Achieving High Throughput with Predictive Resource Allocation

Chuting Yao, Jia Guo, and Chenyang Yang
Beihang University, Beijing, China http://welcom.buaa.edu.cn

Main Works

- Investigate how to exploit the excess resources in the prediction window to improve the load carrying ability with three levels of context information.
- Formulate and solve a resource allocation planning problem to minimize the maximal transmission completion time, and introduce a method for making the transmission plan to help user scheduling.
- Gain of predictive resource allocation in supporting high arrival rate and reducing average waiting time is dramatic over non-predictive resource allocation.

System Model

- N_b-cell network, each equipped with N_t antennas. BSs serve two classes of traffic: real-time (RT) services and content delivery. User with content delivery traffic can only use the residual resources left by RT services. Design predictive allocation for the mobile user (MS) demanding content delivery: each requesting one file with size B bits.

- **Context information assumed known:**

◇ **Application level** : Requests arrival time and the files to be requested.
◇ **User level** : The trajectory of every MS and radio map.
◇ **Network level** : The average residual resources remained at each BS after serving the RT traffic.

- **Channel Model**

◇ The length of the prediction window is T_f frames. Each frame includes T_s time slots.
◇ The large scale channel gains remain constant within each frame and may vary among frames.
◇ The small scale channel gains remain constant within each time slot and vary among time slots independently.

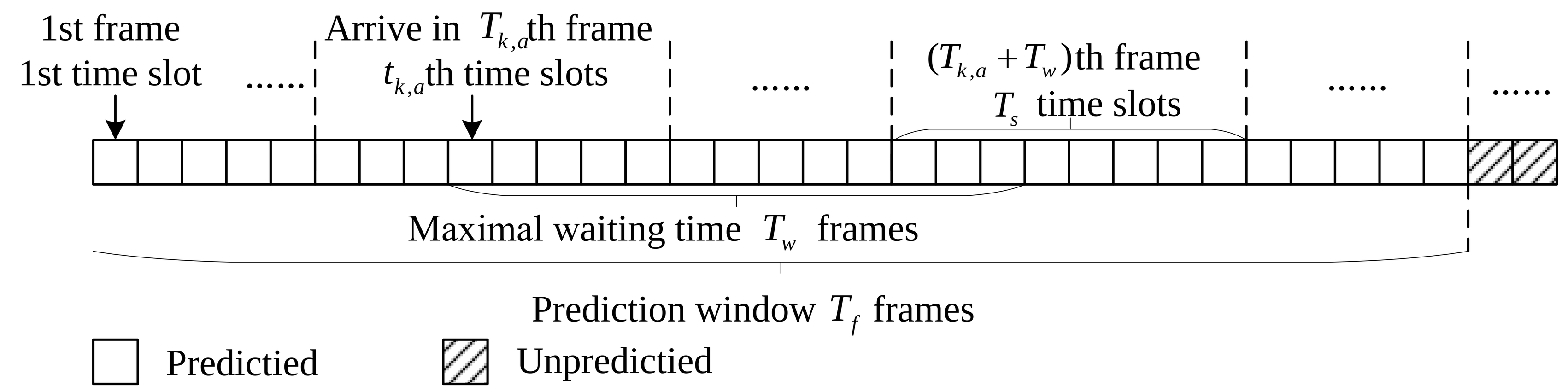
- **Achievable Rate**

$$R_{j,k}^t = W_{j,k}^t \log_2(1 + g_{j,k}^t p_{\max,j,k}^t)$$

◇ $W_{j,k}^t$ and $p_{\max,j,k}^t$ are the residual bandwidth and transmit power available for the MS at the closest BS after serving RT traffic.
◇ $g_{j,k}^t \triangleq \alpha_k^j \|\mathbf{h}_{j,k}^t\|^2/\sigma^2$ is the equivalent channel gain.

Predictive Resource Planning and Allocation

- **Making Resource Allocation Plan**:



request arrival time of MS$_k$: $J_{k,\mathrm{a}} \triangleq ((T_{k,\mathrm{a}}-1)T_s + t_{k,\mathrm{a}})/T_s$;
completion time of MS$_k$: $J_k \triangleq ((T_k-1)T_s + t_k)/T_s$.

◇ **Problem Formulation**:

$$\min_{\mathbf{M}_1,\ldots,\mathbf{M}_K} \quad \max_k \; J_k \tag{1}$$

$$\text{s.t.} \quad \sum_{j=1}^{T_f}\sum_{t=1}^{T_s} m_{j,k}^t R_{j,k}^t \Delta_t = B, \tag{2}$$

$$J_k \le J_{k,\mathrm{mw}} \triangleq \min(J_{k,\mathrm{a}} + T_{\mathrm{w}}, T_f), k = 1,\ldots,K, \tag{3}$$

$$m_{j,k}^t = 0, \forall j > T_k, m_{T_k,k}^t = 0, \forall t > t_k, \tag{4}$$

$$\sum_{k\in\mathcal{K}_{j,i}} m_{j,k}^t \le 1, i = 1,\ldots,N_b. \tag{5}$$

◇ (1): Minimize maximal completion time of all MSs, (2): B bits should be conveyed to each MS, (3): File should be conveyed in prediction window and maximal waiting time, (4): MS won't be served after completion time, (5): Each BS only transmits to one MS in each time slot.
◇ **Assumption**: small scale channel gains, residual bandwidth and transmit power are ergodic and $T_s \to \infty$.
◇ Define $s_{j,k} \triangleq \sum_{t=1}^{T_s} m_{j,k}^t/T_s$ as the fraction of time resource allocated to MS$_k$ at jth frame. Then $\lim_{T_s\to\infty}\sum_{j=1}^{T_f}\sum_{t=1}^{T_s} m_{j,k}^t R_{j,k}^t \Delta_t = \sum_{j=1}^{T_k} s_{j,k}\bar{R}_{j,k}T_s\Delta_t$.
◇ With statistical channel and network status information:

$$\min_{J,\mathbf{s}_1,\ldots,\mathbf{s}_K} \quad J$$

$$\text{s.t.} \quad \sum_{j=1}^{T_f} s_{j,k}\bar{R}_{j,k}T_s\Delta_t = B,$$

$$\sum_{k\in\mathcal{K}_{j,i}} s_{j,k} \le 1, i = 1,\ldots,N_b,$$

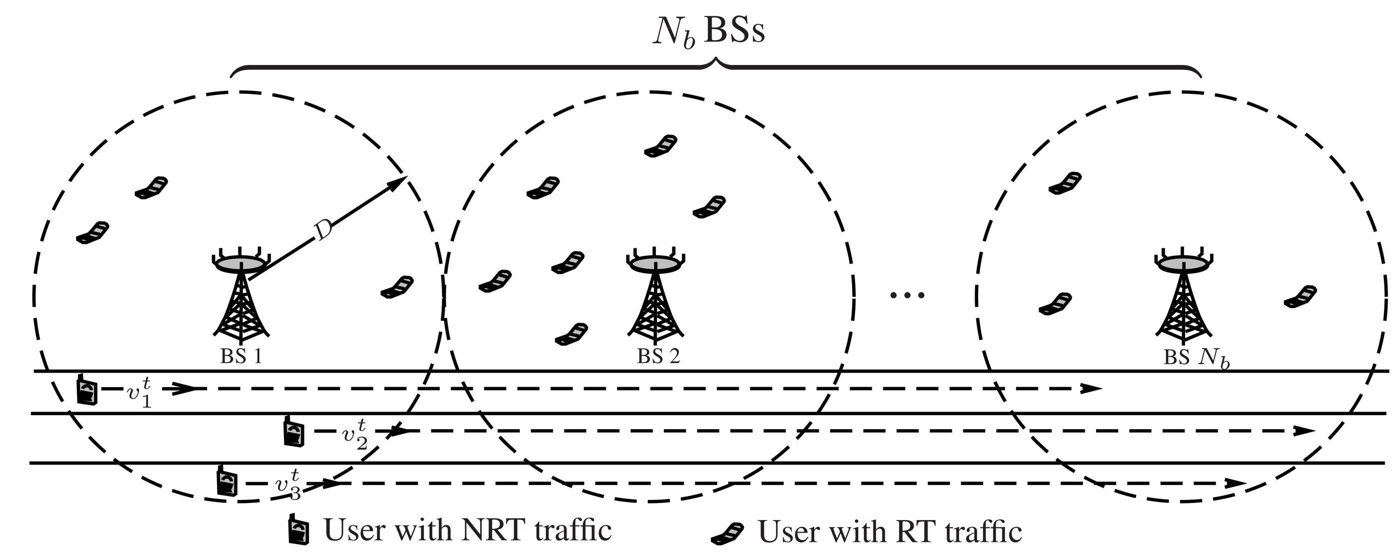
$$s_{j,k} = 0, \forall j > \lceil J_{k,\mathrm{mw}}\rceil, s_{\lceil J_{k,\mathrm{mw}}\rceil} \le J_{k,\mathrm{mw}} - \lceil J_{k,\mathrm{mw}}\rceil + 1,$$

$$s_{j,k} = 0, \forall j > \lceil J\rceil, s_{\lceil J\rceil,k} \le J - \lceil J\rceil + 1.$$

◇ **Transmission According to Plan**: In each time slot, the BSs will transmit file to MS who has not caught up transmission progress with best effort method.

Simulation Results

- **Simulation Setup**

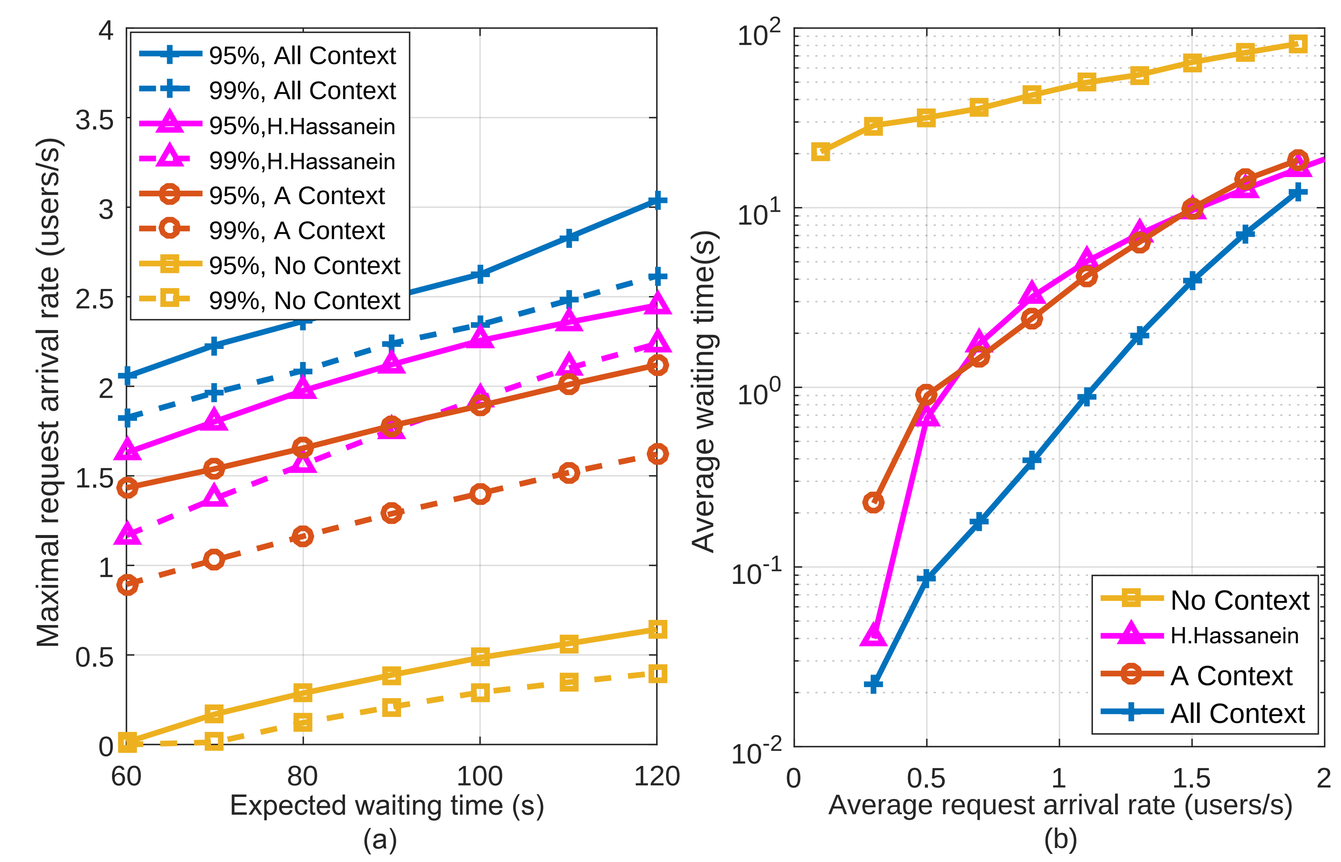


$N_b = 6$ BSs; $N_t = 6$; cell radius $D = 250$m, $P_{\max} = 40$W; $\bar{W}_{\mathrm{busy}} = 1$MHz; $\bar{W}_{\mathrm{idle}} = 10$MHz; $\Delta_t = 10$ms; $T_s = 100$; $B = 30$MBytes; $T_f = 300$ frames; $v \sim \mathcal{U}(2.5, 12.5)$m/s.

- **Transmission Strategies for Contrast**

◇ **All Context**: Proposed predictive resource allocation;
◇ **A Context**: BSs transmit to MSs with best effort no matter if requests arrive (use application level context information);
◇ **No Context**: BSs transmit to MSs with best effort after requests arrive;
◇ **H. Hassanein**: Transmission strategy obtained by minimizing total transmission time (proposed in [14]).

- **Maximal Carrying Traffic Load and Average Waiting Time**



When the user satisfaction rate is high (say 99%) and the expected waiting time is 120 s, the gain in terms of maximal carrying traffic load of "All Context" is 120% over "A Context" and 728% over the traditional transmission.