \max(a^T_p, b^T_p). \tag{78}$$

- 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
$$p_{p,q}^{T*} = \max(p'^T, b_p^T). \tag{79}$$
- 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
$$p_{p,q}^{T*} = a_{p+1}^T. \tag{80}$$
  In this case, we don't need to consider b_p^T explicitly because we've already verified the feasibility of the subregion (At least the point (a_{p+1}^T, a_{q+1}^I) satisfy the constraints, which means there must be $b_p^T \leq a_{p+1}^T$), which ensures that a_{p+1}^T satisfies the constraint.

This completes the proof of Theorem 3.