Evaluating Session-Aware Admission Control Strategies to Improve the Profitability of Service Providers

Authors

Narjess AYARI	RESO – LIP
Denis Barbaron	Orange Labs
Laurent Lefèvre	RESO – INRIA - LIP (UMR 5668 CNRS, ENS, UCB)

Orange Labs (SIRP/ASF/INTL) and LIP/RESO (INRIA/UCBL/ENS de Lyon) – Université de Lyon, France.

The 3rd IEEE Workshop on Enabling the Future Service-Oriented Internet: Towards Socially-Aware Networks (EFSOI'09)

Held in conjunction with IEEE GLOBECOM 2009, Honolulu, HI, USA, December 4, 2009.

- Towards service oriented Internet
- Sessions : an instant aggregation of flows(data and control)
  - http sessions
  - VOIP sessions (when a costumer dialogs with an automatic vocal server, accounting)

- Problem : assuming QoS with a maximum number of sessions
- For service provider : provisioning enough resources
- Internet servers : when overloaded, they must cancel some sessions
- Cancelled sessions :
  - Not charged to customers
  - Waste of resources
  - Increase the number of angry customers

- Most state-of-the-art research on admission control advocate session oblivious mechanisms (packet level, flow level)
- From a service provider point of view, session unawareness results in profit loss & resources' waste

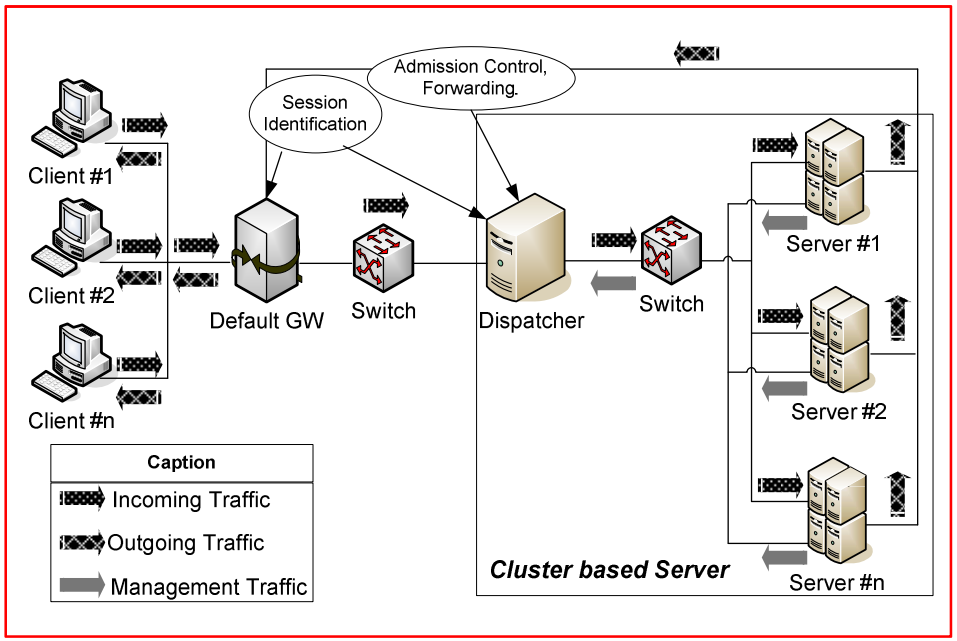
The objective of the operator is the mean monetary equivalent due to the blocking, the completion and the interruption of the offered sessions over a long time scale.

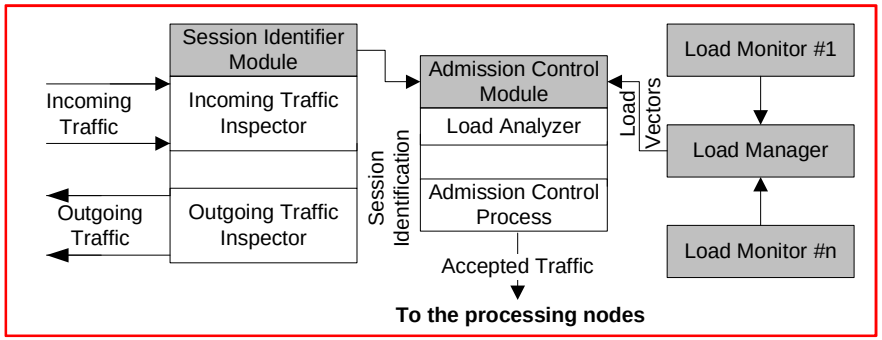
- We associate equivalent monetary values to:
  - ✓ the **good completion** of a session (R_c >0),
  - ✓ the **rejection/blocking** of a session (C_b >0),
  - ✓ the **interruption** of a session (C_i>0),
- **Economic Basis:** $C_i > C_b$

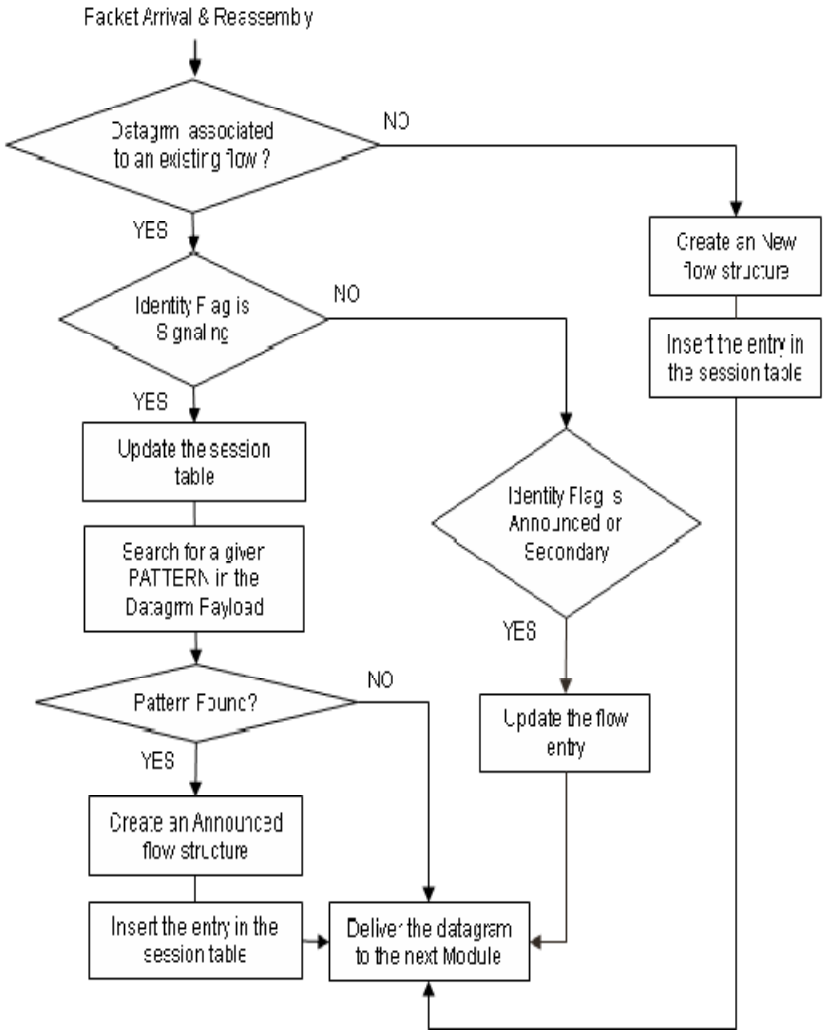
$$\max_p (R_c N_c - C_b N_b - C_i N_i)$$

N_c, N_b and N_i are respectively the mean number of completed, blocked and interrupted sessions per unit of time.

Implementation framework



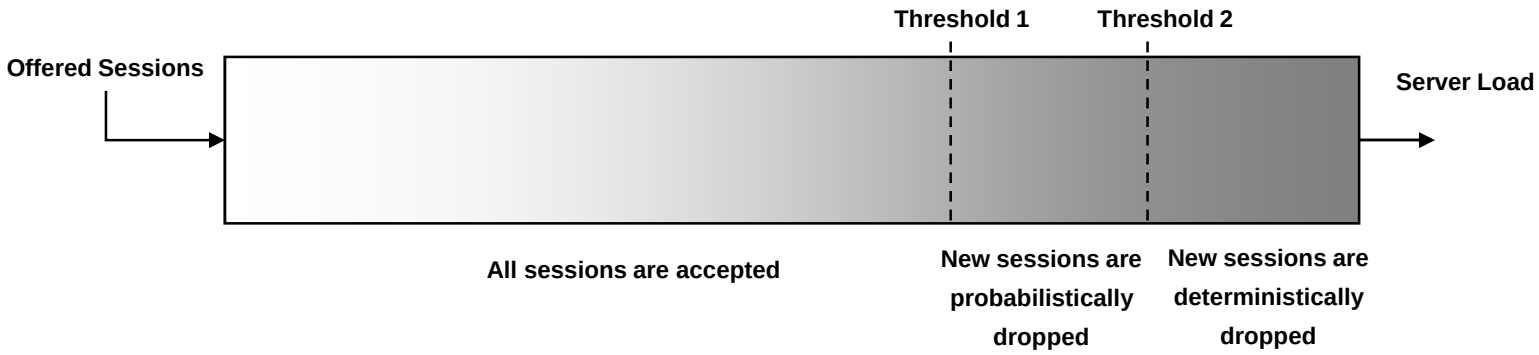




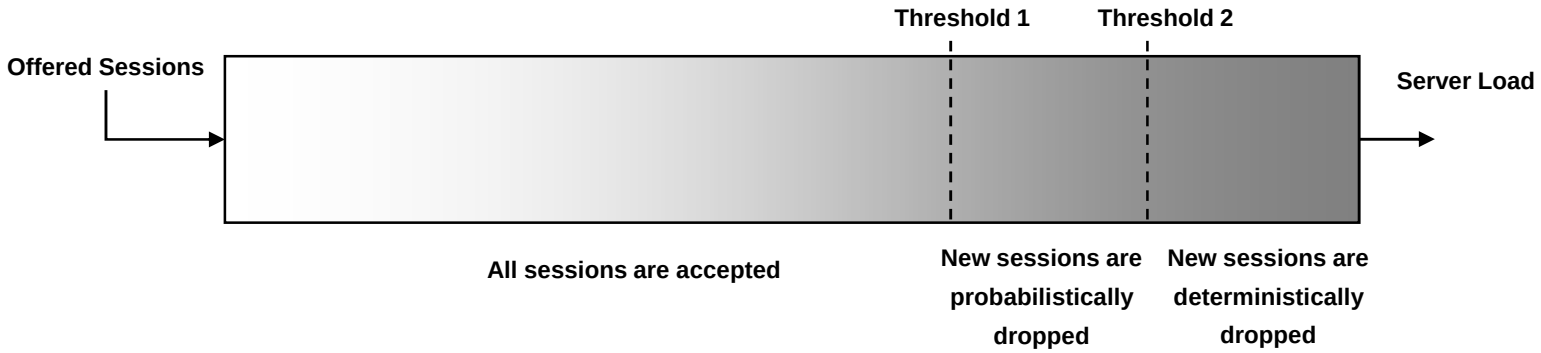
- Traffic pertaining to the active sessions are given a higher priority under overload !

A two threshold based approach

- Better to not accept a new session than cancelling it during its activity



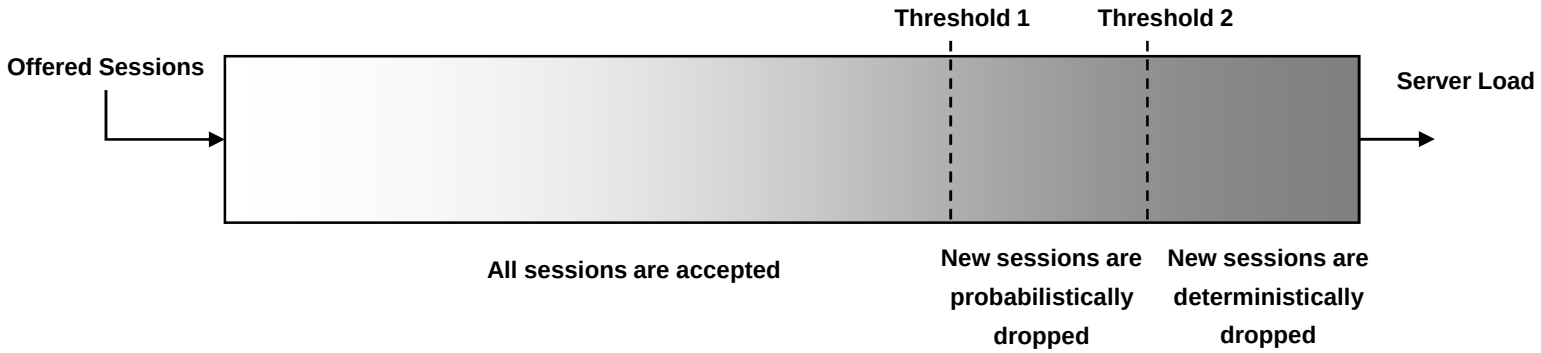
- Traffic pertaining to the active sessions are given a higher priority under overload !



With a low load server (under T1)

sessions are accepted : incoming traffic forwarded to processing entities

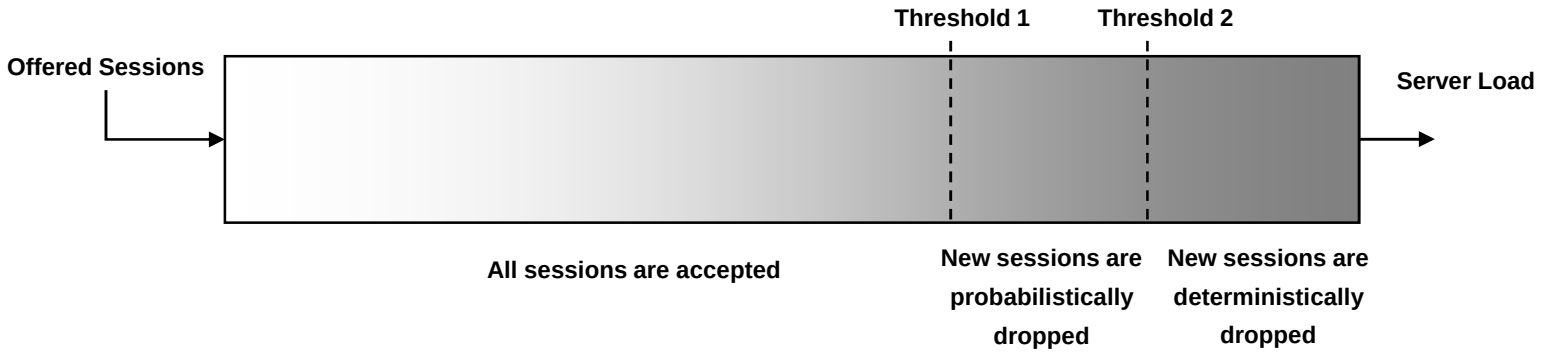
- Traffic pertaining to the active sessions are given a higher priority under overload !



When the server is facing some load (between T1 and T2)

incoming request is dropped with some probabilty

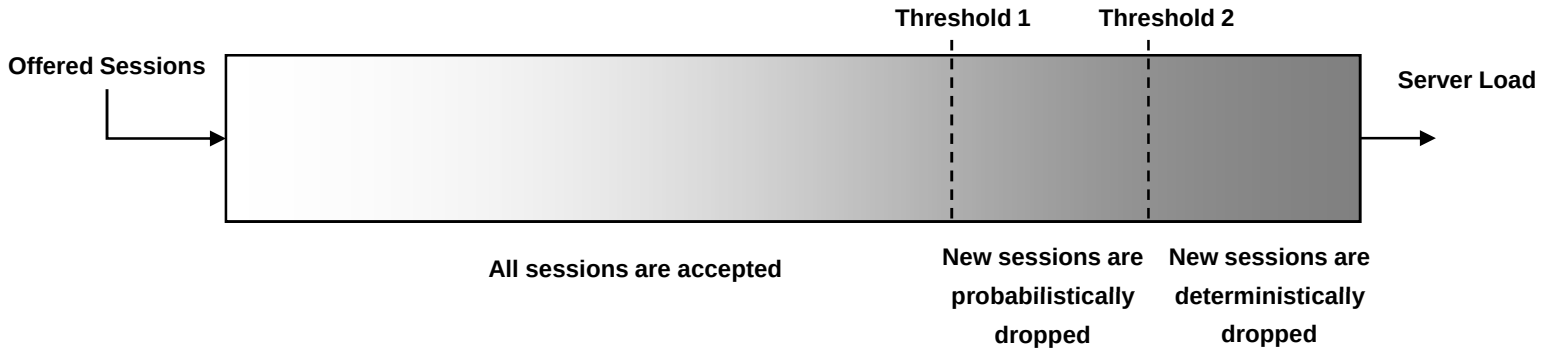
- Traffic pertaining to the active sessions are given a higher priority under overload !



When the server nearly overloaded (more than T2)

only traffic pertaining to already established sessions is processed

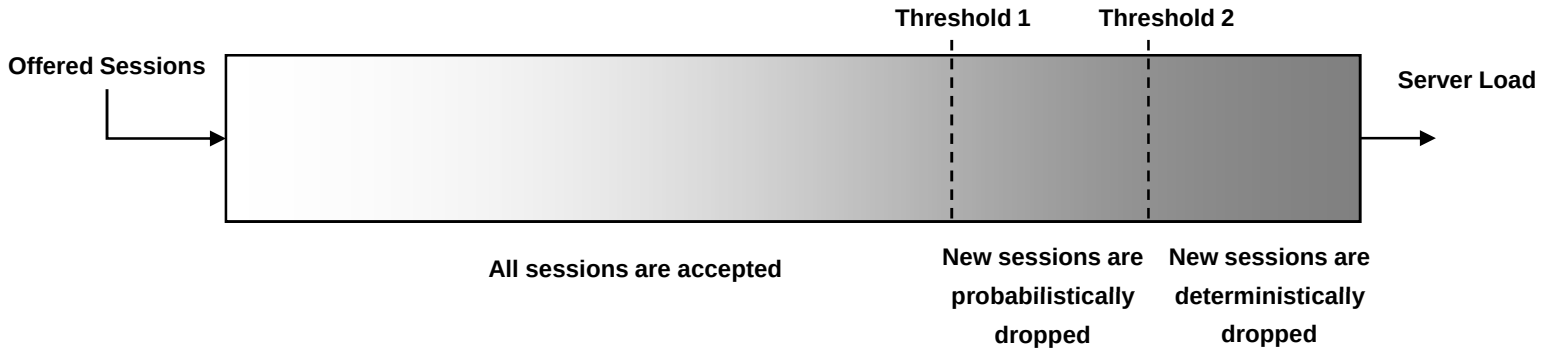
- Traffic pertaining to the active sessions are given a higher priority under overload !



A packet is dropped with a probability *p derived as a function of the instantaneously* measured server's load denoted by (*ls)*

$$p = \begin{cases} 0, & if\ l_S \leq T_1 \\ f(l_S),\ f(x) = \left\| \frac{x - T_1}{T_2 - T_1} \right\|, & if\ (New\ Session)\ and\ T_1 < l_S \leq T_2 \\ 1, & if\ (New\ Session)\ and\ T_2 < l_S \leq C \\ 1, & if\ l_S > C \end{cases},$$

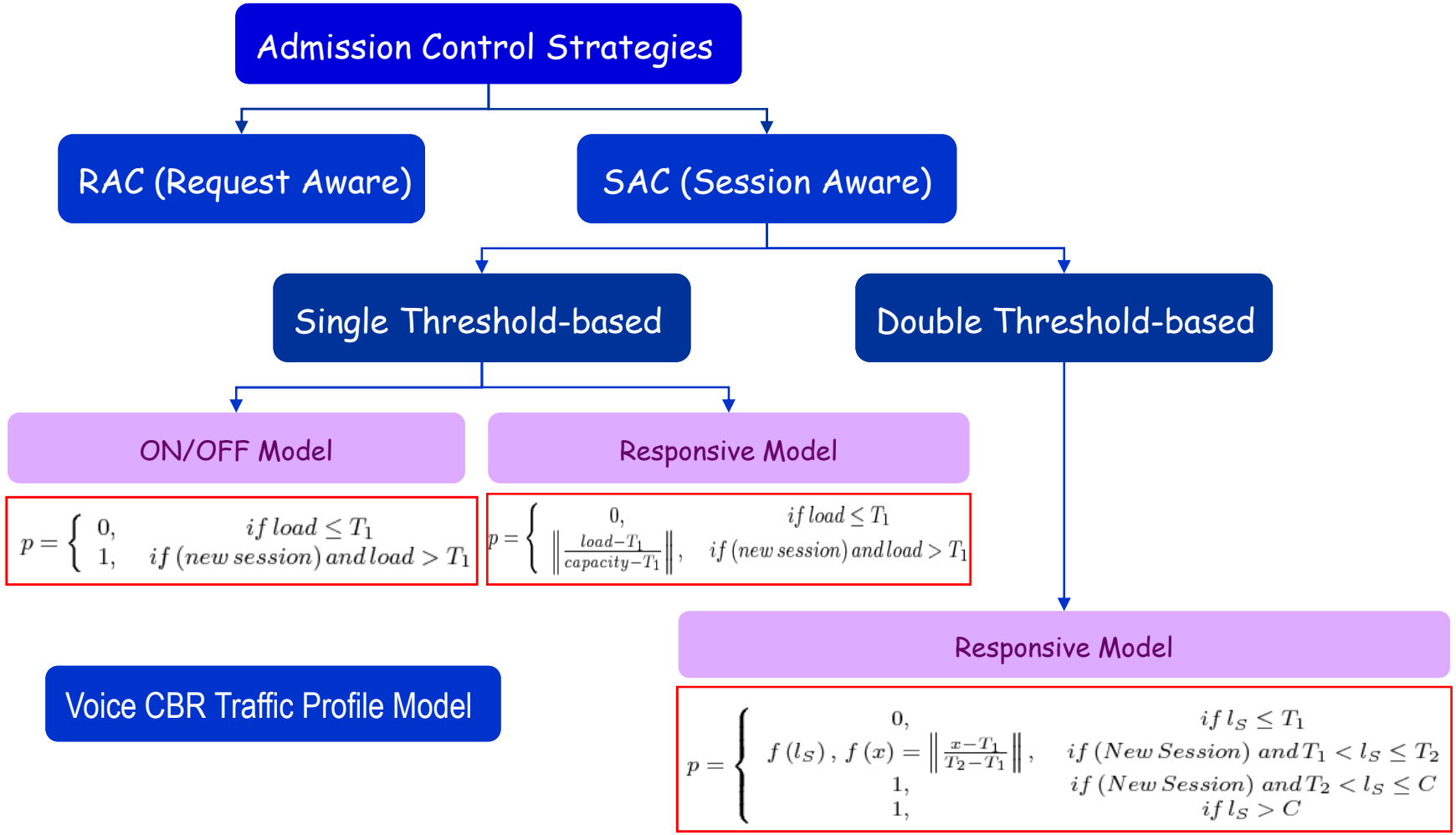
- Traffic pertaining to the active sessions are given a higher priority under overload !



Open questions :

Impact of AC in some scenario ?

Validate a double threshold compared to single one ?


Admission Control Strategies
RAC (Request Aware)
SAC (Session Aware)
Single Threshold-based
Double Threshold-based
ON/OFF Model
Responsive Model
$p = \begin{cases} 0, & if\ load \leq T_1 \\ 1, & if\ (new\ session)\ and\ load > T_1 \end{cases}$
$p = \begin{cases} 0, & if\ load \leq T_1 \\ \left\| \frac{load - T_1}{capacity - T_1} \right\|, & if\ (new\ session)\ and\ load > T_1 \end{cases}$
Responsive Model
Voice CBR Traffic Profile Model
$p = \begin{cases} 0, & if\ l_S \leq T_1 \\ f(l_S),\ f(x) = \left\| \frac{x - T_1}{T_2 - T_1} \right\|, & if\ (New\ Session)\ and\ T_1 < l_S \leq T_2 \\ 1, & if\ (New\ Session)\ and\ T_2 < l_S \leq C \\ 1, & if\ l_S > C \end{cases}$

- We consider two simulation scenarios

  - Generating homogeneous sessions over the same simulation run, having a duration that ranges from medium to long term.

  - Generating mixed sessions over the same simulation run, meaning that both medium term and long term conversations are equitably generated over the simulation time

Simulation Parameters

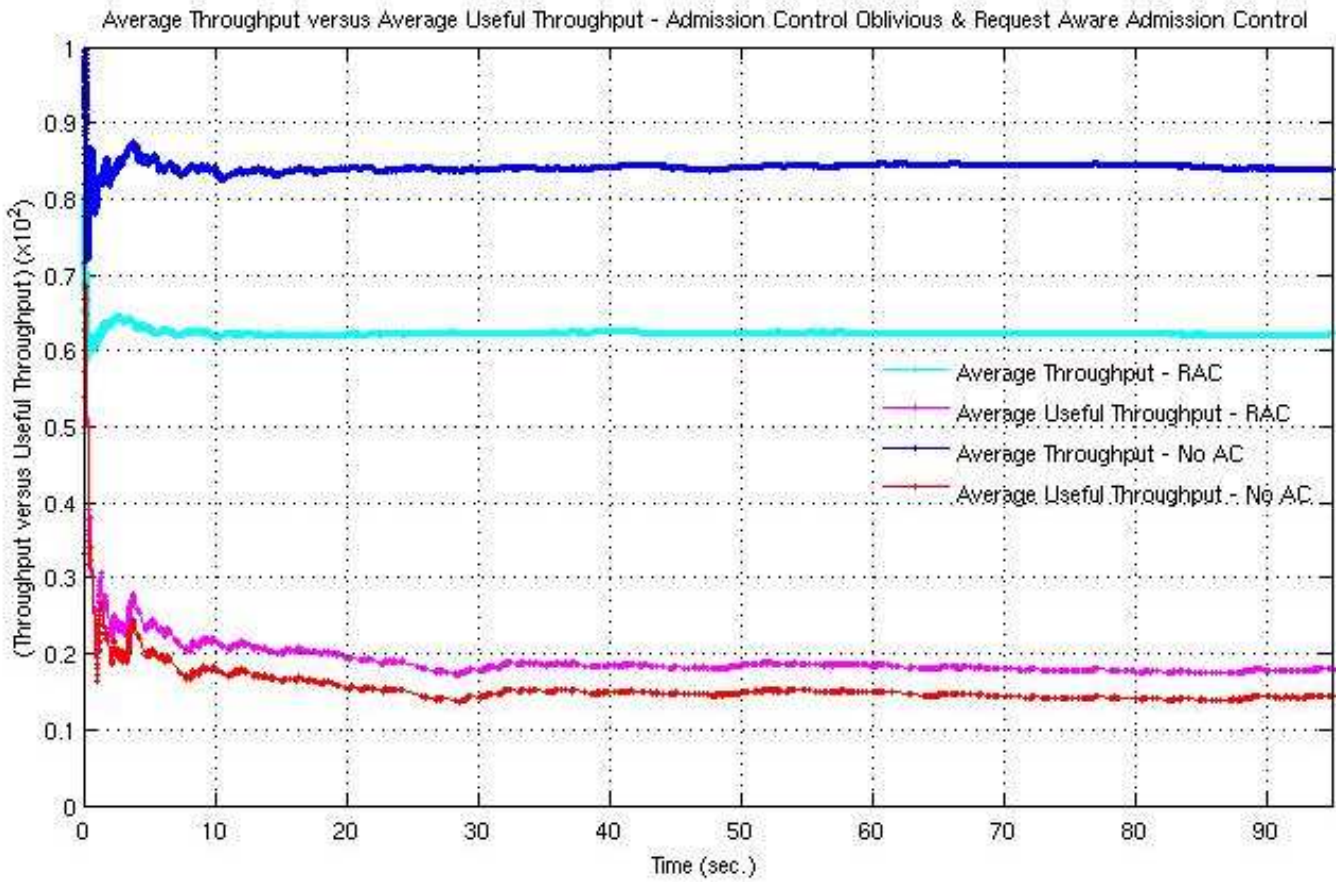
We have simulated the behaviour of a processing server over a duration of 100 sec.

Some simulation parameters

- ✓ Mean packet size (512 Bytes)
- ✓ Small inter-request think time
- ✓ No retransmission
- ✓ Traffic rate of 50 pkts per unit of time.
- ✓ First threshold equals to 75% of the server capacity
- ✓ Second threshold equals to 85% of the server capacity
- ✓ Exponential average talking time
  - ▪ Mixed traffic involves the generation of equal % of medium & long term sessions over the simulation time

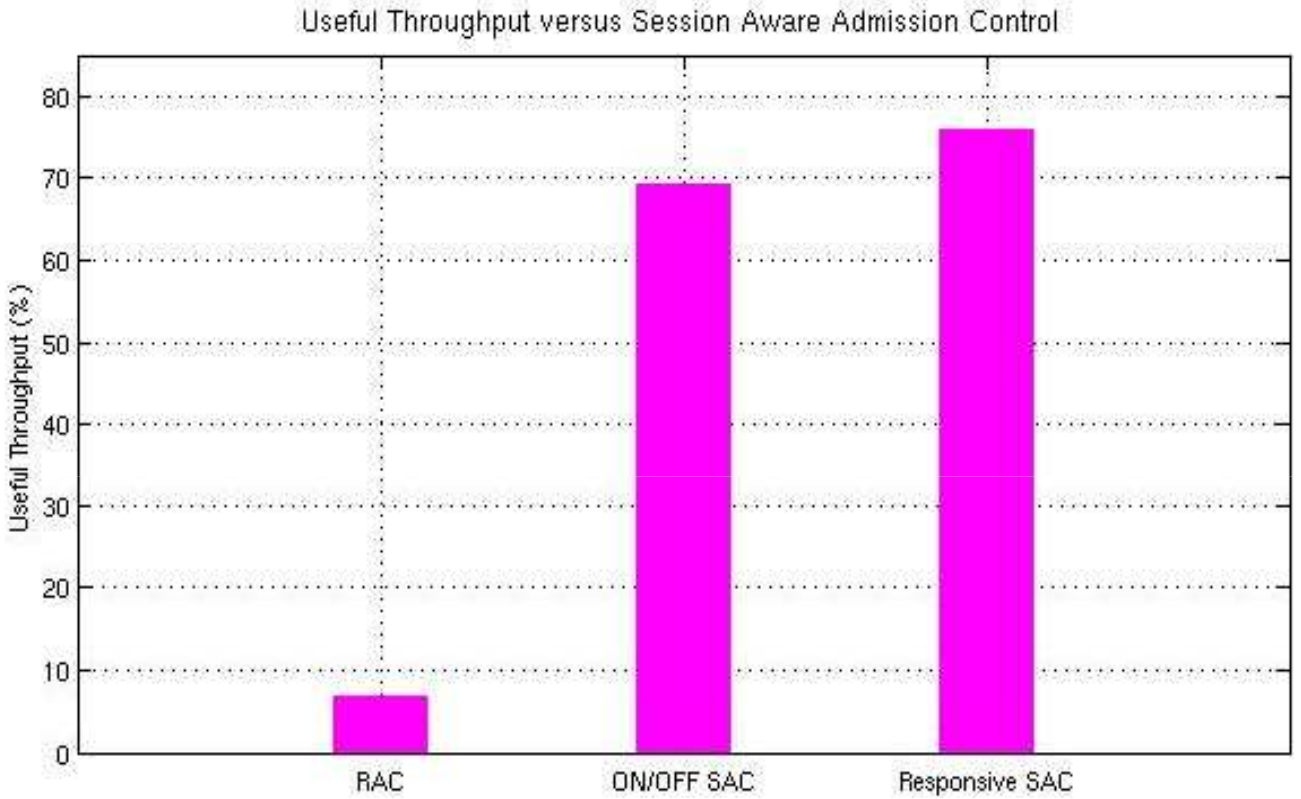
Success rate = (Average number of completed sessions) / (Total number of generated sessions)

Effect of Session-Unaware Admission Control on the Average Useful Throughput - Medium-Term Sessions & Single-Threshold-based RAC (75 % of the Server's Capacity)



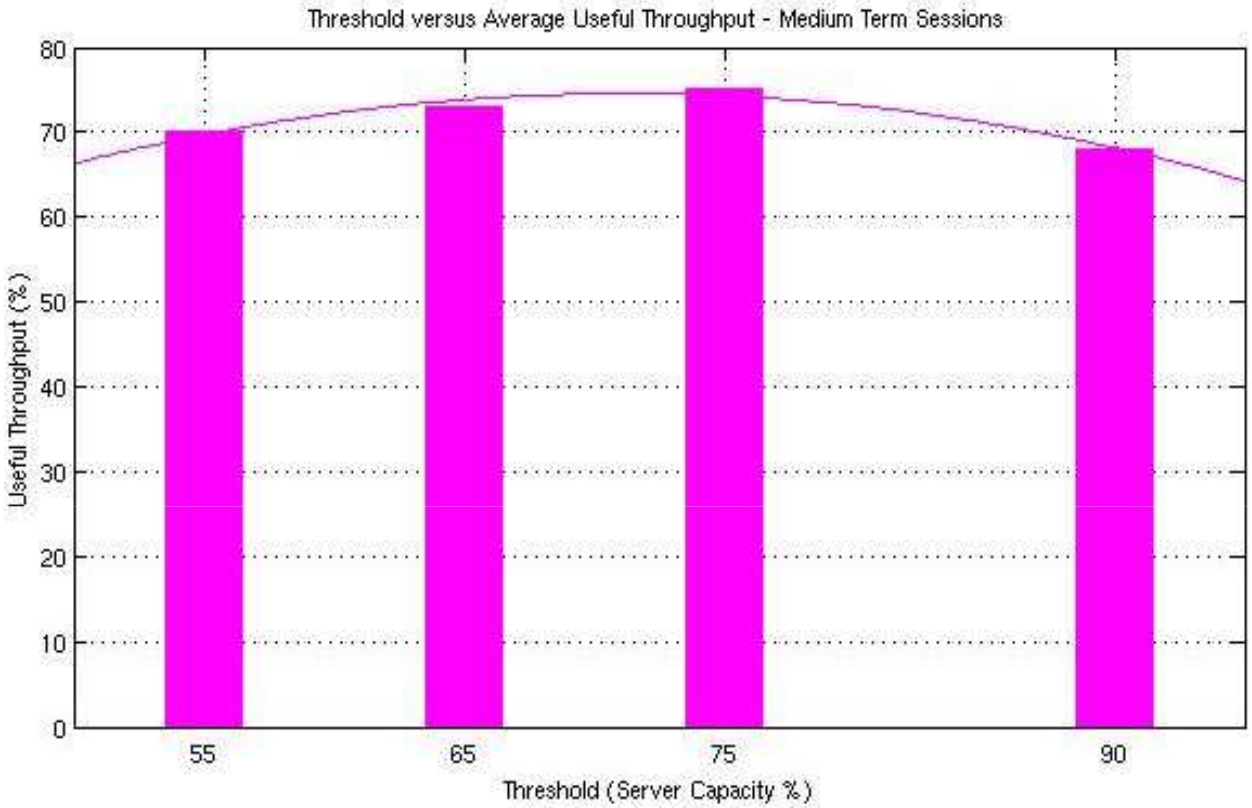
Seems a good / profitable service and condition for provider ! ☺
Not the case for customers ! ☹

Effect of Session-Aware Admission Control on the Average Useful Throughput - Medium-Term Sessions & Single-Threshold-based RAC (75 % of the Server's Capacity)



Session aware admission control is more profitable than Request Aware Admission Control

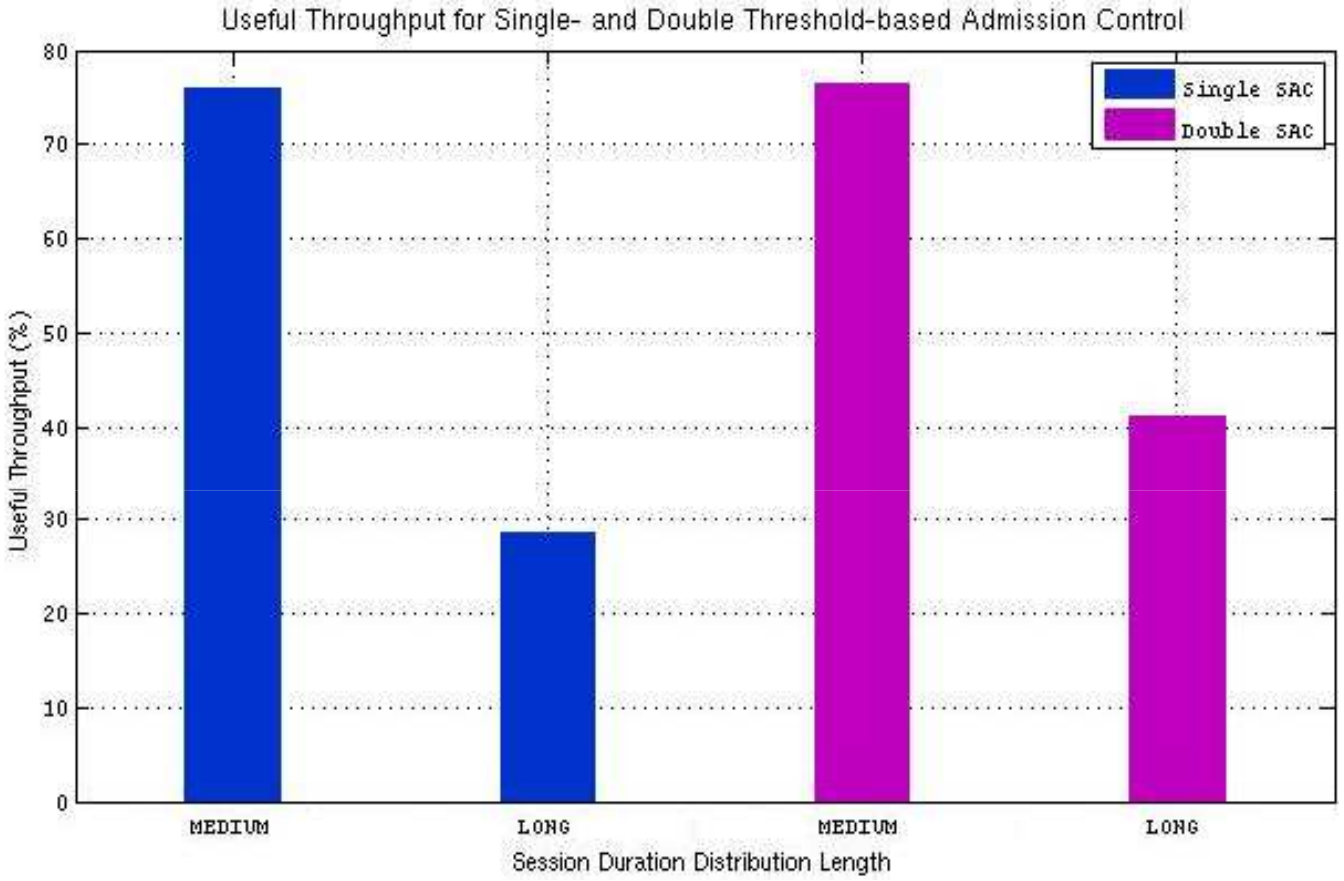
Average Useful Throughput versus Threshold - Medium-Term Sessions & Responsive Session-Aware Admission Control



Why 75 % ?

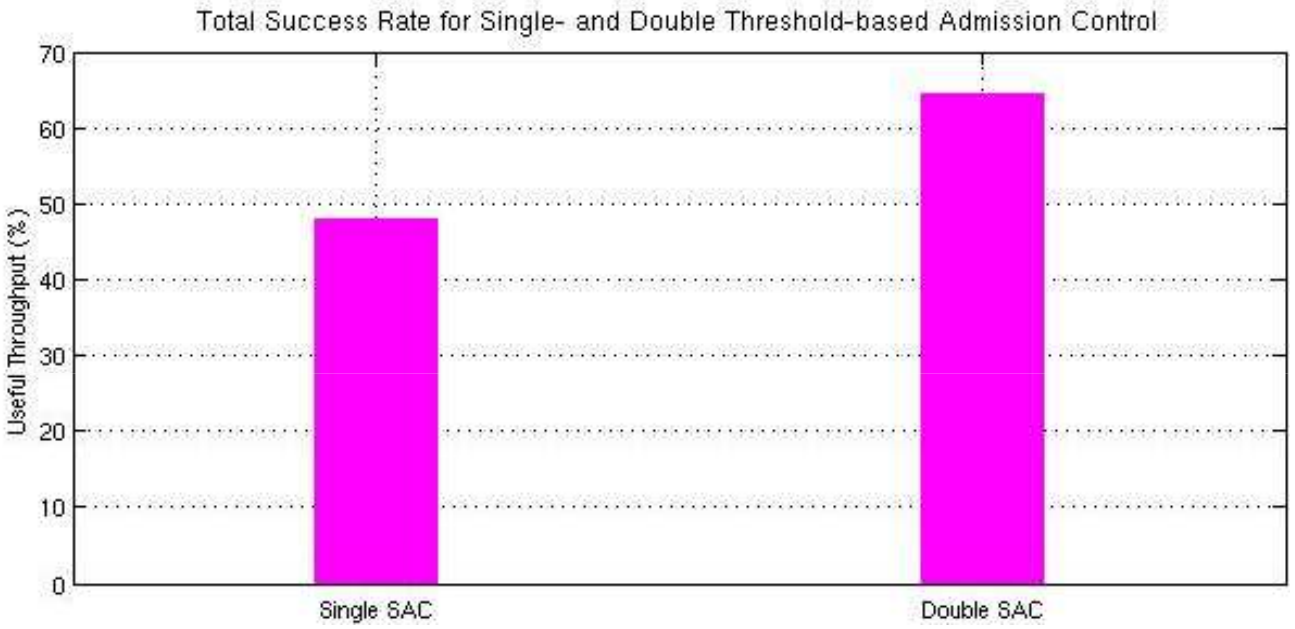
Must be adapted dynamically ! (depending of services, usage, sessions profiles…)

Percentage Histogram of Completed Calls for Different-Admission Control Strategies – Homogeneous Generated Traffic.



Not good to have long sessions in a Single SAC world…

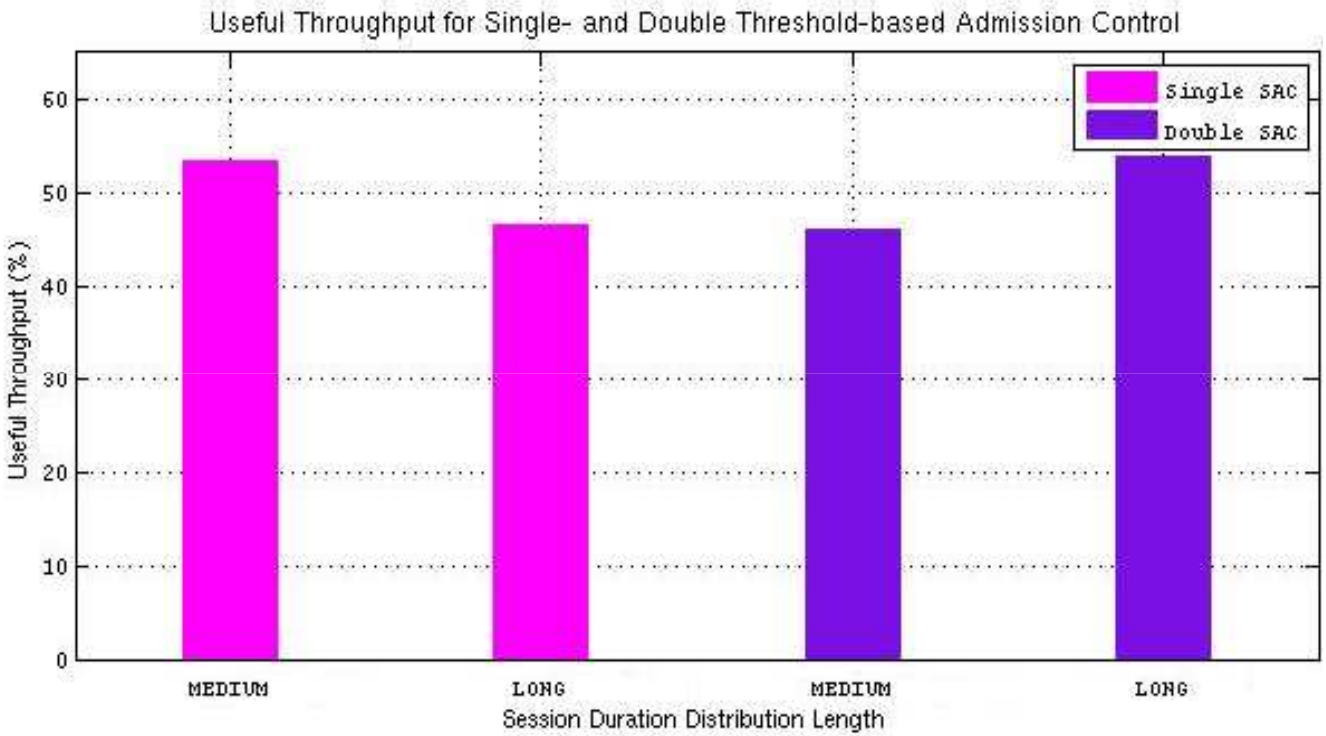
Percentage Histogram of Completed Calls for Different Admission-Control Strategies - Mixed Traffic.



Impact of long sessions in mixed traffic and benefit of double SAC !

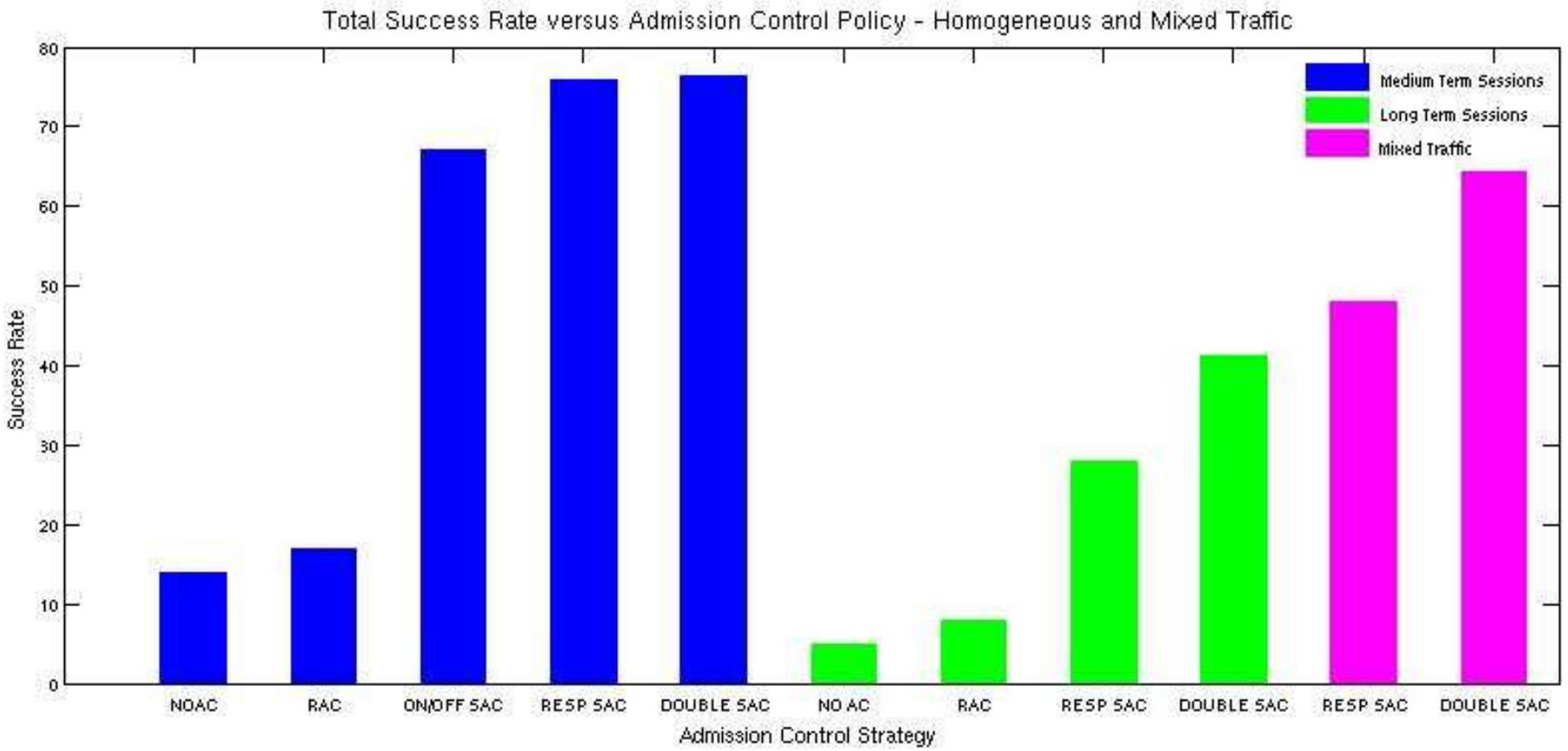
Percentage Histogram of Completed Calls for Different Admission-Control Strategies - Mixed Generated Traffic.

Discrimination Against Long Term Sessions?



Like with homogeneous traffic, impact of single approach to long sessions

Summary: Percentage Histogram of Completed Calls for Different Admission Control Strategies.



Always better to have AC than nothing

Double SAC always provides the best success rate

Only 65 % ? Yes, here the server is 100% used -> to increase this rate, provider must put in place more resources !

Session awareness should be a mandatory approach for service provider to improve customer QoS and satisfactory

Session aware admission control as the means to efficiently prevent a server overload while maximizing the operator profitability.

Responsive session aware admission control is beneficial to increase the performance of a server subjected to long lived sessions

- If we define QoS as the completion of sessions independently of their duration, we can say that double threshold-based session aware admission control improves the QoS provided to subscribers by decreasing discrimination against long lived sessions

- Exploring some approaches for early detection mechanisms (before T1 is reached) and dynamic thresholds
- Extend the advocated session aware admission control model to handle more QoS metrics
  - ✓ Client category, etc.

- First evaluation with homogeneous / mixed traffic
- Enhance the proposed session aware admission control model with the means to address the QoS of highly variable Internet traffic
  - ✓ Use forecasting techniques to improve stability


Thanks ☺!
Any Questions ?
Laurent.lefevre@inria.fr