

ISKE 2009

Choosing Between Several Queuing Policies

pierre.douillet@ensait.fr

École Nationale Supérieure des Arts et Industries Textiles

Roubaix, France

⇒ ●	How to manage a G/GI/n system ?	3
	averaging process	
	criteria	
	assumptions	
	batch mean method	
●	manager's point of view	7
●	customer's point of view	11
●	scaling and pooling	14
●	conclusion	17

How to manage a G/GI/n system ?

averaging process

- customer : "one man, one vote" $E_c(X)$
- manager : "one clock tick, one vote" $E_t(X)$

criteria

- manager : exhaustivity and number of waiting customers
- customer : mean and variance of sojourn time
- fairness, and perceived fairness, are important

assumptions

- n identical servers,
any service is independent from anything else
- total capacity of service $\mu = n / E_B(t)$
- independent (...) arrivals, flow $\lambda = E_A(t)$, $\rho = \lambda / \mu < 1$
- distributions : anything except from M/M
- Here : B is Gamma (svc= 0.4), A is Gamma (svc=1.25),
 $\rho = 0.93$ or 0.97

batch mean method

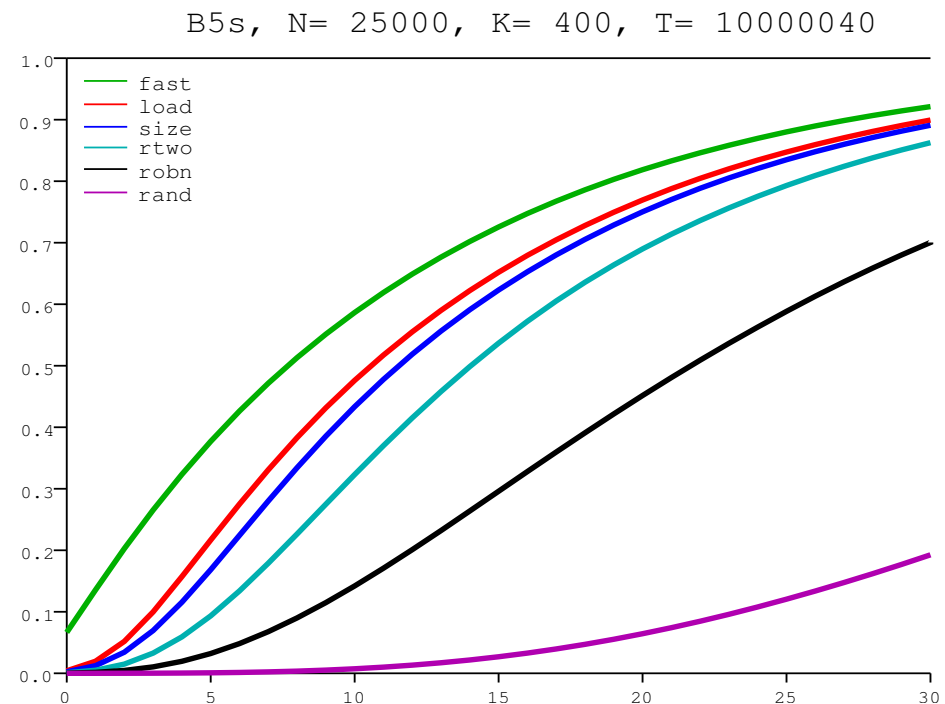
- each result has been obtained with $K = 400$ batches of $N = 50000$ events
- containing rounding errors, allowing parallelization (with suitable random generator)
- estimation of the *sd* of the estimators (and checking for independence)

✓ ●	How to manage a G/GI/n system ?	3
⇒ ●	manager's point of view	7
	ordinary policies	
	number of busy servers	
	jockeying	
●	customer's point of view	11
●	scaling and pooling	14
●	conclusion	17

manager's point of view

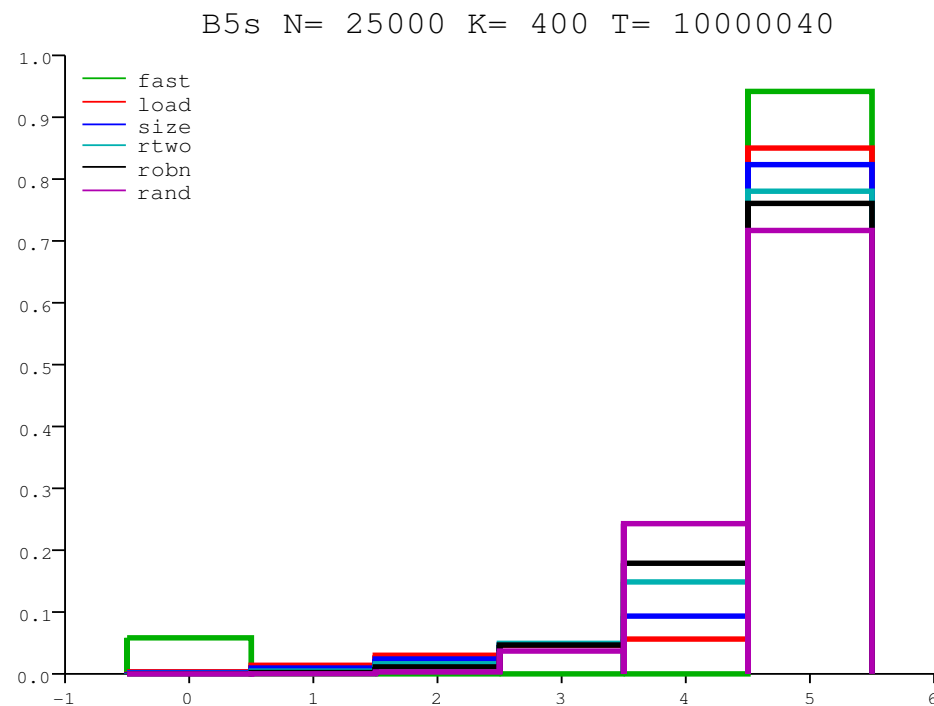
ordinary policies

- *rand*, *robn*,
- *rtwo* (distributed)
- *size* (shortest queue)
- *load* (how can we ... ?)
- *fast* (single, μ , scv)



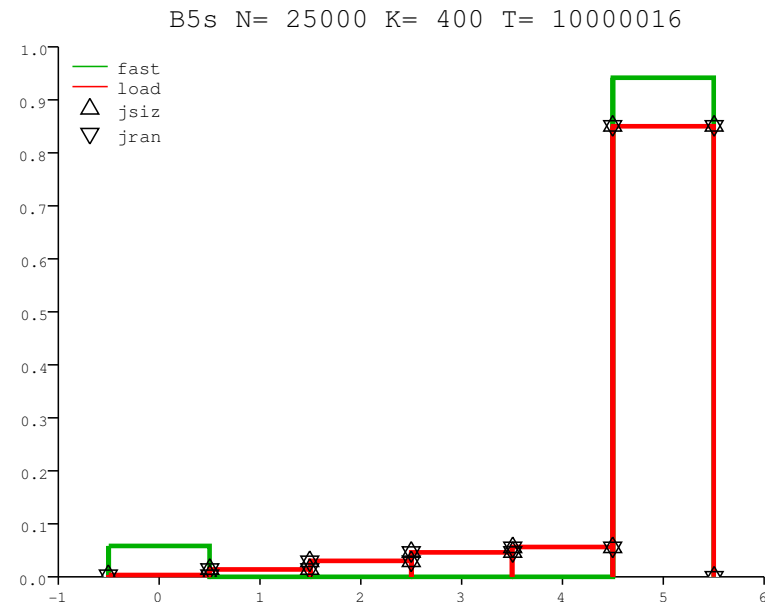
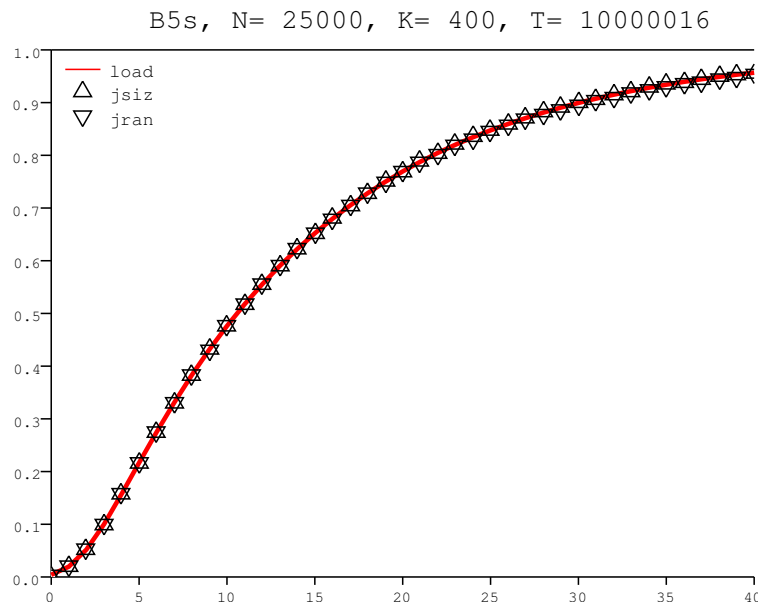
number of busy servers

- *load* is pooling, *size* is not the optimal
- Little : $E_t(n_b) = n\rho$,
doesn't depend on policy
- probability ρ^* of full use
of the capacity of service
- *fast* is $\rho^* = \rho$
- *load* ensures exhaustivity



jockeying

- $jsiz = size$ then jockeying, $jran = rand$ then jockeying
- same distribution of queue length and servers business as *load* : jockeying solves (mostly) the manager's problem.

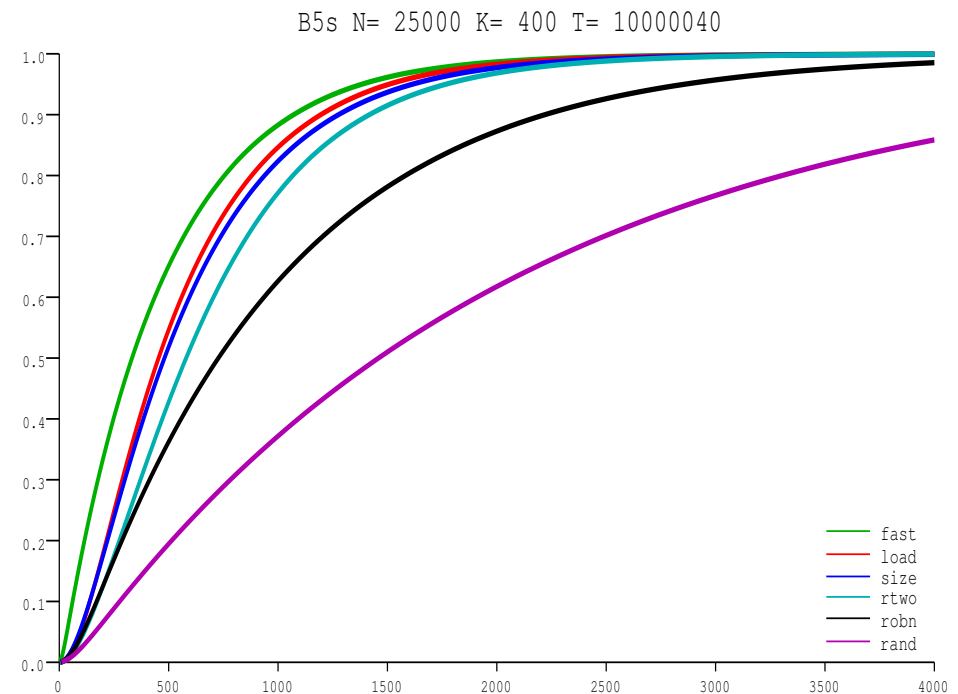


✓ ●	How to manage a G/GI/n system ?	3
✓ ●	manager's point of view	7
⇒ ●	customer's point of view	11
	sojourn time	
	some results	
●	scaling and pooling	14
●	conclusion	17

customer's point of view

sojourn time

- mean sojourn time
- MM1 : parameter $\mu - \lambda$
- variance and fairness



some results

	mean	$\pm 2\sigma$	ratio	sd	$\pm 2\sigma$	ratio	$\tau_{0.1\%}$	τ/μ
fast	470.53	7.00	1.00	445.61	10.01	1.00	3808	8.1
load	586.72	7.36	1.25	463.88	10.60	1.04	4189	7.2
jsiz	606.51	7.89	1.29	496.94	11.40	1.12	4435	7.4
jran	675.53	7.09	1.44	670.32	11.94	1.50	7479	11.2
size	621.48	7.69	1.32	500.77	10.48	1.12	4423	7.1
rtwo	710.71	7.36	1.51	527.89	10.37	1.18	4817	6.8
robn	997.21	13.15	2.12	904.20	18.61	2.03	8793	8.9
rand	2102.66	31.03	4.47	1999.41	45.64	4.49	17560	8.4

τ is the last 1/1000 fractile

✓ ●	How to manage a G/GI/n system ?	3
✓ ●	manager's point of view	7
✓ ●	customer's point of view	11
⇒ ●	scaling and pooling	14
	how to model scaling ?	
	pooling factor	
	pooling reshapes towards Poisson	
●	conclusion	17

scaling and pooling

how to model scaling ?

- shape
- independence (short range)
- independence (long range) ???
- distributed, with coupling ?

pooling factor

- when customer flow increases and the number of servers increases accordingly, the mean sojourn time decreases

		exhaustive			non exhaustive			
n	fast	load	jsiz	jran	size	rtwo	robn	rand
7	339	458	481	550	497	627	921	2070
5	470	586	606	675	621	710	997	2102
3	794	881	898	965	911	948	1242	2141
1	2378	2378	2378	2378	2378	2378	2378	2378

- pooling can also result in reduced staff...

pooling reshapes towards Poisson

- $S_1(z)$, $S_n(z)$, $S_r(z)$ are the mgf of A , n scaled A , and the resulting *rand*-arrivals in a single queue

$$S_1(z) = \int a(t) \exp(tz) dt, \quad S_n(z) = S_1\left(\frac{z}{n}\right), \quad S_r(z) = \frac{S_n(z)}{n - (n-1)S_n(z)}$$

- Expanding in series :

$$S_1(z) = 1 + \frac{z}{\lambda} + \frac{z^2}{2\lambda^2} (1 + scv) + \dots, \quad S_r(z) = \frac{\lambda}{\lambda - z} + \frac{z^2}{2n} \frac{(scv - 1)}{(\lambda - z)^2} + z^3 O\left(\frac{1}{n^2}\right)$$

$$scv(a_r) = 1 + (scv(a) - 1) / n$$

conclusion

- manager/customer perceptions are not based on the same averaging process
- moreover manager's focus is exhaustivity
- customer's focus is variance and fairness
- centralized/distributed systems are different
- aggregation changes drastically the results