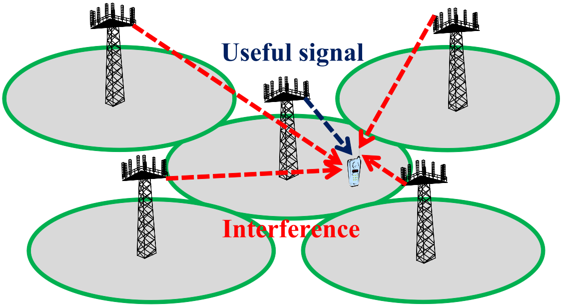
Interference coordination in wireless networks: a flow level perspective

Richard Combes ([1]), Zwi Altman ([2]) and Eitan Altman ([3])

[1]KTH, The Royal Institute of Technology
[2]Orange Labs
[3]INRIA

INFOCOM 2013

The model



- Wireless data network with flow level dynamics
- User arrival rate $\lambda(dr)$, average file size $\mathbb{E}[\sigma]$
- System state: number and locations of users + remaining file sizes
- Action: $\theta(t)$, transmitted powers and frequency allocation
- Objective: minimize the average user delay (proportional to $\sum_s \mathbb{E}[n_s(t)]$ by Little's law)

Related Work

- Optimize a function of the active users data rates ([1]). Convergence/optimality is hard to analyze.
- Throughput optimality: Max-weight scheduling at the flow-level ([2]), Interacting processors ([3]) . Average delay is hard to analyze, and throughput optimality depends on Poisson assumptions.
- Association problem: optimize a function of the loads ([4]).
- Traffic studies: wireless data traffic is not Poisson ...

---

[1] Stolyar et al , Self-Organizing Dynamic Fractional Frequency Reuse for Best-Effort Traffic through Distributed Inter-Cell Coordination , INFOCOM 2009

[2] Van de Ven et al, Spatial inefficiency of MaxWeight scheduling, Wiopt 2011

[3] Borst et al, Interacting queues with server selection and coordinated scheduling - application to cellular data networks, Annals of Operations Research 2009

[4] Kim et al, Distributed α-Optimal User Association and Cell Load Balancing in Wireless Networks, Trans. on Networking 2012

Proposed approach

- "Semi-static" approach:

$$\text{minimize } U(\overline{\rho}) = \sum_{s=1}^{N_s} u(\overline{\rho}_s).$$

- Learning procedure: arrival rates, network geometry and data rates are unknown.
- Parameters are tuned at a time scale $\approx 60s$: slower than arrivals/departures but faster than variations of arrival rates.
- Separability: distributed implementation is possible.

Queuing models

- Model for elastic traffic ([5])
- Round-robin scheduling: instantaneous throughput $R_s(r)/n_s$
- Station Load:

$$\overline{\rho}_s = \mathbb{E}\left[\sigma\right] \int_{\mathbb{A}_s} \frac{\lambda(dr)}{R_s(r)}$$

- Expected number of active users: $E[n_s] = \frac{\overline{\rho}_s}{1-\overline{\rho}_s}$.
- For elastic traffic, minimizing

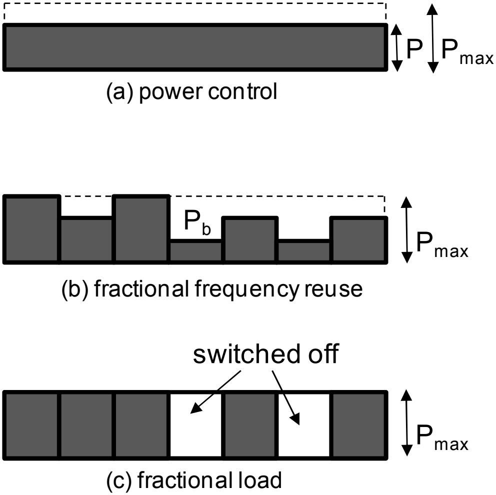
$$\sum_{s=1}^{N_s} \frac{\overline{\rho}_s}{1-\overline{\rho}_s}$$

is delay optimal.

- Similar model for streaming traffic ...

[5] Bonald et al, Wireless Downlink Data Channels: User Performance and Cell Dimensioning, Mobicom 2003

Interference coordination schemes



Fractional load and fractional frequency reuse

- Fractional load:

$$R_s(r) = \theta_s \mathbb{E}\left[\phi(S_s(r))\right],$$

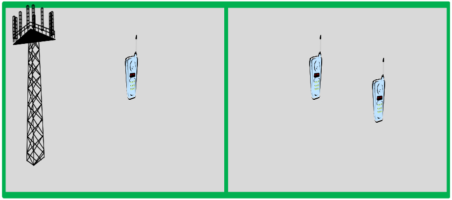
$$S_s(r) = \frac{P_{max} h_s(r)}{N_0^2 + \sum_{s' \neq s} P_{max} h_{s'}(r) X_{s'}}, \; X_{s'} \equiv \text{Bernouilli}(\theta_{s'}).$$

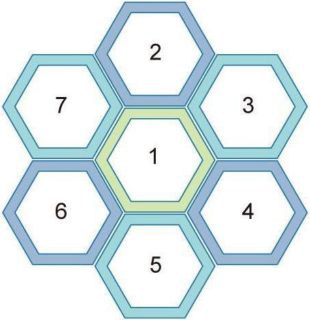
- Fractional frequency reuse:

$$R_s(r) = \sum_b \phi(S_{s,b}(r))$$

$$S_{s,b}(r) = \frac{\theta_{s,b} h_s(r)}{N_0^2 + \sum_{s' \neq s} \theta_{s',b} h_{s'}(r)}.$$

Soft frequency reuse





$$R_{s,edge}(r) = \sum_{b \in edge} \phi(S_{s,b}(r))$$

$$R_{s,center}(r) = \sum_{b \in center} \phi(S_{s,b}(r))$$

Each station is equivalent to two queues *in parallel*.

Load estimation

- Time is slotted, n-th slot $[nT, (n+1)T)$
- Load estimate (empirical workload):

$$\rho_s[k] = \frac{1}{T} \sum_{n \in \mathbb{Z}} \frac{\sigma_n}{R_s(r_n)} \mathbf{1}_{[kT,(k+1)T)}(T_n).$$

- Derivative estimate:

$$\nabla_\theta \rho_s[k] = -\frac{1}{T} \sum_{n \in \mathbb{Z}} \sigma_n \frac{\nabla_\theta R_s(r_n)}{R_s(r_n)^2} \mathbf{1}_{[kT,(k+1)T)}(T_n).$$

- Unbiased estimators: $\mathbb{E}\left[\rho_s[k]\right] = \rho_s(\theta[k])$, $\mathbb{E}\left[\nabla_\theta \rho_s[k]\right] = \nabla_\theta \rho_s(\theta[k])$.
- Load estimation is model-free (works for non-Poisson input).

A stochastic gradient algorithm

A generic algorithm:

$$C_s[k+1] = (1-\delta)C_s[k] + \delta\rho_s[k], \text{ (filtered loads)}$$
$$Y[k] = \sum_{1\le s\le N_s} \nabla_\theta \rho_s[k] u'(C_s[k]) \text{ (noisy gradient + bias)}$$
$$\theta[k+1] = \pi_{\mathcal{P}}[\theta[k] - \epsilon Y[k]] \text{ (projected gradient descent)}$$

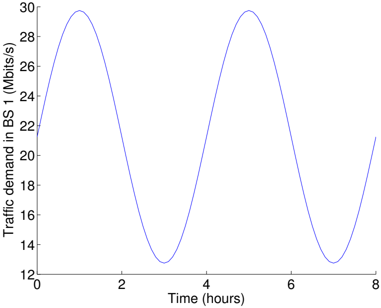
Theorem

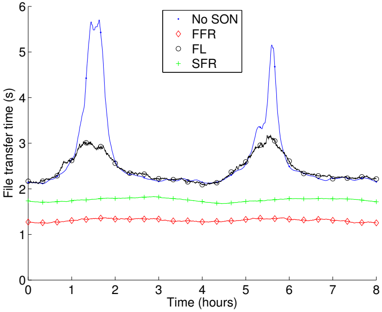
$\{\theta[k]\}_{k\in\mathbb{N}}$ converges in distribution to $\mathcal{U}$, the set of local minima of U on the constraint set $\mathcal{P}$ when $\epsilon \to 0$, $\delta \to 0$ and $\frac{\epsilon}{\delta} \to 0$. Namely, for all $\beta > 0$:

$$\limsup_k \mathbb{P}\left[d_{\mathcal{U}}(\theta[k]) > \beta\right] \underset{\epsilon,\delta,\frac{\epsilon}{\delta}\to 0}{\to} 0, \tag{1}$$

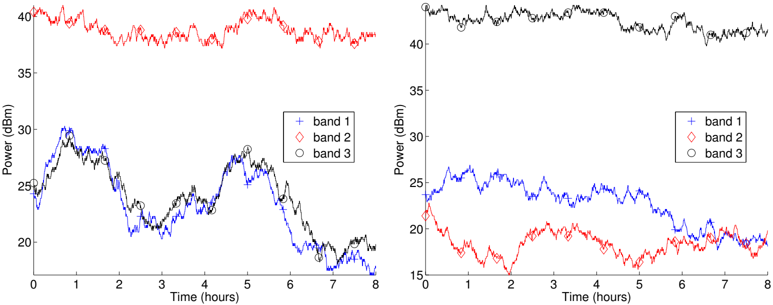
with $d_{\mathcal{U}}(\theta) = \inf_{u\in\mathcal{U}} \|\theta - u\|$ the distance to set $\mathcal{U}$.

Tracking performance

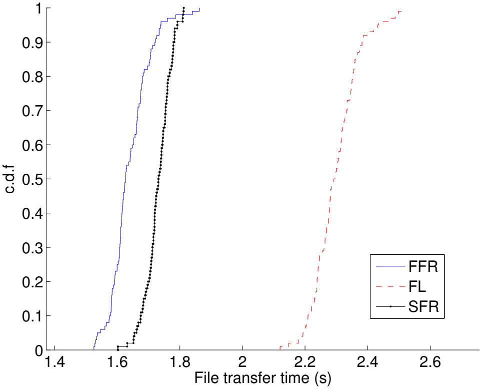




Apparition of dynamic reuse patterns



Global vs local optima



Advantages of the proposed approach

- Low signaling ($\approx 10 bits/s$) and delay requirements (BS to neighbors interface delay $\approx 50ms >> T \approx 60s$).
- Valid for all stationary ergodic input (“model free approach”)
- Delay optimal for some queuing models
- Fast enough to adapt to daily traffic patterns
- Gradient-type method: simple convergence analysis.

Questions ?

Thank you for your attention, any questions ?