DIFFUSION LIMITS AND CONTROL

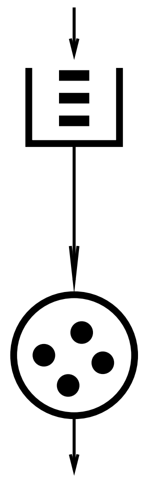
Rami Atar

Technion - Israel Institute of Technology

Department of Electrical Engineering

With Avishai Mandelbaum, Gennady Shaikhet, Adam Shwartz

HEAVY TRAFFIC APPROACH TO QUEUEING



Arrival rate: λ_n. Number of servers: N_n. Individual service rate: μ_n.

Heavy traffic approach = critically loaded system $\qquad \lambda_n \approx N_n \mu_n$

+ diffusive scaling

‘CRITICALLY LOADED’ MEANS $\lambda_n \approx N_n \mu_n$

which can be achieved as follows

$$\lambda_n \approx n, \qquad N_n \approx n^{\alpha}, \qquad \mu_n \approx n^{1-\alpha}, \qquad \text{where } 0 \le \alpha \le 1$$

‘DIFFUSIVE SCALING’ MEANS SECOND ORDER APPROXIMATIONS

Let $X_n(t) = \#$ customers in the system at time t,

$$\hat{X}_n(t) = \frac{X_n(t)}{\sqrt{n}}$$

FACT: *For the whole range* $0 \le \alpha < 1$, $\hat{X}_n$ *converges, under appropriate assumptions, to a reflecting Brownian motion, as* $n \to \infty$.

CONSEQUENCE: At most times, all servers are busy, queue is non-empty

THE CASE $\alpha = 1$: $\lambda_n \approx n$, $N_n \approx n$, $\mu_n \approx 1$

The queue is empty for a nontrivial fraction of time. This behavior is seen in applications.

More precisely, take

$$\lambda_n = \lambda n + \hat{\lambda}\sqrt{n} + l.o.t. \qquad N_n = n, \qquad \mu_n = \mu + \frac{\hat{\mu}}{\sqrt{n}} + l.o.t.,$$

with exponential service time distribution. Assume $\lambda = \mu$.

HALFIN AND WHITT'S RESULT ('81):

$$\hat{X}_n(t) := \frac{X_n(t) - N_n}{\sqrt{n}} \Rightarrow \xi, \quad \text{where} \quad d\xi = dw + (\hat{\lambda} - \hat{\mu})dt + \mu\xi^- dt$$

$I_n(t)$:=# idle servers at time t, $\quad \hat{I}_n(t) := \dfrac{I_n(t)}{\sqrt{n}} \Rightarrow \xi^-$.

Decrease of processing rate due to idleness: $\mu\xi^-$.

REMARKS

Without the exponential assumption the result is not valid. Reason: from the viewpoint of an individual server, time is not accelerated, hence past is important to keep track of.

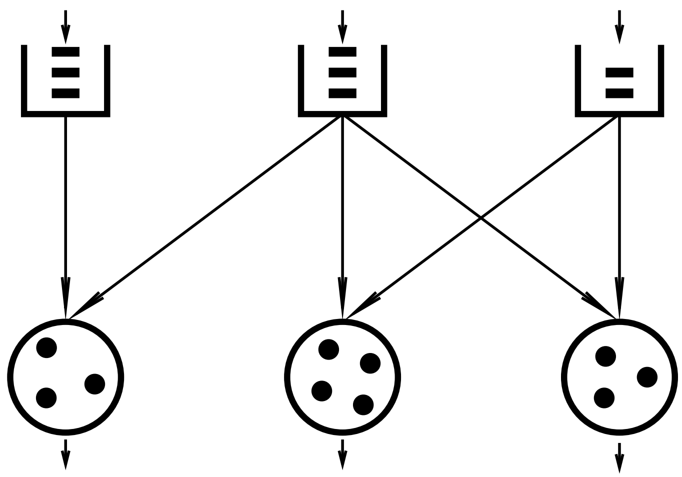
Results beyond the exponential service time distribution:

* Puhalski and Reiman ('00): Phase type service time distribution, diffusive scaling at 'near stationarity' - **higher dimensional diffusion**

* Reed ('07): General service time distribution, fluid and diffusion limit, one dimensional evolution equation in **function space**

* Kaspi and Ramanan ('07): General service time distribution, fluid limit, evolution equations in **measure space**

HETEROGENOUS SERVERS, MULTIPLE POOL APPROACH



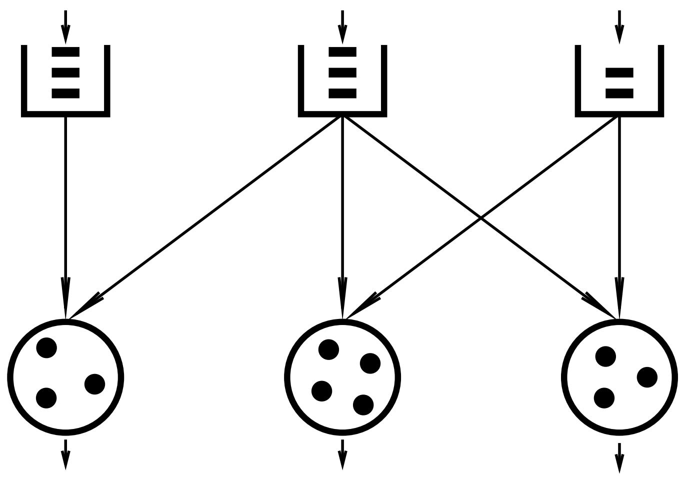
- I classes of customers; arrivals at rates λ_i, $i = 1, \ldots, I$; independent renewal processes.

- J service stations; station $j = 1, \ldots, J$ consists of N_j identical servers working independently. Exponential service times.

- Rate of service: μ_{ij} for class-i customer at station j.

DYNAMIC CONTROL PROBLEM (routing and scheduling):

Choose a service station for each customer, and when to start its service...



...so as to minimize a cost.

Fixed policy setting

Queueing process —— **Diffusion process**

Rigorous relation: Weak convergence

Control setting

Queueing control problem —— **Diffusion control problem**

= Controlled queueing model
+ Cost criterion

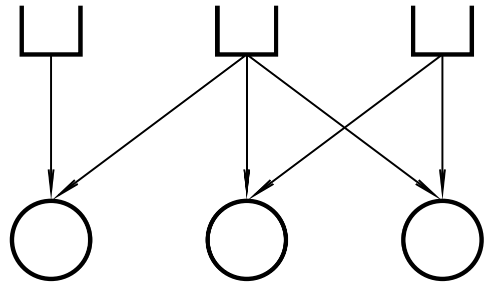
= Controlled diffusion model
+ same cost criterion

Rigorous relations:
1) Convergence of value
2) Identify asymptotically optimal policy (AO)

Understand DCP solution $->$ Propose policy $->$ Prove AO

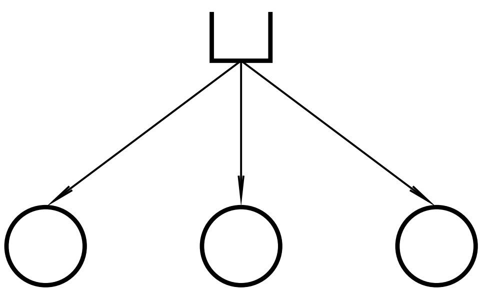
BRIDGE: Optimal control theory, Hamilton-Jacobi-Bellman equations

HALFIN-WHITT REGIME, GENERAL $I \times J$

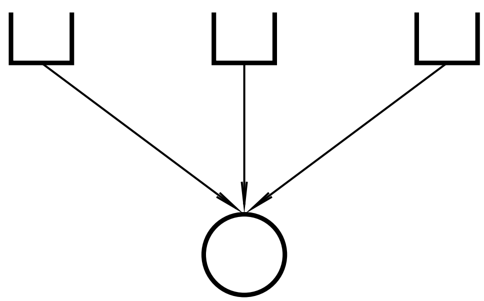


Much studied partial models include:

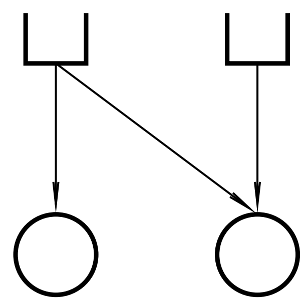
$1 \times J$



$I \times 1$



2×2



$I \times J$ MODEL IN THE HALFIN-WHITT REGIME

Recall: in the 1×1 case the process X^n was centered about n:

$$\hat{X}^n(t) = n^{-1/2}(X^n(t) - n) \tag{1}$$

and the "fluid-level" parameters were assumed to satisfy

$$\lambda = \mu \tag{2}$$

Before introducing the rescaled processes for the $I \times J$ case we must define a fluid model about which to center (substitute for (1)), and to study it in heavy traffic (substitute for (2)).

FLUID MODEL IN HEAVY TRAFFIC: DEFINITION

Static allocation problem: *Minimize* $\rho \in \mathbb{R}$ *subject to*

$$\sum_{j=1}^{J} \mu_{ij}\xi_{ij} = \lambda_i \quad \forall i, \qquad \sum_{i=1}^{I} \xi_{ij} \le \rho \quad \forall j, \qquad \xi_{ij} \ge 0 \quad \forall i, j$$

Heavy traffic condition [Harrison & López (1999)]: *There exists a unique optimal solution* (ξ^*, ρ^*) *to the linear program. Moreover,* $\rho^* = 1$, *and* $\sum_i \xi^*_{ij} = 1$ *for all* $j \in \mathcal{J}$.

Denote $x^*_i = \sum_{j=1}^{J} \xi_{ij}$, $i = 1, \ldots, I$, let $x^* = (x_i)_{i=1,\ldots,I}$

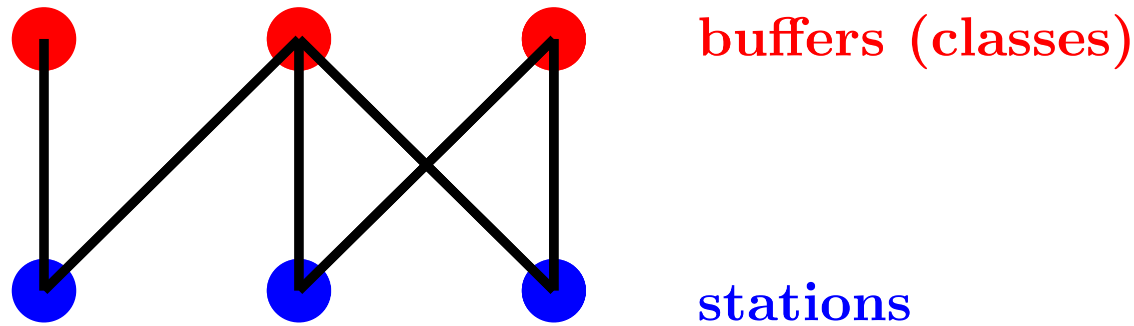
$X^n = (X^n_i)$ $i = 1, \ldots, I$: class-i population in system
$\Psi^n = (\Psi^n_{ij})$ $i = 1, \ldots, I$, $j = 1, \ldots, J$: population in activity (i, j).

Rescaling: $\hat{X}^n(t) = n^{-1/2}(X^n(t) - nx^*)$

FLUID MODEL IN HEAVY TRAFFIC: ACTIVITIES

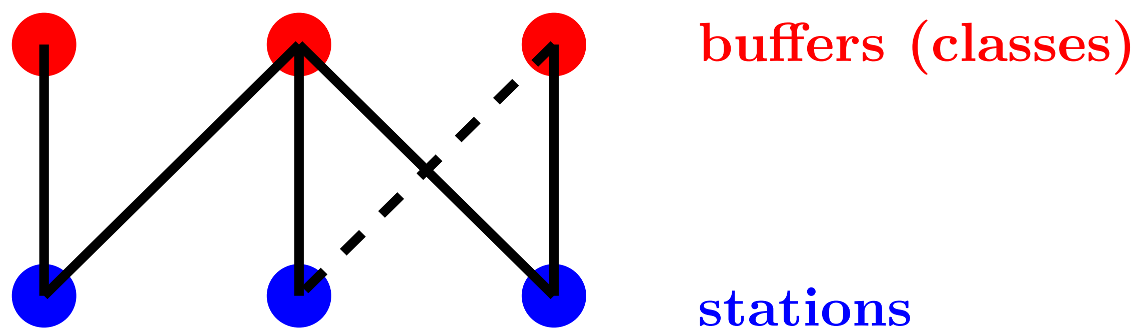
Activity: *a pair* (i, j) *such that station* j *can serve class* i*.*

Graph of activities



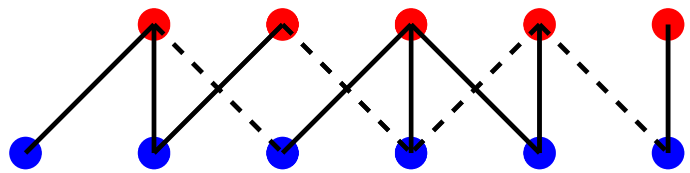
An activity (i, j) *is* **basic** *if* $\xi^*_{ij} > 0$ *and* **non-basic** *if* $\xi^*_{ij} = 0$.

Graph of basic activities*

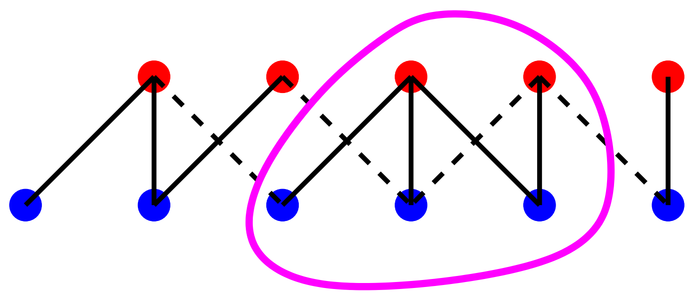


(*) Non-basic activities shown in dashed line

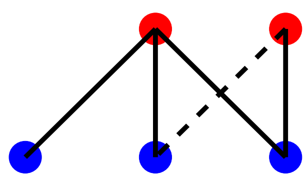
FACT (Williams 2000): *The graph of basic activities contains no cycles.*



We will focus on a single component



Thus we assume *the basic activities form a tree* (known as the complete resource pooling assumption)



DIFFUSION MODEL

The diffusion model obtained for weak limits of $\hat{X}^n$:

$$X(t) = x + W(t) + \int_0^t b(X(s), U(s))ds + \sum_{k=1}^{K} m_k \eta_k(t),$$

X - a diffusion in $\mathbb{R}^I$,
W - an I-dimensional Brownian motion,
b - drift coefficient,
$\{m_k\}$ - fixed vectors,
$(U, \{\eta_k\})$ - control process:

- $U(t)$ takes values in a compact set,
- η_k is a non-decreasing process, $k = 1, \ldots, K$.

Main point: *Diffusion with singular control.*

ROUTING AND SCHEDULING CONTROL PROBLEM

Preemptive: Service to a customer can be stopped and resumed at a later time, possibly in a different station.
Scheduling decisions are made by continuously selecting Ψ.

Nonpreemptive: Customers complete service with the server they are first assigned
Scheduling decisions are made by selecting a server for each customer, and when to begin service.

RESULTS ON ASYMPTOTIC OPTIMALITY

Cost for the queueing scheduling problem:

$$E \int_0^\infty e^{-\gamma t} \tilde{L}(\hat{X}^n(s), \hat{\Psi}^n(s))ds,$$

$$E \int_0^T \tilde{L}(\hat{X}^n(s), \hat{\Psi}^n(s))ds + g(\hat{X}^n(T)).$$

Theorem: In each of the above cases where the PDE analysis is possible, the optimal cost of the queueing scheduling problem converges to the diffusion control problem's value. Moreover, one can construct **preemptive** and **nonpreemptive** scheduling controls under which the cost converges to the same limit.

The optimal control U for the diffusion problem is given as

$$\Psi_s = h(X_s),$$

where h is a function determined by the PDE solution.

For the preemptive problem, it is possible to let

$$\hat{\Psi}_s^n = h(\hat{X}_s^n)$$

and argue by weak convergence.
For the nonpreemptive problem there is no direct control over the $\hat{\Psi}$ process.
Apply a **tracking mechanism:**

- Compute $h(\hat{X}_s^n)$
- Declare activities with $\hat{\Psi}_s^n > h(\hat{X}_s^n)$ as "over-populated"
- Block over-populated activities. The population drops rapidly; population in the other activities increases rapidly
- As a result

$$\hat{\Psi}_s^n \sim h(\hat{X}_s^n)$$

RESULTS ON THE HJB EQUATION

The equations for the infinite and resp., finite horizon problems are

$$(\Delta - \gamma) f + H(x, \nabla f) = 0; \qquad H(x,p) = \inf_{U \in \mathbb{U}} [b(x,U) \cdot p + L(x,U)]$$

$$(\frac{\partial}{\partial t} + \Delta_x) f + \tilde{H}(x, \nabla_x f) = 0; \qquad \tilde{H}(x,p) = \inf_{U \in \mathbb{U}} [b(x,U) \cdot p + L(x,U)];$$

with growth condition

$$\exists C \quad |f(x)|, |f(t,x)| \leq C(1 + \|x\|^C), \quad x \in \mathbb{R}^I$$

and terminal condition

$$f(T, \cdot) = g$$

Assumptions on L and g: Some regularity, and polynomial growth in x.

Infinite horizon discounted problem:

Theorem: Unique solvability by the value function, under each of the following conditions:

1) μ_{ij} depends only on i; or μ_{ij} depends only on j; no abandonments.

2) The tree $\mathcal{G}$ is of diameter 3 at most, and $\forall (i,j) \in A\ \theta_i \leq \mu_{ij}$

3) $L(x,U) \sim \|x\|^m$, and $\exists (i,j) \in A\ \theta_i \leq \mu_{ij}$

4) $L(x,U)$ is bounded.

Finite horizon problem:

Theorem: Unique solvability by the value function holds in general.

Back to the queueing model: We will be interested in properties (1) and (2) below.

Let $Q^n(s)$ = *total # customers in queues (i.e., not being served) at time s.*

Null-controllability (NC): *The ability to maintain a system with* **empty queues** *on arbitrarily long time intervals, i.e.,*

one can find a control policy for the queueing system under which
$$P(Q^n(s) = 0\ \forall\, s \in [\varepsilon, t]) \longrightarrow 1 \quad \textit{as } n \to \infty, \quad \forall\, 0 < \varepsilon < t < \infty. \tag{1}$$

Weak null-controllability (WNC): *The ability to maintain empty queues "almost always" within long time intervals, i.e.,*

one can find a control policy for the queueing system under which
$$\int_0^t 1_{\{Q^n(s)>0\}} ds \longrightarrow 0 \quad \textit{in probability, as } n \to \infty, \quad \forall\, 0 < t < \infty. \tag{2}$$

NULL-CONTROLLABILITY RESULTS

Main result 1 (with Mandelbaum & Shaikhet): *We find classes of systems that exhibit NC and WNC.*

This identifies a class of queueing systems that are critically loaded (and in heavy traffic in the usual sense) but behave as if they are underloaded.

Main tool: The singular control formulation of the diffusion model.

Main result 2 (with Shaikhet): *A necessary and sufficient condition for WNC in terms of a property of the fluid model that we refer to as* **throughput optimality.**

This is explained in what follows.

THROUGHPUT OPTIMALITY OF THE FLUID MODEL

Recall the LP: *Minimize* $\rho \in \mathbb{R}$ *subject to*

$$\sum_{j=1}^{J} \mu_{ij}\xi_{ij} = \lambda_i \quad \forall i, \qquad \sum_{i=1}^{I} \xi_{ij} \leq \rho \quad \forall j, \qquad \xi_{ij} \geq 0 \quad \forall i, j$$

Unique optimal solution (ξ^*, ρ^*), and $\rho^* = 1$. $x_i^* = \sum_{j=1}^{J} \xi_{ij}$, $i = 1, \ldots, I$.

Define the maximal throughput of the fluid model as

$$T := \max \left\{ \sum_{i,j} \xi_{ij}^* \mu_{ij} : \sum_i \xi_{ij} \leq 1, \sum_j \xi_{ij} \leq x_i^*, \xi_{ij} \geq 0 \right\}.$$

T represents maximum amount of fluid that can be processed.

The fluid model is said to be throughput optimal if $T = \sum_i \lambda_i$.

Null-controllability and throughput optimality

Main result 2 (rephrased): *One can find a control policy for the queueing system under which*

$$\int_0^t 1_{\{Q^n(s)>0\}} ds \longrightarrow 0 \quad \textit{in probability, as } n \to \infty, \qquad \forall\, 0 < t < \infty,$$

(i.e., WNC holds) if and only if the fluid model is not throughput optimal.

Major open issues

1. Null-controllability and throughput optimality can be defined for a much broader class of queueing networks; are they related in general?

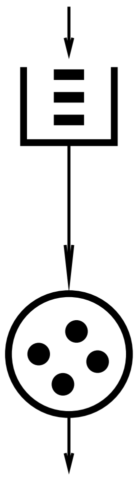
2. Rigorous relation under genuine singular control problem.

CONTROL UNDER MODEL UNCERTAINTY

1) Models where parameters are random

2) Models where information on service time distribution is given only through samples from the distribution

3) Models where service rates are completely unknown



$$\lambda_n = \lambda n + \hat{\lambda}\sqrt{n} + l.o.t.$$

N_n servers, iid rates $\mu_1, \ldots, \mu_k$, dist. m

Critical load condition: $\lambda = \mu := \int x dm$

RANDOM PARAMETERS; MAXIMUM THROUGHPUT POLICY

Service rates are known to the router. The policy is *to route to the fastest server available.*

Policy studied by Armony ('05) in a deterministic environment setting.

Main new ingredient in random environment setting: random drift term that comes from a CLT limit for the service rates.

THEOREM: *Assume* $\operatorname{ess\,sup} m < \infty$. *Under policy I,* $\hat{X}_n \Rightarrow \xi$, *a diffusion with a random drift coefficient given by*

$$d\xi = \sigma dw + \beta dt + \mu_* \xi^- dt,$$

where $\beta = \hat{\lambda} - \int x d\hat{m} + \zeta - \mu\nu$, $\quad \zeta \sim \mathcal{N}(0, \int (x-\mu)^2 dm)$, $\quad \sigma^2 = \lambda(C^2 + \mu)$, *and* ζ, ν, w *are independent*

Main reason for a 1-d diffusion limit: most of the free servers are those that are slowest.

SAMPLING FROM SERVICE TIME DISTRIBUTION

Choose at random $n^{1/2+\varepsilon}$ servers.
Take a single sample from service time distribution of each.
Order servers according to the sample.

THEOREM (with Shwartz): *Weak limit is given by*

$$d\xi = \sigma dw + \beta dt + \mu_* \xi^- dt;$$

Under any other policy, all weak subsequential limits dominate ξ pathwise.

UNKNOWN RATES. FAIR JOB ALLOCATION

Service rates are not known to the router. The policy is *to route to the server who has been idle for the longest time since it last served.*

THEOREM: Assume $\int x^2 dm < \infty$. Then $\hat{X}_n \Rightarrow \xi$, where

$$d\xi = \sigma dw + \beta dt + \gamma \xi^- dt,$$

where

$$\gamma = \frac{\int x^2 dm}{\int x dm}$$

Why 1-d diffusion? Why γ?

The main reason is very different from the one in the previous case.

HEURISTIC EXPLANATION OF RESULT

With the relation $I(t;\Delta x) \approx Nm([x, x+\Delta x))xh(t)$

and $N \approx n$, we have

$$
\begin{aligned}
F^n(t) &:= \text{the (normalized) decrease of processing rate due to idleness} \\
&= n^{-1/2}\sum_{k=1}^{N}\mu_k I_k(t) \approx n^{-1/2}\int xI(t;dx) \approx n^{1/2}h(t)\int x^2 m(dx), \\
\widehat{I}^n(t) &= n^{-1/2}\sum_{k=1}^{N} I_k(t) \approx n^{1/2}h(t)\int xm(dx)
\end{aligned}
$$

Hence

$$
F^n(t) \approx \frac{\displaystyle\int x^2 m(dx)}{\displaystyle\int xm(dx)}\int_0^t \widehat{I}^n(s)ds = \gamma\int_0^t X^n(s)^- ds
$$

which explains the limit

BEHAVIOR UNDER 'FAIR JOB ALLOCATION' POLICY

1) 1-dimensional diffusion

2) 'Local' fairness is achieved:

Given a time t, the load is balanced among all servers that at or near t have become available, in such a way that they all experience an idle period of approximately the same length.

3) Sample path Little's law:

- Number of idle servers is asymptotic (in diffusion scale) to ξ^-
- Average rate at which servers become idle is (in fluid scale) μ
- Time servers "wait for a job" is asymptotic (at diffusion scale) to $\mu^{-1}\xi^-$

SOME OPEN ISSUES

- ‘Global’ fairness under natural policy?

- Diffusion limits under natural policies:
    - Random routing
    - Routing to server that has shown best performance first

- Dynamic versions of sampling result