APPENDIX A
PROOF OF LEMMA 1

The problem is a standard linear program with a separable objective function and linear constraints. The optimal allocation follows from the relative costs of the two GPU types. We allocate high-performance GPUs to training tasks whenever possible, while assigning low-performance GPUs to inference tasks to achieve optimal matching. The piecewise linearity of C^* follows from the parametric nature of the linear program. Therefore, we can calculate the optimal total cost C^*.

Lemma 6: The minimal operational cost of ADC C^* depends on all clients' demand $(\mathbf{g^T}, \mathbf{g^I})$, which can be denoted by:

$$C^*(\mathbf{g^T}, \mathbf{g^I}) = \begin{cases} c^{TH}\sum_{n=1}^{N} g_n^T + c^{IL}\sum_{n=1}^{N} g_n^I, \\ \quad \text{if} \quad \sum_{n=1}^{N} g_n^T \le G^H \text{ and } \sum_{n=1}^{N} g_n^I \le G^L \\ c^{TH}G^H + c^{TL}(\sum_{n=1}^{N} g_n^T - G^H) + c^{IL}\sum_{n=1}^{N} g_n^I, \\ \quad \text{if} \quad \sum_{n=1}^{N} g_n^T > G^H \text{ and } \sum_{n=1}^{N} g_n^I \le G^L \\ c^{TH}\sum_{n=1}^{N} g_n^T + c^{IL}G^L + c^{IH}(\sum_{n=1}^{N} g_n^I - G^L), \\ \quad \text{if} \quad \sum_{n=1}^{N} g_n^T \le G^H \text{ and } \sum_{n=1}^{N} g_n^I > G^L \\ +\infty, \\ \quad \text{if} \quad \sum_{n=1}^{N} g_n^T > G^H \text{ and } \sum_{n=1}^{N} g_n^I > G^L \end{cases}. \tag{36}$$

Proof: For different cases in Lemma 1, we have:

- If $\sum_{n=1}^{N} g_n^T \le G^H$ and $\sum_{n=1}^{N} g_n^I \le G^L$, the optimal allocation is:

$$\begin{cases} x_n^{TH*} = g_n^T, & x_n^{TL*} = 0, \\ x_n^{IH*} = 0, & x_n^{IL*} = g_n^I. \end{cases} \tag{37}$$

and the minimal cost is:

$$C^*(\mathbf{g^T}, \mathbf{g^I}) = c^{TH}\sum_{n=1}^{N} g_n^T + c^{IL}\sum_{n=1}^{N} g_n^I. \tag{38}$$

- If $\sum_{n=1}^{N} g_n^T > G^H$ and $\sum_{n=1}^{N} g_n^I \le G^L$, the optimal allocation is:

$$\begin{cases} x_n^{TH*} = g_n^T \frac{G^H}{\sum_{n=1}^{N} g_n^T}, & x_n^{TL*} = g_n^T\left(1 - \frac{G^H}{\sum_{n=1}^{N} g_n^T}\right), \\ x_n^{IH*} = 0, & x_n^{IL*} = g_n^I. \end{cases} \tag{39}$$

and the minimal cost is:

$$C^*(\mathbf{g^T}, \mathbf{g^I}) = c^{TH}G^H + c^{TL}(\sum_{n=1}^{N} g_n^T - G^H) + c^{IL}\sum_{n=1}^{N} g_n^I. \tag{40}$$

- If $\sum_{n=1}^{N} g_n^T \le G^H$ and $\sum_{n=1}^{N} g_n^I > G^L$, the optimal allocation is:

$$\begin{cases} x_n^{TH*} = g_n^T, & x_n^{TL*} = 0, \\ x_n^{IH*} = g_n^I\left(1 - \frac{G^L}{\sum_{n=1}^{N} g_n^I}\right), & x_n^{IL*} = g_n^I \frac{G^L}{\sum_{n=1}^{N} g_n^I}. \end{cases} \tag{41}$$

and the minimal cost is:

$$C^*(\mathbf{g^T}, \mathbf{g^I}) = c^{TH}\sum_{n=1}^{N} g_n^T + c^{IL}G^L + c^{IH}(\sum_{n=1}^{N} g_n^I - G^L). \tag{42}$$

- If $\sum_{n=1}^{N} g_n^T > G^H$ and $\sum_{n=1}^{N} g_n^I > G^L$, we have:

$$\sum_{n=1}^{N} g_n^T + \sum_{n=1}^{N} g_n^I > G^H + G^L = G, \tag{43}$$

which is not satisfied with the constraint in (1).

■