Dynamic Resource Allocation for Delay-Sensitive Applications

Ishai Menache†, Asuman Ozdaglar† and Nahum Shimkin‡

† Department of Electrical Engineering and Computer Science, MIT

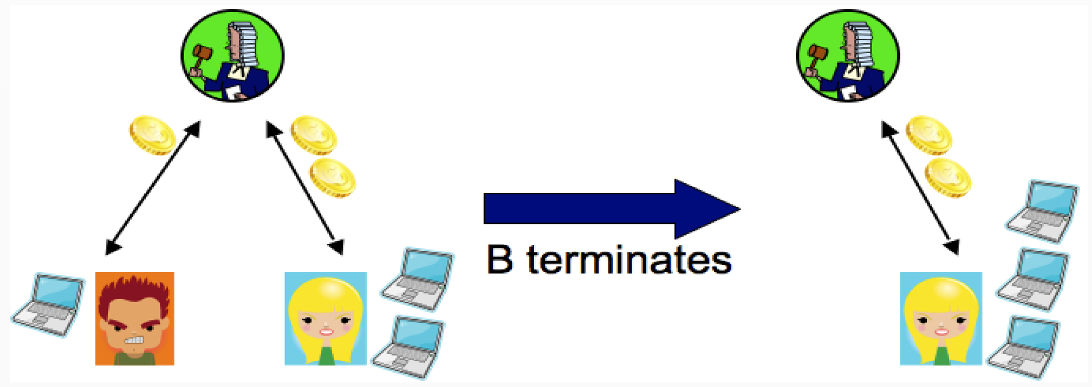
‡ Faculty of Electrical Engineering, Technion, Israel

Motivation

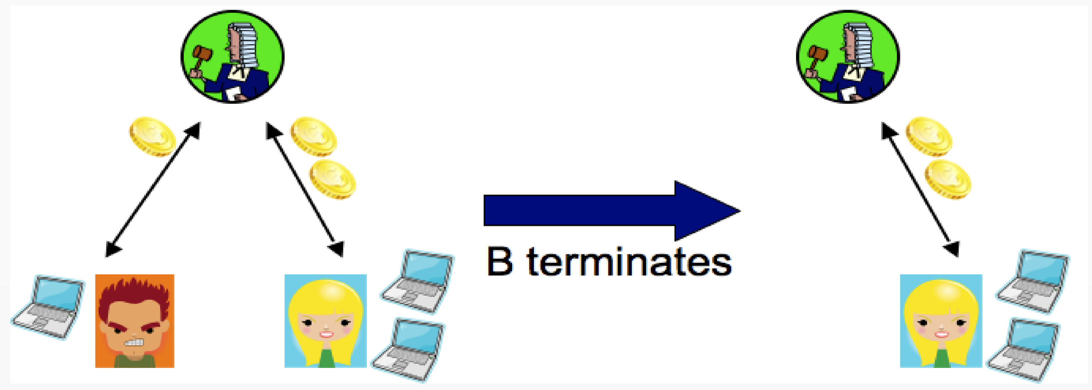
- Dynamic allocation of resources. Examples:
  - Cloud computing
  - Wireless-spectrum access
  - More...
- Delay or completion-time as a central performance metric.
- Twofold objective: (i) Design simple allocation mechanisms, (ii) Develop tools for their analysis

The Allocation Mechanism

- Ideal scheme: $z_i(t) = Z_0 \frac{w_i}{w_i + \sum_{j \in J_{-i}(t)} w_j}$, where $J_{-i}(t)$ is the set of other jobs that are active at time t.

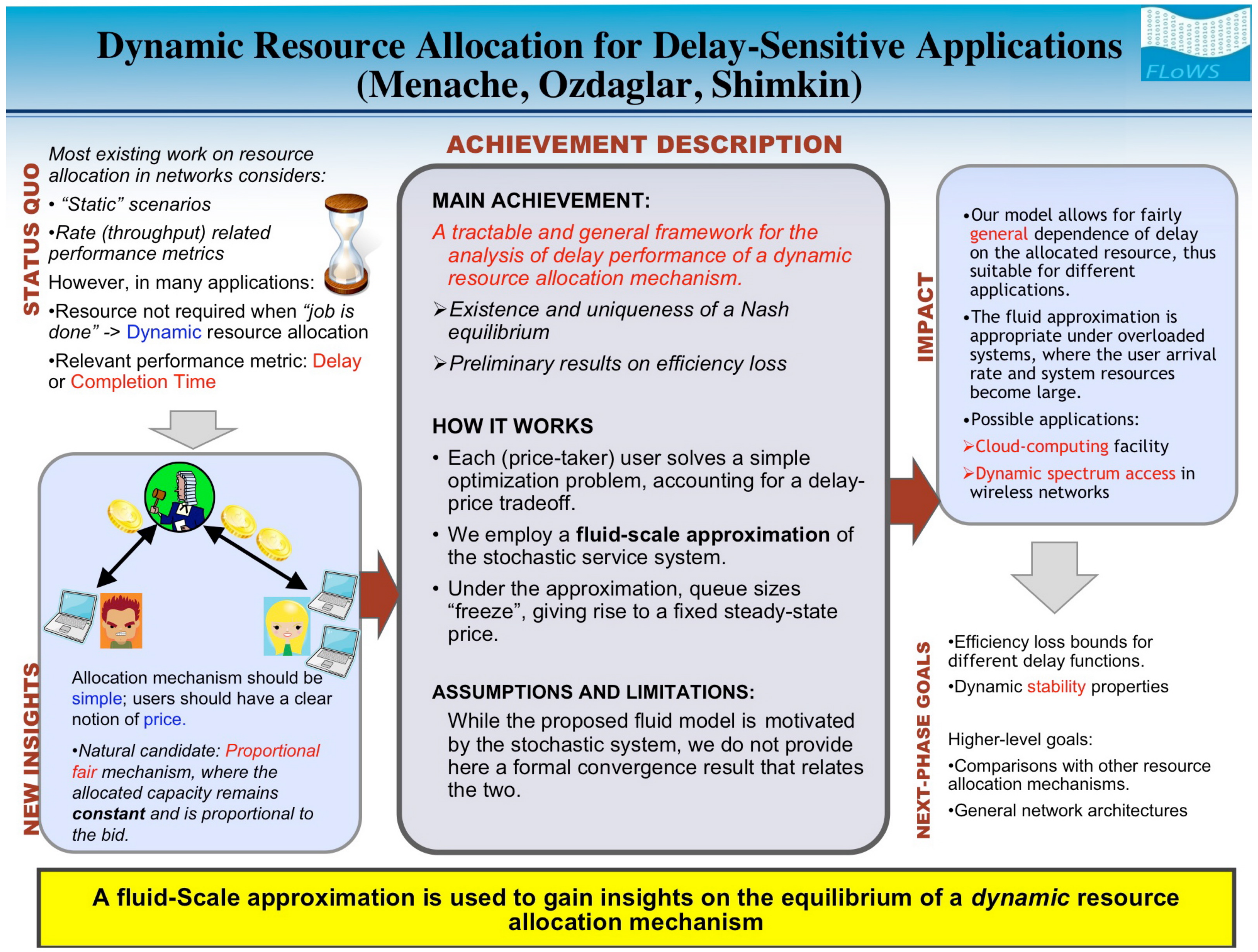


- Implementable scheme: $z_i = \frac{w_i}{P}$. The price P is determined according to $P = \frac{1}{Z_0} \sum_{j \in J_a(i)} w_j$, where $J_a(i)$ is the set of other active jobs at job i's arrival moment.



- The implementable scheme approximates the ideal scheme under plausible conditions.
- Total monetary transfer: $w_i T_i(z_i)$, where T_i is the delay (completion time).

General Summary



Delay, User-Cost and Demand

- **Assumption 1** [Marginal effectiveness of adding resources is decreasing]: Let $\mu_i(z_i) = \frac{1}{T_i(z_i)}$ be the effective service rate. We assume that $\mu_i(z_i)$ is a differentiable, strictly concave and strictly increasing function of $z_i \geq 0$, with $\mu_i(0) = 0$ and $\mu_i(\infty) < \infty$. Consequently, the delay $T_i(z_i)$ is convex-decreasing in z_i.
  - Example: $T_i(z_i) = a_i + \frac{D_i}{z_i}$.
- Assume a finite-set of user (or job) classes. The cost function J_s for a class-s is given by

$$J_s(w_i) = (c_s + w_i) T_s(z_i),$$

where c_s is the delay-disutility parameter.

- **Assumption 2** [Effective arrival rates decrease with price]: For every service class s, the arrival rate $\lambda_s(P)$ is continuous and strictly decreasing in $P \geq 0$, and $\lambda_s(P) \to 0$ as $P \to \infty$.

Fluid Scaling and Nash Equilibrium

- With n a large scaling factor, let the arrival rate of class s be $n\lambda_s$, the system resources nZ_0. After re-scaling, the arrival stream may be approximated by a deterministic rate λ_s (in fluid-units per unit time).
- Let Q_s denote the queue-size (in fluid units) of class-s users in the system. Corresponding dynamics:

$$\frac{dQ_s(t)}{dt} = \lambda_s(t) - Q_s(t)\mu_s(t).$$

- A class-homogeneous Nash equilibrium is characterized by the following equations:

$$Q_s = \lambda_s T_s, \qquad P = \frac{1}{Z_0} \sum_s Q_s w_s,$$

where $J_s(w_i) = (c_s + w_i) T_s(\frac{w_i}{P})$ and $w_s \in \arg\max_{w_i \geq 0} J_s(w_i)$

Summary of Results

Theorem 1. *Under Assumptions 1 and 2, there exists a unique class-homogeneous Nash equilibrium.*

CASE STUDY: Let $T_s(z_s) = a_s + \frac{D_s}{z_s}$. Consider demand functions of threshold type: Users of class s join if $J_s(T_s, w_s) \equiv c_s T_s + w_s T_s \leq v_s$, where v_s is a 'value of service' parameter. Then:

Theorem 2.
(i) The equilibrium decision of whether to join the system or not is a simple index-rule:

$$\sqrt{P} \gtrless \frac{\sqrt{v_s} - \sqrt{a_s c_s}}{\sqrt{D_s}}.$$

(ii) For $a_s \to 0$, the equilibrium delays are zero for all classes ($T_s = 0$); additionally, the equilibrium coincides with the socially optimal working point, i.e., no efficiency loss.