원래 model : K blocks

$\boxed{1}\ \boxed{2}\ \boxed{3}\ \cdots\ \boxed{K}$

Model split : K_D blocks @ device : $1, 2, \cdots, K_D$

K_E blocks @ ES : $K, K-1, \cdots, K-K_E+1$

Spec : $1 \sim k$번째 block을 거쳤을 때 $\begin{cases} \text{평균 성능 } P_k \\ \text{entropy } E_k(*) \end{cases}$

B_k의 computation 양 : C_k

comp. resource $\begin{cases} \text{device} : u_D \\ \text{ES} : u_E \end{cases}$

B_k output hidden rep.의 size : S_k

Offloading : ① computing resource 제한

② entropy – confidence

③ Energy 소모 제한 (computing + comm.)

device → ES offloading 할 layer : $l_D(t) \in \{K-K_E, K-K_E+1, \cdots, K\}$

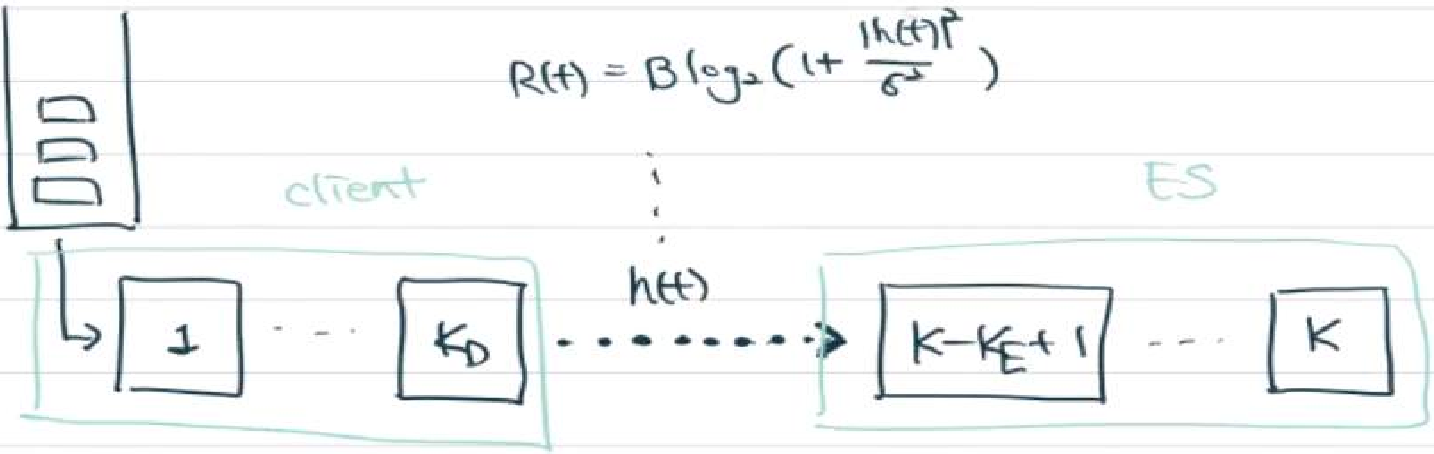
최종 출력 layer : $l_E(t) \in \{l_D(t)+1, \cdots, K\}$

offloading indicator : $b(t) \in \{0, 1\}$

Objective ① Delay minimization

② Accuracy maximization (Entropy minimization)

③ computation resource 소모.



Problem $\mathcal{L}_0$

- image가 B_1에 들어가는 순간 channel $h(t)$가 다 처리될 때까지 유지 — block fading
- Decision : $u_D(t)$, $l_D(t)$, $b(t)$, $l_E(t)$, $P(t)$
- Delay : $$\frac{\sum_{k=1}^{l_D(t)} c_k}{u_D(t)} + b(t)\cdot\left[\frac{S_{l_D(t)}}{R(t)} + \frac{\sum_{k=l_D(t)+1}^{l_E(t)} c_k}{u_E}\right] = \tau(n)$$
- Entropy : $H_{l_E(t)}(x)$

$K(u_D(t))^2 \sum_{k=1}^{l_D(t)} c_k$

- Energy : $E(t) = E_c(u_D(t), l_D(t)) + E_T(P(t))$

$P(t)\,\frac{S_{l_D(t)}}{R(P(t))}$

$$\min \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[\tau(n)]$$

$$\text{s.t.} \quad \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[H_{l_E(n)}(x_n)\right] \leq H_0$$

target accuracy

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[E(n)\right] \leq E_0$$

$$l_D(t) \in \{K-K_E, K-K_E+1, \cdots, K_D\}$$

$$l_E(t) \in \qquad \cdots, K\}$$

$$b(t) \in \{0, 1\}$$

$$u_D(t) \in \{0, 1, \cdots, \overline{u}_D\}$$

$$0 \leq P(t) \leq P_0$$

Virtual queues :

① $Z(t+1) = \max\{Z(t) + H_{l_E(n)}(x_n) - H_0, 0\}$

② $Y(t+1) = \max\{Y(t) + E(n) - E_0, 0\}$

Define $\Theta(t) = [Z(t), Y(t)]$

$$L(t) = \frac{1}{2}\left(Z(t)^2 + Y(t)^2\right), \quad \Delta(t) = \mathbb{E}\left[L(t+1) - L(t) \mid \Theta(t)\right]$$

$$L(t+1) - L(t) = \frac{1}{2}\left(Z(t+1)^2 - Z(t)^2 + Y(t+1)^2 - Y(t)^2 \right)$$

$$\leq Z(n)\left(H_{l_E(n)}(x_n) - H_0\right) + \frac{1}{2}\left(H_{l_E(n)}(x_n) - H_0\right)^2$$

$$+ Y(n)\left(E(n) - E_0\right) + \frac{1}{2}\left(E(n) - E_0\right)^2$$

DPP

$$\text{minimize} \quad \Delta(t) + V \cdot \mathbb{E}\left[\tau(n) \mid \Theta(n)\right]$$

$$\mathbb{E}\left[Z(n) H_{l_E(n)}(x_n)\right] + \mathbb{E}\left[Y(n) E(n)\right] + V \cdot \mathbb{E}\left[\tau(n) \mid \Theta(n)\right]$$

$$\boxed{\min. \quad Z(n) H_{l_E(n)}(x_n) + Y(n) E(n) + V \cdot \tau(n)}$$

$$\min. \quad Z(n) H_{l_E(n)}(x_n) + Y(n)\left\{ \kappa u_D(t)^2 \sum_{k=1}^{l_D(t)} C_k + p(t)\, b(t) \cdot \frac{S_{l_D}(t)}{R(p(t))} \right\}$$

$$+ V \cdot \left[\frac{\sum_{k=1}^{l_D(t)} C_k}{u_D(t)} + b(t) \cdot \left\{ \frac{S_{l_D}(t)}{R(p(t))} + \frac{\sum_{k=l_D(t)+1}^{l_E(t)} C_k}{u_E} \right\} \right]$$

$$\text{s.t.} \quad l_D(t) \in \langle K - K_E, K - K_E + 1, \cdots, K \rangle$$

$$l_E(t) \in \langle l_D(t) + 1, \cdots, K \rangle$$

$$b(t) \in \langle 0, 1 \rangle$$

$$u_D(t) \in \langle 0, 1, \cdots, \overline{u_D} \rangle$$

$$0 \leq p(t) \leq P_0$$

1번째 → $F_1(x)$

2번째 →

p,l,b