WCFS Queues:
A new analysis framework

Isaac Grosof
Mor Harchol-Balter
Alan Scheller-Wolf
Unexpected Similarity between Queues

Discover commonality

Use commonality to tightly characterize
Warmup: M/G/1

Poisson arrivals (memoryless), arrival rate: $\lambda$

General job size distribution: $S$, i.i.d.

Fraction of time busy:
Load: $\rho = \lambda E[S] < 1$

Response time: $T$

$E[S]$, $S \sim \text{Hyperexp}$

$E[T] = E[S](1-\rho)$

Load vs. $E[T]$: 0, 0.25, 0.5, 0.75, 1

Load vs. $E[T](1-\rho)$: 0, 0.25, 0.5, 0.75, 1

$E[S^2] = 2E[S]$
$M/G/k$

$\rho \to 1$ limit: $\frac{E[S^2]}{2E[S]}$

[Kollerstrom ‘84]
Mean response time behavior:
Open problem!

Heterogeneous M/G/k

Speed $v_1$
Speed $v_2$
Speed $v_3$
Speed $v_k$

Mean response time behavior:
Open problem!

$E[T](1-\rho)$

$E[S^2]$

$2E[S]$

$E[G/1]$

$E[G/k]$

$E[M/1]$

$E[M/k]$
M/G/1/Limited Processor Sharing

Open problem!
(Unlimited) Processor Sharing:
Known, different limit: $E[S]$
Threshold Parallelism
Threshold Parallelism
Threshold Parallelism
Threshold Parallelism

![Diagram showing Threshold Parallelism](image)

- **Het. M/G/k**
- **M/G/k**
- **M/G/1**
- **LPS**
- **TP FCFS**
- **TP Many first**
- **TP Few first**

Open problem!
Few first or many first: Different limits
All existing policies: open problem!
All different limits, many unstable
We create new policy: ServerFilling
Result: Response time bound

Theorem: All models \( \pi \in WCFS^* \):

\[
\lim_{{\rho \to 1}} E[T^\pi](1 - \rho) = \frac{E[S^2]}{2E[S]}
\]

Even stronger theorem:

\[
E[T^\pi] - E\left[T^{M/G/1}\right] \in [c_\ell^\pi, c_h^\pi]
\]

Goals: Define WCFS, prove result

(Subject to minor condition on job size distribution S)
Service policy serves jobs at some service rate
Work completes at service rate, age increases
Service policy serves jobs at some service rate
Work completes at service rate, age increases
Job completes when age reaches size
Base Queueing Model

- Service policy serves jobs at some service rate
- Work completes at service rate, age increases
- Job completes when age reaches size
- Convention: normalize maximum service rate to 1.
Finite Skip

Finite skip: Only serve jobs among the n oldest jobs in arrival order.
Work conserving (for Finite Skip policies)
Work conserving: If $\geq n$ jobs present, total service rate 1 (maximum)
WCFS Policies

Polices that converge:
- M/G/1
- M/G/k
- Heterogeneous M/G/k
- Limited Processor Sharing
- Threshold Parallelism FCFS
- Multiserver-job ServerFilling

Polices that don’t converge:
- M/G/1/SRPT
- (Unlimited) Processor Sharing
- Threshold Parallelism, most servers first
- Multiserver-job FCFS
WCFS Example: Heterogeneous M/G/k

Is this model WCFS? Yes!
WCFS Example: Multiserver-Job FCFS

Is this model WCFS?
No, not work conserving.
Define ServerFilling for Multiserver-Job model

ServerFilling setting: requirements powers of 2, k is power of 2.

1. Find minimal prefix M containing jobs that require ≥k servers
2. Among prefix M, serve largest requirement first

Finite skip: |M| ≤ k

Work conserving: Theorem: ServerFilling always fills all k servers.

Is this model WCFS? Yes!
Key ideas behind response time bound

Want to prove: $E[T^\pi] - E[T^{M/G/1}] \in [c_l^{\pi}, c_h^{\pi}]$

Key ideas based on work W.
1. $E[T^\pi] \approx E[W^\pi]$
2. $E[W^\pi] \approx E[W^{M/G/1}]$
Idea 1: $E[T] \approx E[W]$

Q: Starred job’s response time?
Idea 1: $E[T] \cong E[W]$

Max time in queue: $W^A$, because completion rate 1.
Min time in queue: $W^A - n$ jobs
Differ by constant in expectation.

Thm: Time in front bounded in expectation over all jobs.
Idea 2: $E[W] \approx E \left[ W^{M/G/1} \right]$ - Intuition

- Job arrives: +S work
- All servers full: rate 1
- Too few jobs: lower rate

Identical to M/G/1!
Proof of $E[W] \cong E\left[W^{M/G/1}\right]$

Consider $W(t)^2$

Consider $\frac{d}{dt} W(t)^2$

Expected stationary rate of change: $E\left[\frac{d}{dt} W^2\right] = 0$

Increases: Jump up by $S$, at rate $\lambda$.

$E[\text{inc.}]: \lambda E[(W + S)^2 - W^2] = 2\lambda E[W]E[S] + \lambda E[S^2] = 2\rho E[S] + \lambda E[S^2]$

Decreases: Work completes at rate $B(t)$. Rate 1 if $\geq n$ jobs.

$E[\text{dec.}]: 2E[WB] = 2E[W] - 2E[W(1 - B)]$

$$E[W] = \frac{\rho E[S^2]}{1 - \rho 2E[S]} + \frac{E[W(1 - B)]}{1 - \rho} = E\left[W^{M/G/1}\right] + \frac{E[W(1 - B)]}{1 - \rho}$$
Proof of $E[W] \cong E \left[ W^{M/G/1} \right]$

$E[W] = E \left[ W^{M/G/1} \right] + \frac{E[W(1-B)]}{1-\rho}$

If $B < 1$, $W$ consists of $< n$ jobs, so $W < c$.

$E[W(1-B)] \leq c E[1-B] = c(1-\rho)$

$E[W] \leq E \left[ W^{M/G/1} \right] + c$

$E[W] \cong E \left[ W^{M/G/1} \right]$

Completes proof that $E[T^\pi] - E \left[ T^{M/G/1} \right] \in [c^\pi_l, c^\pi_h]$
Empirical validation

\[ E[T^π] - E[T^{M/G/1}] \]

- Limited Processor Sharing
- Heterogeneous M/G/k
- MSJ ServerFilling
- Threshold Parallelism
Future questions

• Work conserving and finite skip, relative to a different job ordering, e.g. SRPT?
• Finite skip but not work conserving?
• Finite skip in expectation?
Conclusion

Explained unexpected similarity between queues
Defined work conserving finite skip models
Tightly bounded mean response for all WCFS models.
Extra: Condition on S

Bounded expected remaining size:

Exists constant $c$ such that for all ages $a$,

$$E[S - a \mid S > a] < c$$
Extra: DivisorFilling

Like ServerFilling, MSJ policy that fills all servers if enough jobs available. In particular, WCFS policy.

DivisorFilling works whenever all requirements divide the number of servers $k$. If $k$ jobs present, will fill all servers.

If jobs can have requirements that don’t divide $k$, WCFS not possible.