다계층 컴퓨팅 시스템을 위한 작업 분할 및 셀탈피 전송의 통합 최적화

김도곤, 박석환
전북대학교

kdg8728@jbnu.ac.kr, seokhwan@jbnu.ac.kr

Joint Optimization of Task Splitting and Cell-Free Transmission for Multi-Tier Computing Systems

Dogon Kim, Seok-Hwan Park
Jeonbuk National University

요 약

본 논문은 제한된 계산 자원을 갖고 있는 Internet of Things Device(ID)의 Local Server(LS)에서 계산하기 힘든 일들을 하위 작업으로 분할한 뒤 Edge Server(ES)와 Cloud Server(CS)로 보내 처리하는 방식을 다룬다. 완료 시간과 에너지 소모를 최소화하기 위해 작업 분할 비율과 전송 전략을 공동으로 최적화한다. 수치 결과는 제안된 다중 계층 계산 시스템의 효과를 입증한다.

Ⅰ. 서 론

저지연 통신 및 처리는 ID 의 제한된 컴퓨팅 능력에 의해 어려움을 겪게 된다. 이 문제에 대한 해결책으로 계산 작업을 ID 들에서 ES 나 CS 로 오프로딩하는 방식을 제안한다. 작업 오프로딩은 ID 에서 수행되는 계산 작업을 부분적으로 ES 와 CS 에서 수행하게 하여 전체 실행 시간과 에너지 소모를 최소화할 수 있다. 모든 ID 들의 작업 분할 비율을 각각 최적화하고 ID 에서 Edge Node(EN)로의 무선 액세스 링크 및 EN 에서 Cloud Processor(CP)로의 프론트홀 링크에서의 전송 전략을 함께 최적화한다.

Ⅱ. 시스템 모델 및 최적화 문제

모두 한 개의 안테나를 가지고 있고 ID 들과 EN 들이 각각 N_I, N_E 개씩 무작위로 배치되어 있는 환경을 고려한다. ID 와 EN 의 집합을 각각 $\mathcal{N}_I = \{1,2,\ldots,N_I\}$ 와 $\mathcal{N}_E = \{1,2,\ldots,N_E\}$ 로 정의하였다. k 번째 ID 의 LS 는 $f_{L,k}$ cycles/s 의 속도로 작업을 수행한다고 가정하였고 $f_{L,k} \in [0, F_{L,k}]$ 을 만족하며 $F_{L,k}$ 는 LS 의 최대 계산 속도를 나타낸다. 동일하게 ES 와 CS 의 계산 속도는 각각 $f_{E,i}$ 와 f_C 이며 $f_{E,i} \in [0, F_{E,i}]$ 와 $f_C \in [0, F_C]$ 을 만족한다. 본 논문에서는 모든 ES 와 CS 는 오프로딩된 작업들을 계산하기에 충분한 에너지를 갖고 있다고 가정한다. 따라서 시스템 설계에서 에너지를 많이 소모하는 ID 의 에너지 소비만을 고려하여 $f_{E,i} = F_{E,i}, i \in \mathcal{N}_E$ 와 $f_C = F_C$ 로 고정한다. k 번째 ID 의 작업은 $T_k = (b_k, V_k)$ 로 나타낼 수 있다. 여기서 b_k 는 입력 데이터의 비트 수이고 V_k 는 비트당 필요한 CPU cycle 수이다. ID 는 한 개의 EN 과 통신하는데 이를 나타내는 파라미터가 $\{i_k\}_{k\in\mathcal{N}_I}$, $i_k \in \mathcal{N}_E$ 이다. EN i 와 연결된 ID 들의 집합은 $\mathcal{N}_{I,i} = \{k | k \in \mathcal{N}_I, i_k = i\}$ 로 정의된다. EN 는 CP 로 오프로딩할 때 프론트홀 링크를 이용하는데 이때 용량은 C_F bit/s/Hs 이다. EN i 의 수신 신호는 $y_i = \sum_{k\in\mathcal{N}_I} \mathbf{h}_{i,k} x_k + \mathbf{z}_i$ 이다. x_k 는 k 번째 ID 의 수신 신호이고, $\mathbf{h}_{i,k} \in \mathbb{C}^{n_E \times 1}$ 은 k 번째 ID 와 EN i 의 채널 벡터이다. $\mathbf{z}_k \sim \mathcal{CN}(\mathbf{0}, \sigma_z^2 \mathbf{I})$ 는 잡음 신호이다. 각 수신 신호는 $\mathbb{E}\left[|x_k|^2\right] \le P_{tx}$ 을 만족한다.

Ⅲ. 다중 계층 계산

A.작업 분할과 계산 모델

k 번째 ID 에 대한 LS, ES, CS 각각의 작업 분할 비율을 $\boldsymbol{\alpha}_k = [\alpha_{L,k}, \alpha_{E,k}, \alpha_{C,k}]$ 로 나타낸다. $\alpha_{L,k} + \alpha_{E,k} + \alpha_{C,k} = 1$ 이고 $\alpha_{X,k} \ge 0, X \in \{L, E, C\}$ 이다. 하위 작업은 $T_{L,k} = (\alpha_{L,k} b_k, V_k), T_{E,k} = (\alpha_{E,k} b_k, V_k), T_{C,k} = (\alpha_{C,k} b_k, V_k)$ 로 정의된다. k 번째 ID 의 LS 에서의 계산 시간을 구하면 $\tau_{L,k}^{exe} = \alpha_{L,k} b_k V_k / f_{L,k}$ 이고 에너지 소모는 $E_k^{exe} = \varsigma_k \alpha_{L,k} b_k V_k f_{L,k}^2$ 이다. 여기서 ς_k 는 칩 아키텍처에 따라 달라지는 효과적인 스위치 용량이다 [1]. $\tau_{E,k}^{exe} = \alpha_{E,k} b_k V_k / f_{E,k}, \tau_{C,k}^{exe} = \alpha_{C,k} b_k V_k / f_{C,k}$ 이다. k 번째 ID 는 가우시안 채널 코딩을 사용한 속도 분할 통신을 이용한다. 따라서 $x_k = \sqrt{p_{E,k}} s_{E,k} + \sqrt{p_{C,k}} s_{C,k}$ 이고, 여기서 $s_{E,k} \sim \mathcal{CN}(0,1)$ 와 $s_{C,k} \sim \mathcal{CN}(0,1)$ 는 각각 $T_{E,k}$ 와 $T_{C,k}$ 의 입력 데이터를 부호화 하는 데이터

심볼이다. $p_{E,k} \geq 0$와 $p_{C,k} \geq 0$는 전력 제어 변수를 나타내고 $p_{E,k} + p_{C,k} \leq P_{tx}$를 만족한다. $s_{E,k}$와 $s_{C,k}$는 각각 EN i_k와 CP에서 디코딩 된다. EN에서 신호를 복호화 할 때 successive interference cancellation(SIC) 복호화 하는 걸 가정한다. 복호화 순서는 $s_{E,\pi_i(1)} \rightarrow s_{E,\pi_i(2)} \rightarrow \ldots \rightarrow s_{E,\pi_i(N_{I,i})}$이고 순열 $\pi_i : \{1, \ldots, N_{I,i}\}$이다. 신호 $s_{E,\pi_i(k)}$의 최대 데이터 속도는 $R_{E,\pi_i(k)} = r_{E,\pi_i(k)}(\mathbf{p})$

$= \log_2 \det(\mathbf{I} + p_{E,\pi_i(k)} \mathbf{N}^{-1}_{E,i,\pi_i(k)} \mathbf{h}_{i,\pi_i(k)} \mathbf{h}^H_{i,\pi_i(k)})$

이다. 여기서 $\mathbf{p} = \{p_{E,k}, p_{C,k}\}_{k \in \mathcal{N}_I}$이고, $\mathbf{N}_{E,i,\pi_i(k)}$은 양자화 잡음 공분산 행렬이다. $s_{E,k}$를 복호화 한 후 EN i는 복호화 된 신호를 뺀 나머지 신호를 양자화를 한 뒤 프론트홀 링크를 통해 CP로 보낸다. 양자화된 신호는 다음과 같다. $\tilde{\mathbf{y}}_i \leftarrow \mathbf{y}_i - \sum_{k \in \mathcal{N}_{I,i}} \mathbf{h}_{i,k} \sqrt{p_{E,k}} s_{E,k}$. CP는 수신된 양자화된 신호를 복구하여 다음과 같이 얻을 수 있다. $\hat{\mathbf{y}}_i = \tilde{\mathbf{y}}_i + \mathbf{q}_i$, 여기서 $\mathbf{q}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i)$는 양자화 과정에서 발생하는 왜곡 신호로 가우시안 코드북을 사용하여 압축한다. EN에서 CP까지의 프론트홀 전송시간은 $\tau_F^{tx} = \max_{i \in \mathcal{N}_E} \tau_W^{tx} \cdot g_i(\mathbf{p}, \mathbf{\Omega}_i) / C_F$이다.

$g_i(\mathbf{p}, \mathbf{\Omega}_i)$는 양자화된 신호 $\hat{\mathbf{y}}_i$를 표현하는 데 사용되는 샘플 당 비트 수이다. CP에서도 EN와 동일하게 받은 신호를 복호화 할 때 SIC 복호화를 한다. 신호 $s_{C,\pi_C(k)}$의 최대 데이터 속도는

$R_{C,\pi_C(k)} = r_{C,\pi_C(k)}(\mathbf{p}, \mathbf{\Omega})$

$= \log_2 \det(\mathbf{I} + p_{C,\pi_C(k)} \mathbf{N}^{-1}_{C,\pi_C(k)} \mathbf{h}_{\pi_C(k)} \mathbf{h}^H_{\pi_C(k)})$ 이다.

$\mathbf{N}_{C,\pi_C(k)}$은 간섭 및 잡음 항의 공분산 행렬이고 $\mathbf{h}_k = [\mathbf{h}^H_{1,k} \cdots \mathbf{h}^H_{N_E,k}]^H$, $\bar{\mathbf{\Omega}} = blkdiag(\{\mathbf{\Omega}_i\}_{i \in \mathcal{N}_E})$, $\mathbf{g}_{i,k} = \mathbf{h}_{i,k} \cdot 1(i \neq i_k)$, $\mathbf{g}_k = [\mathbf{g}^H_{1,k} \cdots \mathbf{g}^H_{N_E,k}]$이다.

따라서 무선 채널 통신 시간 τ_W^{tx}는 $\tau_W^{tx} = \max_{k \in \mathcal{N}_I} \{\alpha_{E,k} b_k / R_{E,k}, \alpha_{C,k} b_k / R_{C,k}\}$이고 k번째 ID가 무선 통신을 할 때 소모되는 에너지는 $E_k^{tx} = (p_{E,k} + p_{C,k}) \tau_W^{tx}$이다. 통신을 하는데 걸리는 총 시간 $\tau^{total} = \max\{\tau_L^{exe}, \tau_W^{tx} + \max\{\tau_E^{exe}, \tau_F^{tx} + \tau_C^{exe}\}\}$이다.

Ⅳ. 문제 정의와 최적화 문제

우리의 목표는 $\boldsymbol{\alpha}_k$, $\mathbf{f} = \{f_{L,k}, f_{E,k}, f_{C,k}\}_{k \in \mathcal{N}_I}$, $\mathbf{p}$, $\mathbf{\Omega}$을 공동으로 최적화하여 총 완료 시간 τ^{total}과 ID의 최대 에너지 소비 $E^{total} = \max_{k \in \mathcal{N}_I} (E_k^{exe} + E_k^{tx})$의 합을 최소화하는 것이다. 위 문제는 다음과 같이 공식화할 수 있다.

$$\min_{\substack{\mathbf{p},\mathbf{\Omega},\boldsymbol{\tau},\boldsymbol{\alpha},\mathbf{f},\mathbf{R},\mathbf{E}, \\ \boldsymbol{\theta},\boldsymbol{\Sigma},\boldsymbol{\gamma},\boldsymbol{\nu},\boldsymbol{\Phi},\boldsymbol{\beta}}} \tau^{total} + E^{total},$$

$$s.t.\ \alpha_{L,k} + \alpha_{E,k} + \alpha_{C,k} = 1,\ k \in \mathcal{N}_I,$$

$$\alpha_{X,k} \geq 0,\ X \in \{L, E, C\},\ k \in \mathcal{N}_I,$$

$$\tau_X^{exe} \geq \alpha_{X,k} b_k V_k / f_{X,k},\ X \in \{L, E, C\},\ k \in \mathcal{N}_I,$$

$$f_{L,k} \in [0, F_{L,k}],\ k \in \mathcal{N}_I,$$

$$\sum_{k \in \mathcal{N}_{I,i}} f_{E,k} = F_{E,i},\ i \in \mathcal{N}_E,\ f_{E,k} \geq 0,\ k \in \mathcal{N}_I,$$

$$\sum_{k \in \mathcal{N}_I} f_{C,k} = F_C,\ f_{C,k} \geq 0,\ k \in \mathcal{N}_I,$$

$$\tau_W^{tx} \geq \alpha_{X,k} b_k / R_{X,k},\ X \in \{E, C\},\ k \in \mathcal{N}_I,$$

$$\tau_F^{tx} \geq W \tau_W^{tx} \cdot g_i(\mathbf{p}, \mathbf{\Omega}_i) / C_F,\ i \in \mathcal{N}_E,$$

$$R_{E,\pi_i(k)} \leq r_{E,\pi_i(k)}(\mathbf{p}),\ i \in \mathcal{N}_E,\ k \in \{1, \ldots, N_{I,i}\},$$

$$R_{C,\pi_C(k)} \leq r_{C,\pi_C(k)}(\mathbf{p}, \mathbf{\Omega}),\ k \in \mathcal{N}_I,$$

$$p_{E,k} \geq 0,\ p_{C,k} \geq 0,\ p_{E,k} + p_{C,k} \leq P_{tx},\ k \in \mathcal{N}_I,$$

블록 조정 하강 접근법[2]을 사용하여 $\{\mathbf{p}, \mathbf{\Omega}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \mathbf{f}, \mathbf{R}, \mathbf{E}\}$와 $\{\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\Phi}\}$을 번갈아 최적화함으로써 최적해를 찾는다.

Ⅴ. 모의실험 결과

파라미터 $N_I = 8, N_E = 2, C_F = 10 bps/Hz$ 및 10dB의 SNR을 고려하여 최적화된 다중 계층 컴퓨팅 시스템의 완료 시간과 에너지 소모량의 합을 평가한다. 그래프는 고정된 LS 계산 속도 f_L에 대한 처리 속도 비율인 $r = f_E / f_L = f_C / f_E$을 증가시키며 값을 나타낸다. 그래프는 고정된 값을 갖는 LS를 제외하고는 r이 증가함에 따라 개선된 평균값을 얻는 것을 보여준다. 다중 계층 계산 시스템은 단일 계층 계산 시스템과 비교하여 완료 시간과 에너지 소모를 크게 줄이는 것을 관찰할 수 있다.

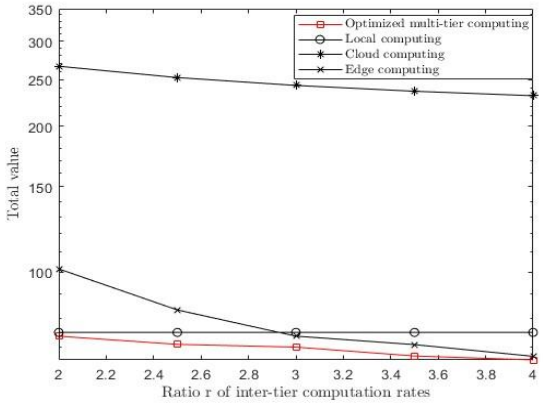

그림 1. 계산 속도 비율 r 대비 계산 기법 성능 비교

ACKNOWLEDGMENT

이 성과는 정부(교육부, 과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1A6A1A09031717, RS-2023-00238977).

참 고 문 헌

[1] D. V. Huynh and et al., "Joint communication and computation offloading for ultra-reliable and low-latency with multi-tier computing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 2, pp. 521-537, Feb. 2023.

[2] D. P. Bertsekas, *Nonlinear Programming*, 2nd edition, Athena Scientific, 1999.