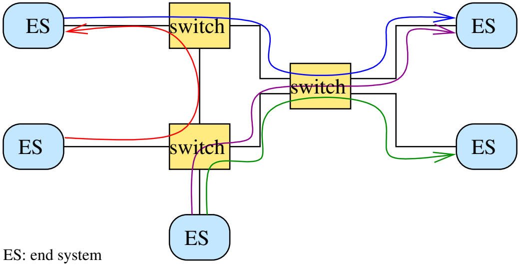
Some advances in Stochastic Network Calculus

Anne Bouillard
July 6, 2022

Objective of Network Calculus

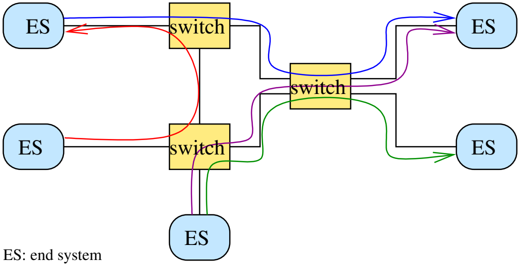


- classes of flows with hard delay constraints

Objective: deterministic performance guarantees

Compute the maximum time it takes for a packet to cross the system (Worst-case delay)

Objective of Network Calculus



- classes of flows with hard delay constraints
- classes of flows with not so hard delay constraints

Objective: stochastic performance guarantees

Compute the time it takes for 99.99% of the packets to cross the system.

Network calculus

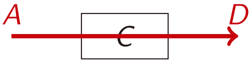
- Theory introduced in the 1990's by R.L. Cruz and Alain Jean-Marie then developed and popularized by C.S. Chang and J.-Y. Le Boudec.
- Filtering theory in the (min,plus) algebra.
- Applications:
  - Internet: video transmission (VoD),
  - Load-balancing in switches [Birkhoff-von Neumann switches, C.S. Chang]
  - Embedded systems: AFDX (Avionics Full Duplex) [Rockwell-Collins software used to certify A380], Networks-on-chip
- Recent trend to using this in 5G network that have strong latency and reliability requirements.
- Extentions / variations:
  - Real-Time Calculus [L. Thiele, S. Chakraborty]
  - Extended to Stochastic network calculus [C.S. Chang, Y.M. Jiang, F. Ciucu, J. Schmitt, M. Fidler]

Network calculus

- Theory introduced in the 1990's by R.L. Cruz and Alain Jean-Marie then developed and popularized by C.S. Chang and J.-Y. Le Boudec.
- Filtering theory in the (min,plus) algebra.
- Applications:
  - Internet: video transmission (VoD),
  - Load-balancing in switches [Birkhoff-von Neumann switches, C.S. Chang]
  - Embedded systems: AFDX (Avionics Full Duplex) [Rockwell-Collins software used to certify A380], Networks-on-chip
- Recent trend to using this in 5G network that have strong latency and reliability requirements.
- Extentions / variations:
  - Real-Time Calculus [L. Thiele, S. Chakraborty]
  - Extended to Stochastic network calculus [C.S. Chang, Y.M. Jiang, F. Ciucu, J. Schmitt, M. Fidler]

Contents

Server model



Arrival process

$A(s,t)$ amount of arrivals between time s and t: if a_i is the amount of data arriving during time slot i, $i \geq 1$,

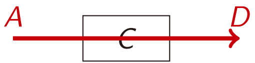
$$A(s,t) = \sum_{i=s}^{t-1} a_i.$$

Service process

$C(s,t)$ amount of service offered between time s and t: if c_i is the amount of data arriving during time slot i, $i \geq 1$,

$$C(s,t) = \sum_{i=s}^{t-1} c_i.$$

Server model



Arrival process

$A(s,t)$ amount of arrivals between time s and t: if a_i is the amount of data arriving during time slot i, $i \geq 1$,

$$A(s,t) = \sum_{i=s}^{t-1} a_i.$$

Service process

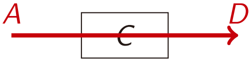
$C(s,t)$ amount of service offered between time s and t: if c_i is the amount of data arriving during time slot i, $i \geq 1$,

$$C(s,t) = \sum_{i=s}^{t-1} c_i.$$

Queue length (or backlog):

$q(t)$

Server model



Arrival process

$A(s,t)$ amount of arrivals between time s and t: if a_i is the amount of data arriving during time slot i, $i \geq 1$,

$$A(s,t) = \sum_{i=s}^{t-1} a_i.$$

Service process

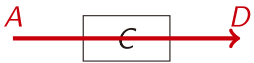
$C(s,t)$ amount of service offered between time s and t: if c_i is the amount of data arriving during time slot i, $i \geq 1$,

$$C(s,t) = \sum_{i=s}^{t-1} c_i.$$

Queue length (or backlog):

$$q(t) = \max(q(t-1) + a_{t-1} - c_{t-1}, 0)$$

Server model



Arrival process

$A(s,t)$ amount of arrivals between time s and t: if a_i is the amount of data arriving during time slot i, $i \geq 1$,

$$A(s,t) = \sum_{i=s}^{t-1} a_i.$$

Service process

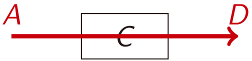
$C(s,t)$ amount of service offered between time s and t: if c_i is the amount of data arriving during time slot i, $i \geq 1$,

$$C(s,t) = \sum_{i=s}^{t-1} c_i.$$

Queue length (or backlog):

$$q(t) = \max(q(t-1) + a_{t-1} - c_{t-1}, 0) = \cdots = \max_{0 \leq s \leq t} A(s,t) - C(s,t)$$

Server model



Arrival process

$$A(s,t) = \sum_{i=s}^{t-1} a_i.$$

Service process

$$C(s,t) = \sum_{i=s}^{t-1} c_i.$$

Queue length (or backlog):

$$q(t) = \max(q(t-1) + a_{t-1} - c_{t-1}, 0) = \cdots = \max_{0 \le s \le t} A(s,t) - C(s,t)$$

Input/Output relation:

$$D(0,t) = A(0,t) - q(t) \ge \inf_{0 \le s \le t} A(0,s) + C(s,t).$$

Performance bounds

Backlog $q(t)$

$$q(t) > B \Leftrightarrow A(0,t) - D(0,t) > B \Leftrightarrow \inf_{s \le t} A(s,t) - C(s,t) > B.$$

Delay $d(t)$

$$d(t) > T \Leftrightarrow A(0,t) > D(0,t+T) \Leftrightarrow \inf_{s \le t} A(s,t) - C(s,t+T) > 0.$$

Characterization of the departure process $D(s,t)$

$$D(s,t) = \sup_{u \le s} A(u,t) - C(s,t)$$

Different models of Stochastic Network Calculus (SNC)

1. Tailbound-SNC:

$$\mathbb{P}(A(t) - A(s) > r(t-s) + b) \leq \epsilon(b, t-s).$$

- Introduced in [Yaron and Sidi, 93], and focus of the book [Jiang 08]
- (min,plus) computations can be adapted with smart use of the Boole inequality.
- loosy error bounds

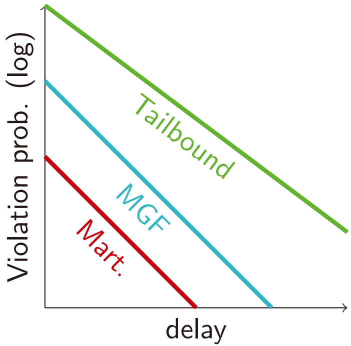
2. Moment-Generating Function-SNC:

$$\mathbb{E}[e^{\theta A(s,t)}] \leq e^{\theta(\sigma_A(\theta)+\rho_A(\theta)(t-s))}.$$

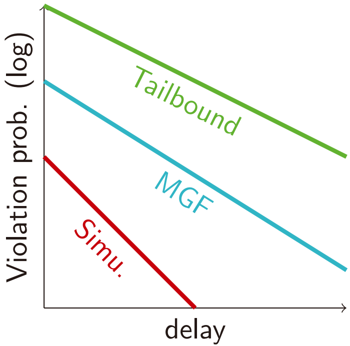
- Introduced in [Chang, 00], and focus of [Rizk and Fidler, 12]
- Martingale and Doob's inequality approach [Ciucu, 15]
- Not independent processes, network analysis more complex (Hölder's inequality) [Beck, 16; Nikolaus, Schmitt, 17]

Comparison of the models

Model	Tailbound	MGF	Martingale
Quality of the bound	Very poor	A little better	Quasi-tight
Topologies	Feed-forward	Feed-forward	1 server
Scheduling policies	All from DNC	Blind	FIFO, SP, EDF



one server



several servers

Moment generating functions and $(\sigma(\theta), \rho(\theta))$-constraints

Arrival $(\sigma_A(\theta), \rho_A(\theta))$-constraint

$$\forall s \leq t,\ \mathbb{E}[e^{\theta A(s,t)}] \leq e^{\theta(\sigma_A(\theta)+\rho_A(\theta)(t-s))}.$$

Service $(\sigma_C(\theta), \rho_C(\theta))$-constraint

$$\forall s \leq t,\ \mathbb{E}[e^{-\theta C(s,t)}] \leq e^{\theta(\sigma_C(\theta)-\rho_C(\theta)(t-s))}.$$

Examples of arrival processes

I.i.d. processes

- $\mathbb{E}[e^{\theta A(s,t)}] = \mathbb{E}[e^{\theta a_1}]^{t-s}$
- $\sigma(\theta) = 0$ and $\rho(\theta) = \frac{\ln \mathbb{E}[e^{\theta a_1}]}{\theta}$

Examples of arrival processes

I.i.d. processes

- $\mathbb{E}[e^{\theta A(s,t)}] = \mathbb{E}[e^{\theta a_1}]^{t-s}$
- $\sigma(\theta) = 0$ and $\rho(\theta) = \frac{\ln \mathbb{E}[e^{\theta a_1}]}{\theta}$

On-Off modulated Markov processes (MMOO)

- $\{X(t)\}$ a two-state Markov-Chain (ON and OFF states), $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$
- d_{ON} a distribution
- $a_t = \begin{cases} 0 & \text{if } X(t) = OFF \\ \sim d_{ON} & \text{if } X(t) = ON \end{cases}$
- $\rho(\theta)$ and $\sigma(\theta)$ are computed using the spectral analysis of $\begin{pmatrix} 1-p & p \\ q\mathbb{E}[e^{\theta d_{ON}}] & (1-q)\mathbb{E}[e^{\theta d_{ON}}] \end{pmatrix}$

More generally, the Markov-modulated processes can be analyzed

Computing the violation probability of a backlog bound

$$\mathbb{P}(q(t) > B) = \mathbb{P}(\max_{0 \leq s \leq t} A(s,t) - C(s,t) > B)$$
$$= \mathbb{P}(\cup_{0 \leq s \leq t} \{A(s,t) - C(s,t) > B\})$$

Computing the violation probability of a backlog bound

$$\mathbb{P}(q(t) > B) = \mathbb{P}(\max_{0 \le s \le t} A(s,t) - C(s,t) > B)$$
$$= \mathbb{P}(\cup_{0 \le s \le t} \{A(s,t) - C(s,t) > B\})$$

Some probabilistic inequalities

1. Boole inequality (aka union bound): $\mathbb{P}(\cup_{i \ge 0} E_i) \le \sum_{i \ge 0} \mathbb{P}(E_i)$
2. Chernoff Bound: $\mathbb{P}(X \ge a) \le \inf_{\theta > 0} \mathbb{E}[e^{\theta X}] e^{-\theta a}$

Computing the violation probability of a backlog bound

$$\mathbb{P}(q(t) > B) = \mathbb{P}(\max_{0\leq s\leq t} A(s,t) - C(s,t) > B)$$
$$= \mathbb{P}(\cup_{0\leq s\leq t}\{A(s,t) - C(s,t) > B\})$$

Some probabilistic inequalities

1. Boole inequality (aka union bound): $\mathbb{P}(\cup_{i\geq 0} E_i) \leq \sum_{i\geq 0} \mathbb{P}(E_i)$
2. Chernoff Bound: $\mathbb{P}(X \geq a) \leq \inf_{\theta>0} \mathbb{E}[e^{\theta X}]e^{-\theta a}$

The combination of the two inequalities is powerful for the probabilistic method (prove the existence of combinatorial objects with probabilistic tools): proving that some probabilities tend to 0 with the size of the object is enough.

Computing the violation probability of a backlog bound

$$\mathbb{P}(q(t) > B) = \mathbb{P}(\max_{0 \le s \le t} A(s,t) - C(s,t) > B)$$
$$= \mathbb{P}(\cup_{0 \le s \le t} \{A(s,t) - C(s,t) > B\})$$

Some probabilistic inequalities

1. Boole inequality (aka union bound): $\mathbb{P}(\cup_{i \ge 0} E_i) \le \sum_{i \ge 0} \mathbb{P}(E_i)$
2. Chernoff Bound: $\mathbb{P}(X \ge a) \le \inf_{\theta > 0} \mathbb{E}[e^{\theta X}] e^{-\theta a}$

The combination of the two inequalities is powerful for the probabilistic method (prove the existence of combinatorial objects with probabilistic tools): proving that some probabilities tend to 0 with the size of the object is enough.

For performance evaluation, the goal is to find a good approximation of probabilities, and these two inequalities are not enough.

Moment-generating-function SNC (MGF-SNC)

[Fidler 2006, Schmitt et al. 2010–]

If the arrivals and services are independent:

$$\mathbb{P}(q(t) > B) \leq \mathbb{P}(\cup_{s \leq t}\{A(s,t) - C(s,t) > B\})$$

Moment-generating-function SNC (MGF-SNC)

[Fidler 2006, Schmitt et al. 2010–]

If the arrivals and services are independent:

$$\mathbb{P}(q(t) > B) \leq \mathbb{P}(\cup_{s \leq t}\{A(s,t) - C(s,t) > B\})$$

$$\text{Union Bound} \leq \sum_{s \leq t} \mathbb{P}(\{A(s,t) - C(s,t) > B\})$$

Moment-generating-function SNC (MGF-SNC)

[Fidler 2006, Schmitt et al. 2010–]

If the arrivals and services are independent:

$$\mathbb{P}(q(t) > B) \leq \mathbb{P}(\cup_{s\leq t}\{A(s,t) - C(s,t) > B\})$$

$$\text{Union Bound} \leq \sum_{s\leq t} \mathbb{P}(\{A(s,t) - C(s,t) > B\})$$

$$\text{Chernoff + Independence} \leq \sum_{s\leq t} \mathbb{E}[e^{\theta A(s,t)}]\mathbb{E}[e^{-\theta C(s,t)}]e^{-\theta B}$$

Moment-generating-function SNC (MGF-SNC)

[Fidler 2006, Schmitt et al. 2010–]

If the arrivals and services are independent:

$$\begin{aligned}
\mathbb{P}(q(t) > B) &\leq \mathbb{P}(\cup_{s\leq t}\{A(s,t) - C(s,t) > B\}) \\
\text{Union Bound} &\leq \sum_{s\leq t} \mathbb{P}(\{A(s,t) - C(s,t) > B\}) \\
\text{Chernoff + Independence} &\leq \sum_{s\leq t} \mathbb{E}[e^{\theta A(s,t)}]\mathbb{E}[e^{-\theta C(s,t)}]e^{-\theta B} \\
\text{NC constraints} &\leq \sum_{s\leq t} e^{\theta(\sigma_A(\theta)+\rho_A(\theta)(t-s))} e^{\theta(\sigma_C(\theta)-\rho_C(\theta)(t-s))} e^{-\theta B}
\end{aligned}$$

Moment-generating-function SNC (MGF-SNC)

[Fidler 2006, Schmitt et al. 2010–]

If the arrivals and services are independent:

$$
\begin{aligned}
\mathbb{P}(q(t) > B) &\leq \mathbb{P}(\cup_{s\leq t}\{A(s,t) - C(s,t) > B\}) \\
\text{Union Bound} &\leq \sum_{s\leq t} \mathbb{P}(\{A(s,t) - C(s,t) > B\}) \\
\text{Chernoff + Independence} &\leq \sum_{s\leq t} \mathbb{E}[e^{\theta A(s,t)}]\mathbb{E}[e^{-\theta C(s,t)}]e^{-\theta B} \\
\text{NC constraints} &\leq \sum_{s\leq t} e^{\theta(\sigma_A(\theta)+\rho_A(\theta)(t-s))} e^{\theta(\sigma_C(\theta)-\rho_C(\theta)(t-s))} e^{-\theta B} \\
&\leq \frac{e^{\theta(\sigma_A(\theta)+\sigma_C(\theta))}}{1 - e^{\theta(\rho_A(\theta)-\rho_C(\theta))}} e^{-\theta B}
\end{aligned}
$$

Moment-generating-function SNC (MGF-SNC)

[Fidler 2006, Schmitt et al. 2010–]

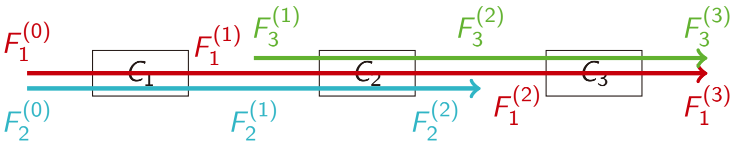
If the arrivals and services are independent:

$$
\begin{aligned}
\mathbb{P}(q(t) > B) &\leq \mathbb{P}(\cup_{s\leq t}\{A(s,t) - C(s,t) > B\}) \\
\text{Union Bound} &\leq \sum_{s\leq t} \mathbb{P}(\{A(s,t) - C(s,t) > B\}) \\
\text{Chernoff + Independence} &\leq \sum_{s\leq t} \mathbb{E}[e^{\theta A(s,t)}]\mathbb{E}[e^{-\theta C(s,t)}]e^{-\theta B} \\
\text{NC constraints} &\leq \sum_{s\leq t} e^{\theta(\sigma_A(\theta)+\rho_A(\theta)(t-s))} e^{\theta(\sigma_C(\theta)-\rho_C(\theta)(t-s))} e^{-\theta B} \\
&\leq \frac{e^{\theta(\sigma_A(\theta)+\sigma_C(\theta))}}{1 - e^{\theta(\rho_A(\theta)-\rho_C(\theta))}} e^{-\theta B}
\end{aligned}
$$

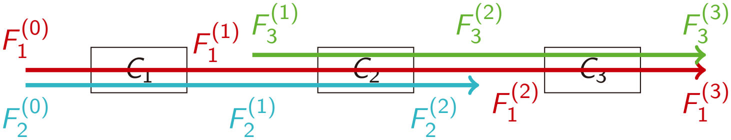
- The larger θ, the larger the decay rate
- The value of θ is constrained by $\rho_A(\theta) - \rho_C(\theta) < 0$

Contents

Previous MGF-SNC approach

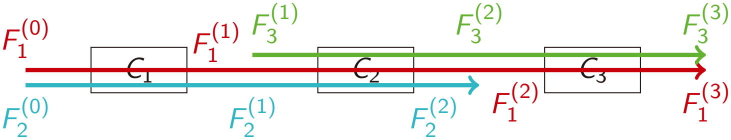


Previous MGF-SNC approach



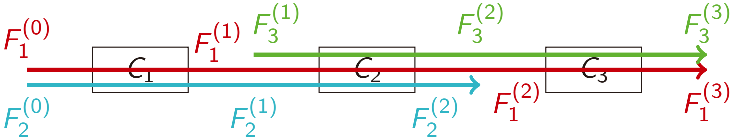
1. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(1)}$ and $F_2^{(1)}$: both depend on $F_1^{(0)}$, $F_2^{(0)}$ and C_1: they are not independent anymore

Previous MGF-SNC approach



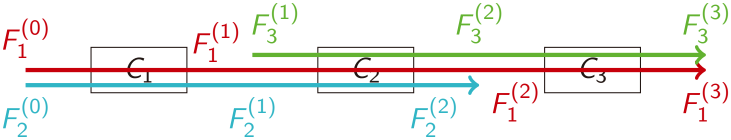
1. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(1)}$ and $F_2^{(1)}$: both depend on $F_1^{(0)}$, $F_2^{(0)}$ and C_1: they are not independent anymore
2. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(2)}$ and $F_3^{(2)}$: both depend on $F_1^{(1)}$, $F_2^{(1)}$, $F_3^{(1)}$ and C_2. The previous computation cannot be used because $F_1^{(1)}$ and $F_2^{(1)}$ are not independent

Previous MGF-SNC approach



1. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(1)}$ and $F_2^{(1)}$: both depend on $F_1^{(0)}$, $F_2^{(0)}$ and C_1: they are not independent anymore
2. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(2)}$ and $F_3^{(2)}$: both depend on $F_1^{(1)}$, $F_2^{(1)}$, $F_3^{(1)}$ and C_2. The previous computation cannot be used because $F_1^{(1)}$ and $F_2^{(1)}$ are not independent
3. Use the Hölder inequality instead. Increase the inaccuracy of the bounds

Previous MGF-SNC approach



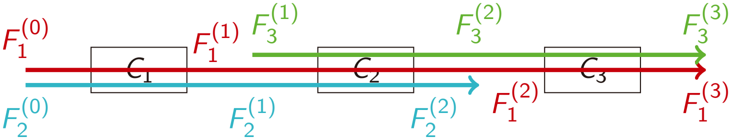
1. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(1)}$ and $F_2^{(1)}$: both depend on $F_1^{(0)}$, $F_2^{(0)}$ and C_1: they are not independent anymore
2. Compute the $(\sigma(\theta), \rho(\theta))$ characterization for $F_1^{(2)}$ and $F_3^{(2)}$: both depend on $F_1^{(1)}$, $F_2^{(1)}$, $F_3^{(1)}$ and C_2. The previous computation cannot be used because $F_1^{(1)}$ and $F_2^{(1)}$ are not independent
3. Use the Hölder inequality instead. Increase the inaccuracy of the bounds

Improvement tricks:

- Flow prolongation
- Lyapunov inequality

Pay multiplexing only once

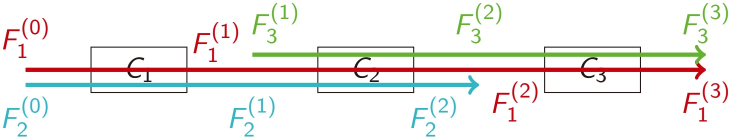
Idea: Directly compute the left-over service curve for flow 1 on its path.



Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.

$\exists \quad t_3 \leq t_4,$

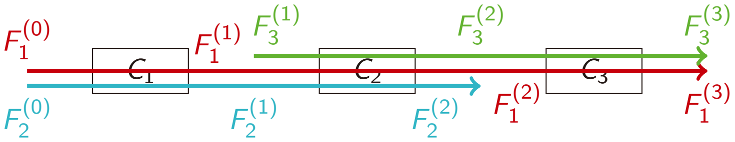


$$F_1^{(3)}(0, t_4) + F_3^{(3)}(0, t_4) \geq F_1^{(2)}(0, t_3) + F_3^{(2)}(0, t_3) + C_3(t_3, t_4)$$

Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.

$\exists \quad t_2 \leq t_3 \leq t_4,$

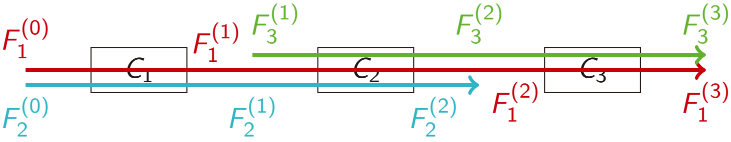


$$F_1^{(3)}(0,t_4) + F_3^{(3)}(0,t_4) \geq F_1^{(2)}(0,t_3) + F_3^{(2)}(0,t_3) + C_3(t_3,t_4)$$

$$F_1^{(2)}(0,t_3) + F_2^{(2)}(0,t_3) + F_3^{(2)}(0,t_3) \geq F_1^{(1)}(0,t_2) + F_2^{(1)}(0,t_2) + F_3^{(1)}(0,t_2) + C_2(t_2,t_3)$$

Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.
$\exists t_1 \leq t_2 \leq t_3 \leq t_4,$

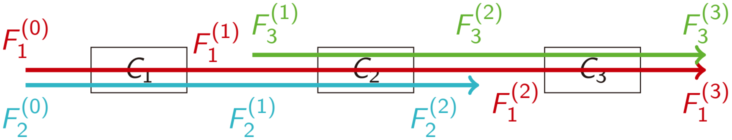


$$F_1^{(3)}(0,t_4) + F_3^{(3)}(0,t_4) \geq F_1^{(2)}(0,t_3) + F_3^{(2)}(0,t_3) + C_3(t_3,t_4)$$
$$F_1^{(2)}(0,t_3) + F_2^{(2)}(0,t_3) + F_3^{(2)}(0,t_3) \geq F_1^{(1)}(0,t_2) + F_2^{(1)}(0,t_2) + F_3^{(1)}(0,t_2) + C_2(t_2,t_3)$$
$$F_1^{(1)}(0,t_2) + F_2^{(1)}(0,t_2) \geq F_1^{(0)}(0,t_1) + F_2^{(0)}(0,t_1) + C_1(t_1,u)$$

Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.

$\exists t_1 \leq t_2 \leq t_3 \leq t_4,$

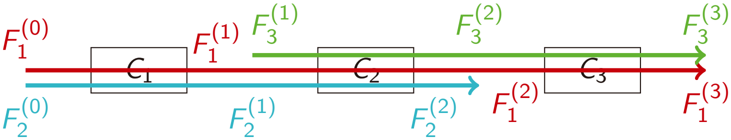


$$
\begin{array}{r}
F_1^{(3)}(0,t_4) + F_3^{(3)}(0,t_4) \geq F_1^{(2)}(0,t_3) + F_3^{(2)}(0,t_3) + C_3(t_3,t_4) \\
F_1^{(2)}(0,t_3) + F_2^{(2)}(0,t_3) + F_3^{(2)}(0,t_3) \geq F_1^{(1)}(0,t_2) + F_2^{(1)}(0,t_2) + F_3^{(1)}(0,t_2) + C_2(t_2,t_3) \\
F_1^{(1)}(0,t_2) + F_2^{(1)}(0,t_2) \geq F_1^{(0)}(0,t_1) + F_2^{(0)}(0,t_1) + C_1(t_1,u) \\
\hline\hline
F_1^{(3)}(0,t_4) + F_3^{(3)}(0,t_4) + F_2^{(3)}(0,t_3) \geq F_1^{(0)}(0,t_1) + F_2^{(0)}(0,t_1) + F_3^{(1)}(0,t_2) \\
+ C_1(t_1,t_2) + C_2(t_2,t_3) + C_3(t_3,t_4)
\end{array}
$$

Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.

$\exists t_1 \leq t_2 \leq t_3 \leq t_4,$



$$F_1^{(3)}(0, t_4) + F_3^{(3)}(0, t_4) \geq F_1^{(2)}(0, t_3) + F_3^{(2)}(0, t_3) + C_3(t_3, t_4)$$

$$F_1^{(2)}(0, t_3) + F_2^{(2)}(0, t_3) + F_3^{(2)}(0, t_3) \geq F_1^{(1)}(0, t_2) + F_2^{(1)}(0, t_2) + F_3^{(1)}(0, t_2) + C_2(t_2, t_3)$$

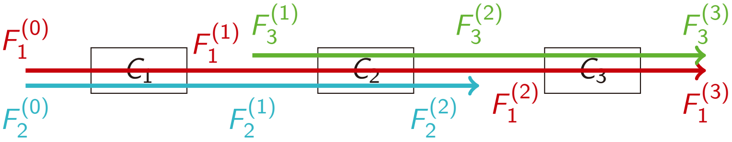
$$F_1^{(1)}(0, t_2) + F_2^{(1)}(0, t_2) \geq F_1^{(0)}(0, t_1) + F_2^{(0)}(0, t_1) + C_1(t_1, u)$$

$$\begin{aligned} F_1^{(3)}(0, t_4) + F_3^{(3)}(0, t_4) + F_2^{(3)}(0, t_3) \geq\ & F_1^{(0)}(0, t_1) + F_2^{(0)}(0, t_1) + F_3^{(1)}(0, t_2) \\ & + C_1(t_1, t_2) + C_2(t_2, t_3) + C_3(t_3, t_4) \end{aligned}$$

$$F_1^{(3)}(0, t_4) \geq F_1^{(0)}(0, t_1) + C_1(t_1, t_2) + C_2(t_2, t_3) + C_3(t_3, t_4) - F_2^{(0)}(t_1, t_3) - F_3^{(0)}(t_2, t_4).$$

Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.



Theorem

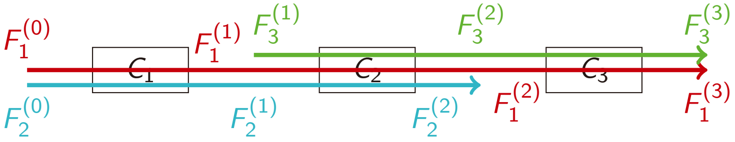
The end-to-end service for flow 1 is

$$C_{e2e}(t_1, t_4) = \Big(\inf_{t_1 \le t_2 \le t_3 \le t_4} A_1(0, t_1) + C_1(t_1, t_2) + C_2(t_2, t_3) + C_3(t_3, t_4) - A_2(t_1, t_3) - A_3(t_2, t_4)\Big)_+$$

$$(x)_+ = \max(0, x)$$

Pay multiplexing only once

Idea: Directly compute the left-over service curve for flow 1 on its path.



Theorem

The end-to-end service for flow 1 is

$$C_{e2e}(t_1,t_4) = (\inf_{t_1 \le t_2 \le t_3 \le t_4} A_1(0,t_1) + C_1(t_1,t_2) + C_2(t_2,t_3) + C_3(t_3,t_4) - A_2(t_1,t_3) - A_3(t_2,t_4))_+$$

$$(x)_+ = \max(0,x)$$

$\mathbb{E}[e^{-\theta C_{e2e}(u,t)}]$ cannot be expressed with a $(\sigma(\theta),\rho(\theta))$-constraint.

Bounding Generating function for the end-to-end service

Use generating functions to generalize the $(\sigma(\theta), \rho(\theta))$-constraints

$$\forall s \leq t,\ \mathbb{E}[e^{-\theta C(s,t)}] \leq e^{-\theta\sigma_C(\theta)+(t-s)\rho_C(\theta)} \quad \rightsquigarrow \quad F_C(\theta, z) = \frac{e^{\theta\sigma(\theta)}}{1 - ze^{-\theta\rho_C(\theta)}}$$

Delay, backlog violation probabilities and departure processes can be expressed using generating functions

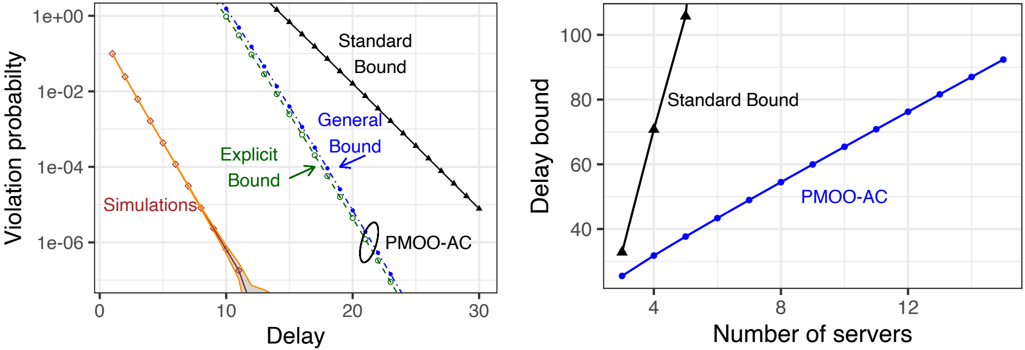
Service bounding generating function for the end-to-end service

$$F_{C_{e2e}}(\theta, z) = \frac{e^{\theta(\sigma_{S_1}+\sigma_{A_2})}}{1 - e^{-\theta(\rho_{C_1}-\rho_{A_2})}z} \frac{e^{\theta(\sigma_{S_2}+\sigma_{A_3})}}{1 - e^{-\theta(\rho_{C_2}-\rho_{A_2}-\rho_{A_3})}z} \frac{e^{\theta\sigma_{S_3}}}{1 - e^{-\theta(\rho_{C_3}-\rho_{A_3})}z}.$$

Computing the performance bound bounds

1. Use $F_{C_{e2e}}(\theta, z)$ for computing the delay or backlog bound
2. For the delay, the formula can be simplified if $F_{C_{e2e}}$ has multiple singularities (basic but tedious computations) to a single singularity
3. optimization of θ:
   - For each server j , there exists a maximum value of θ_j such that $e^{-\theta(\rho_{C_j} - \rho_{A(j)})} \leq 1$
   - One need to choose θ such that $\theta < \min_j \theta_j$, minimizing the performance bound
4. Only one parameter to optimize!

Numerical evaluation



1. The new bound catches the decay rate of the violation probability
2. Easier and faster to compute than the standard ones
3. Still not very satisfactory (the gap with simulation is still too large)

Contents

Martingale SNC

[Chang 2000, Ciucu 2014, Duffield 1994, Kingman 1963]

There exists a vector v such that

$$M(\theta, s, t) = v(X_s)e^{\theta(A(s,t) - C(s,t) - (\rho_A(\theta) - \rho_C(\theta))(t-s))}$$

is a backward martingale: $\mathbb{E}[M(\theta, s-1, t) \mid \mathcal{F}_{u \geq s}] = M(\theta, s, t)$

Martingale SNC

[Chang 2000, Ciucu 2014, Duffield 1994, Kingman 1963]

There exists a vector v such that

$$M(\theta, s, t) = v(X_s)e^{\theta(A(s,t)-C(s,t)-(\rho_A(\theta)-\rho_C(\theta))(t-s))}$$

is a backward martingale: $\mathbb{E}[M(\theta, s-1, t) \mid \mathcal{F}_{u\geq s}] = M(\theta, s, t)$

From Doob's inequality for super-martingales, there exists a relatively small constant $\varphi(\theta)$ (depending on v) such that

$$\begin{aligned}\mathbb{P}(q(t) > B) &\leq \mathbb{P}(\sup_{s\leq t} e^{\theta(A(s,t)-C(s,t))} > e^{\theta B}) \\ &\leq \varphi(\theta)e^{-\theta B}.\end{aligned}$$

This formula is valid for all θ such that $\rho_A(\theta) \leq \rho_C(\theta)$.

HUAWEI

A martingale for the arrivals and service processes

To avoid technicalities/heavy notations, we focus on the i.i.d. cases (but no difficulty for MMPs)

Let $A(s,t)$ be an arrival process, $A(s,t) = \sum_{i=s}^{t-1} a_i$, where (a_i) is i.i.d. Let $e^{\theta\rho_A(\theta)} = \mathbb{E}[e^{\theta a_1}]$ the MGF of a_1.

Martingale associated to A

$$M_A(\theta,s,t) = e^{\theta A(s,t)} e^{-\theta\rho_A(\theta)(t-s)}$$

Backward martingale: $\mathbb{E}[M_A(\theta,s-1,t) \mid \mathcal{F}_{u\geq s}] = M_A(\theta,s,t)$.

A martingale for the arrivals and service processes

To avoid technicalities/heavy notations, we focus on the i.i.d. cases (but no difficulty for MMPs)

Let $A(s,t)$ be an arrival process, $A(s,t) = \sum_{i=s}^{t-1} a_i$, where (a_i) is i.i.d. Let $e^{\theta \rho_A(\theta)} = \mathbb{E}[e^{\theta a_1}]$ the MGF of a_1.

Martingale associated to A

$$M_A(\theta, s, t) = e^{\theta A(s,t)} e^{-\theta \rho_A(\theta)(t-s)}$$

Backward martingale: $\mathbb{E}[M_A(\theta, s-1, t) \mid \mathcal{F}_{u \geq s}] = M_A(\theta, s, t)$.

Let $C(s,t)$ be a service process, $C(s,t) = \sum_{i=s}^{t-1} c_i$, where (c_i) is i.i.d. Let $e^{-\theta \rho_C(\theta)} = \mathbb{E}[e^{-\theta c_1}]$ the MGF of c_1.

Martingale associated to C

$$M_C(\theta, s, t) = e^{-\theta C(s,t)} e^{\theta \rho_C(\theta)(t-s)}$$

HUAWEI

Product of independent martingales

Product of independent martingales

PMOO formula for the backlog:

$$\sup_{t_1 \le t_2 \le t_3} A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3)$$

$$\{q(t_3) > B\} \le \{\exists t_1 \le t_2 \le t_3,\ A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3) > B\}$$

Product of independent martingales

PMOO formula for the backlog:

$$\sup_{t_1 \le t_2 \le t_3} A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3)$$

$$\{q(t_3) > B\} \le \{\exists t_1 \le t_2 \le t_3,\ A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3) > B\}$$

- A martingale cannot be defined directly for $A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3)$
- if t_2 is fixed, $(M_A(\theta, t_1, t_3) M_{C_1}(\theta, t_1, t_2))_{t_1 \le t_2}$ is a backward martingale

Backlog bound

Martingale-SNC

$$\mathbb{P}(q(t_3) > B) \leq \sum_{t_2 \leq t_3} \mathbb{P}(\sup_{t_1 \leq t_2} A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3) > B)$$

$$\leq \frac{e^{-\theta B}}{1 - e^{-\theta(\rho_{C_2} - \rho_A)}}.$$

Backlog bound

Martingale-SNC

$$\mathbb{P}(q(t_3) > B) \leq \sum_{t_2 \leq t_3} \mathbb{P}(\sup_{t_1 \leq t_2} A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3) > B)$$
$$\leq \frac{e^{-\theta B}}{1 - e^{-\theta(\rho_{C_2} - \rho_A)}}.$$

MGF-SNC

$$\mathbb{P}(q(t_3) > B) \leq \sum_{t_1 \leq t_2 \leq t_3} \mathbb{P}(A(t_1, t_3) - C_1(t_1, t_2) - C_2(t_2, t_3) > B)$$
$$\leq \frac{e^{-\theta B}}{(1 - e^{-\theta(\rho_{C_2} - \rho_A)})(1 - e^{-\theta(\rho_{C_1} - \rho_A)})}.$$

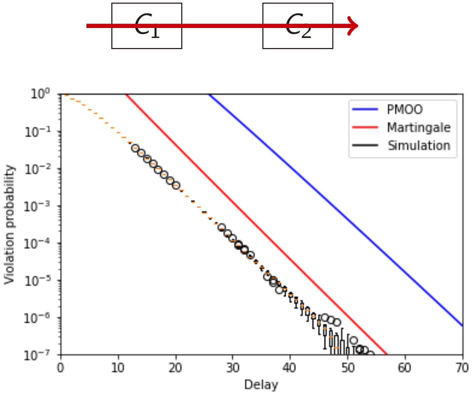
General framework the mixed martingale/MGF SNC

For any tandem network:

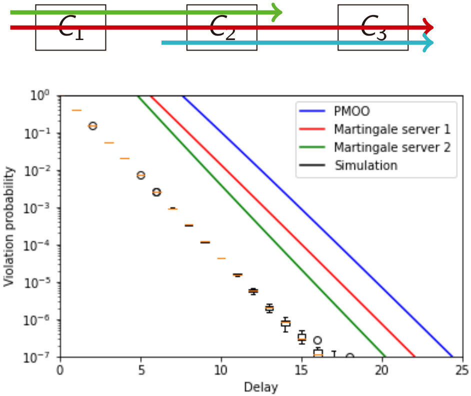
- Write the PMOO service
- Use the Doob's inequality can be applied locally at one server
  - at the first server is always possible
  - under some additional hypothesis (e.g. for sink trees with deterministic service) at any server
  - the best choice is to choose the server that minimizes $\sup\{\theta \mid \rho_{C_j}(\theta) \geq \sum_{i \in j} \rho_{A_i}(\theta)\}$
- Use the union bound at other places

Numerical evaluation

Two servers in tandem



Interleaved (bottlenck is server 2)



HUAWEI

Conclusion

Summary

- A new PMOO formula to improve the MGF-SNC:
  - More accurate bounds (catch the decay rate of the violations probabilities)
  - Faster to compute (no Hölder inequality, only one parameter to optimize)
- A first step to generalize the use of martingales in the multiple server case
  - Application on the first server
  - In some cases, possible to apply it to another server (e.g. in sink-tree tandems with constant-rate servers

Future work

- Generalize the PMOO to other service policies (for example to FIFO networks)
- Generalize to more processes (Auto-regressive)
- General application of the martingale bound to another server? to several servers?

  - New probabilistic tools needed...

Thank you.

Bring digital to every person, home and organization for a fully, connected, intelligent world.