



Análisis de sensibilidad en Filas de Espera

Rodrigo Maranzana

Repaso: costos en filas de espera

Costo de oportunidad (C_o): costo por no despachar unidades y tenerlas dentro del sistema.

$$\begin{aligned} C_o &= \lambda * W_s * e \\ &= L_s * e \end{aligned}$$

λ : Tasa de arribos.

W_s : Tiempo de espera en el sistema.

L_s : Cantidad de agentes en el sistema.

e : Costo por no despachar (o ganancia obtenida por cada despacho)

Costo operativo (C_E): costo por mantener la infraestructura de filas de espera.

$$C_E = M * C_m$$

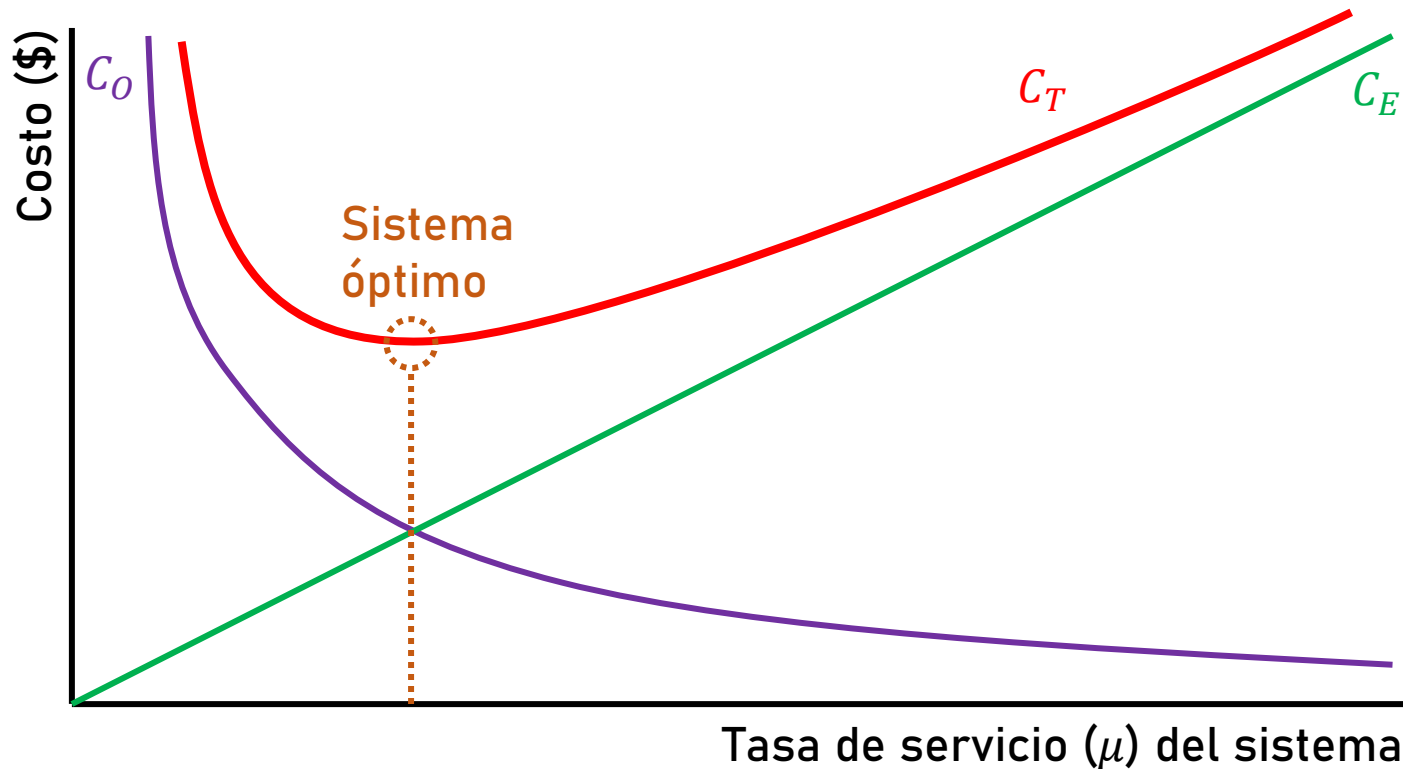
M : Cantidad de servidores.

C_m : Costo de operación de cada servidor.

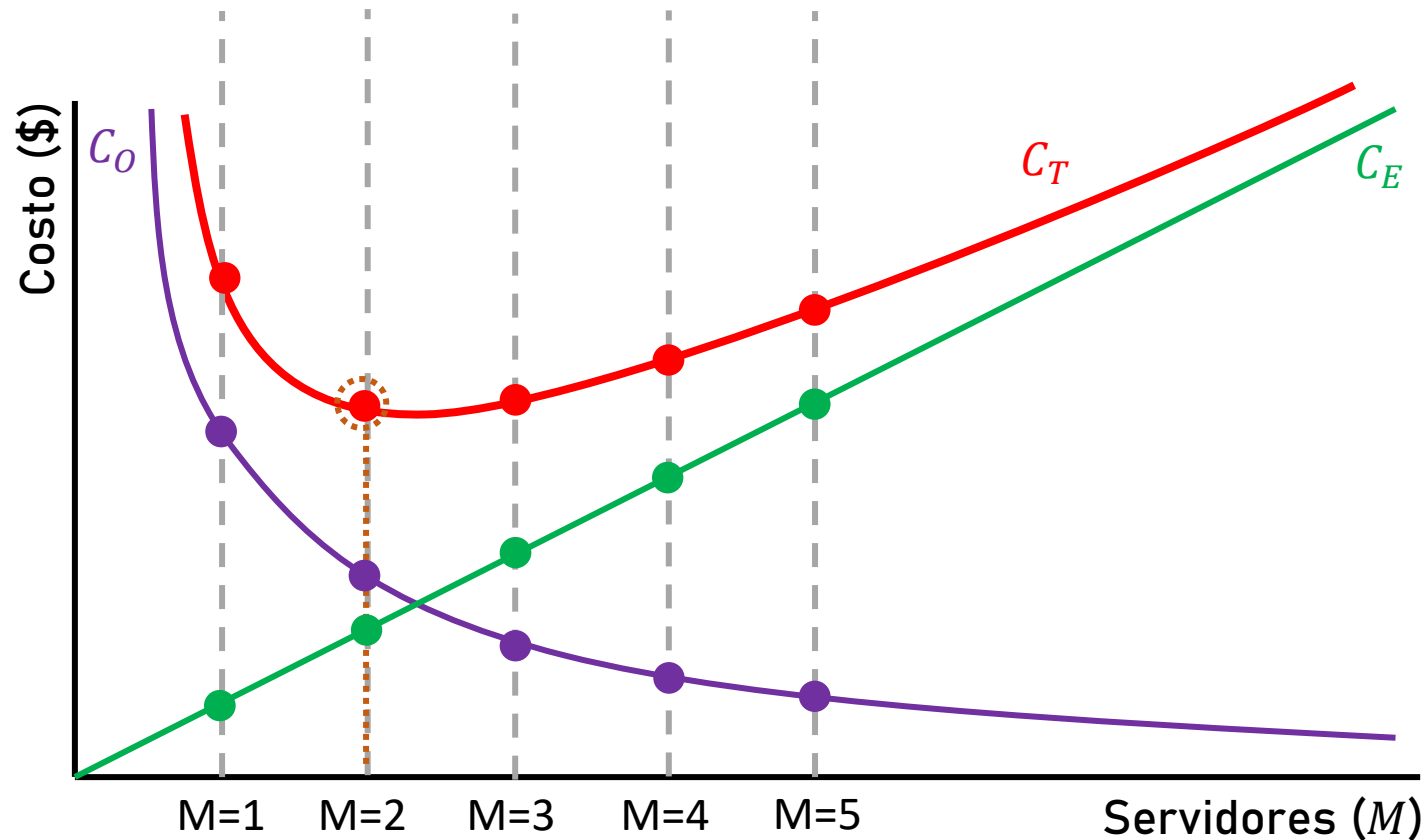
Repaso: costos en filas de espera

Ambos componentes forman parte de una **función de costo total**:

$$C_T = C_O + C_E$$



Repaso: costos en filas de espera



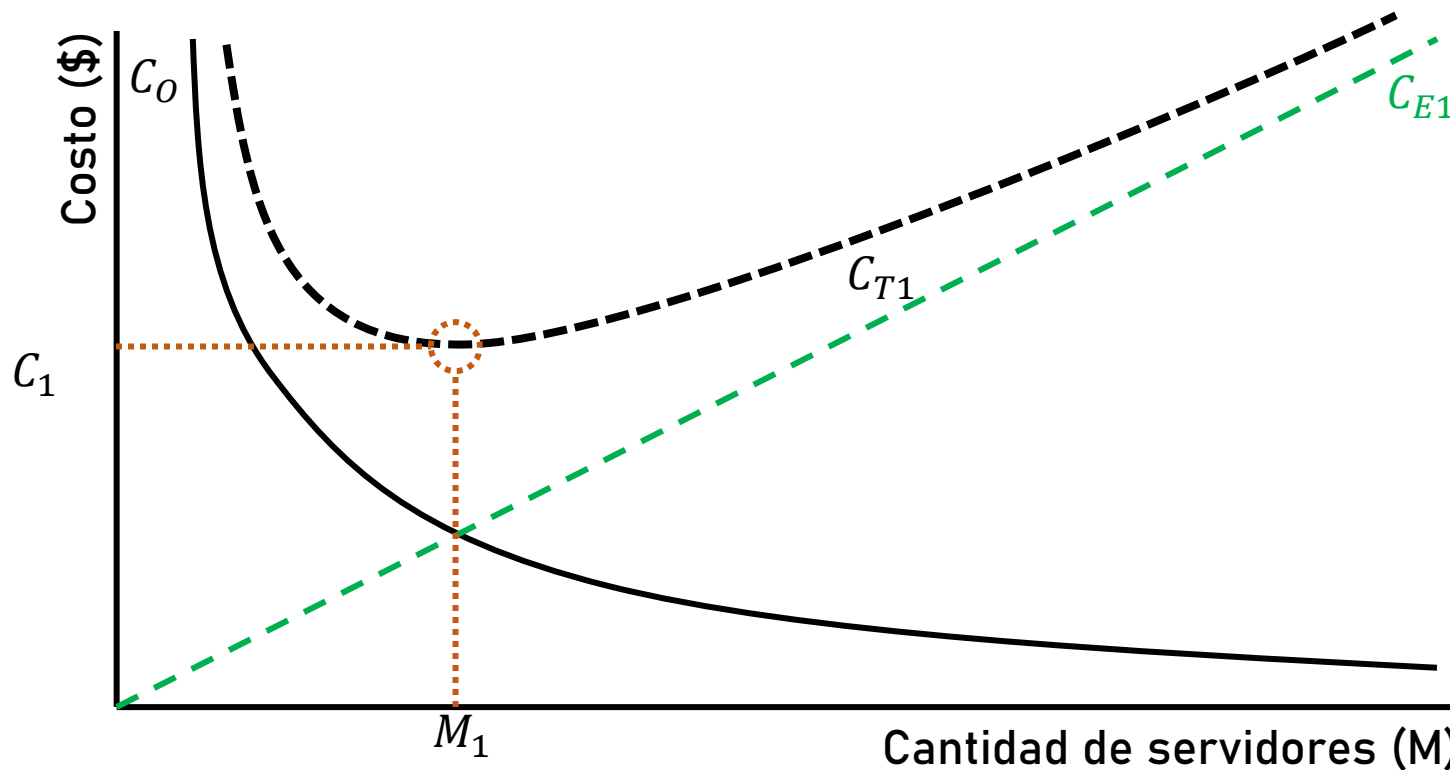
Análisis de sensibilidad de costos

Cambio del **óptimo** si cada una de las **variables** se modifican:

- Cambio de costo operativo C_m .
- Cambio de dinámica de servicio μ .
- Cambio de dinámica de arribos λ .

Cambio de costo operativo : ¿Qué pasa si aumenta?

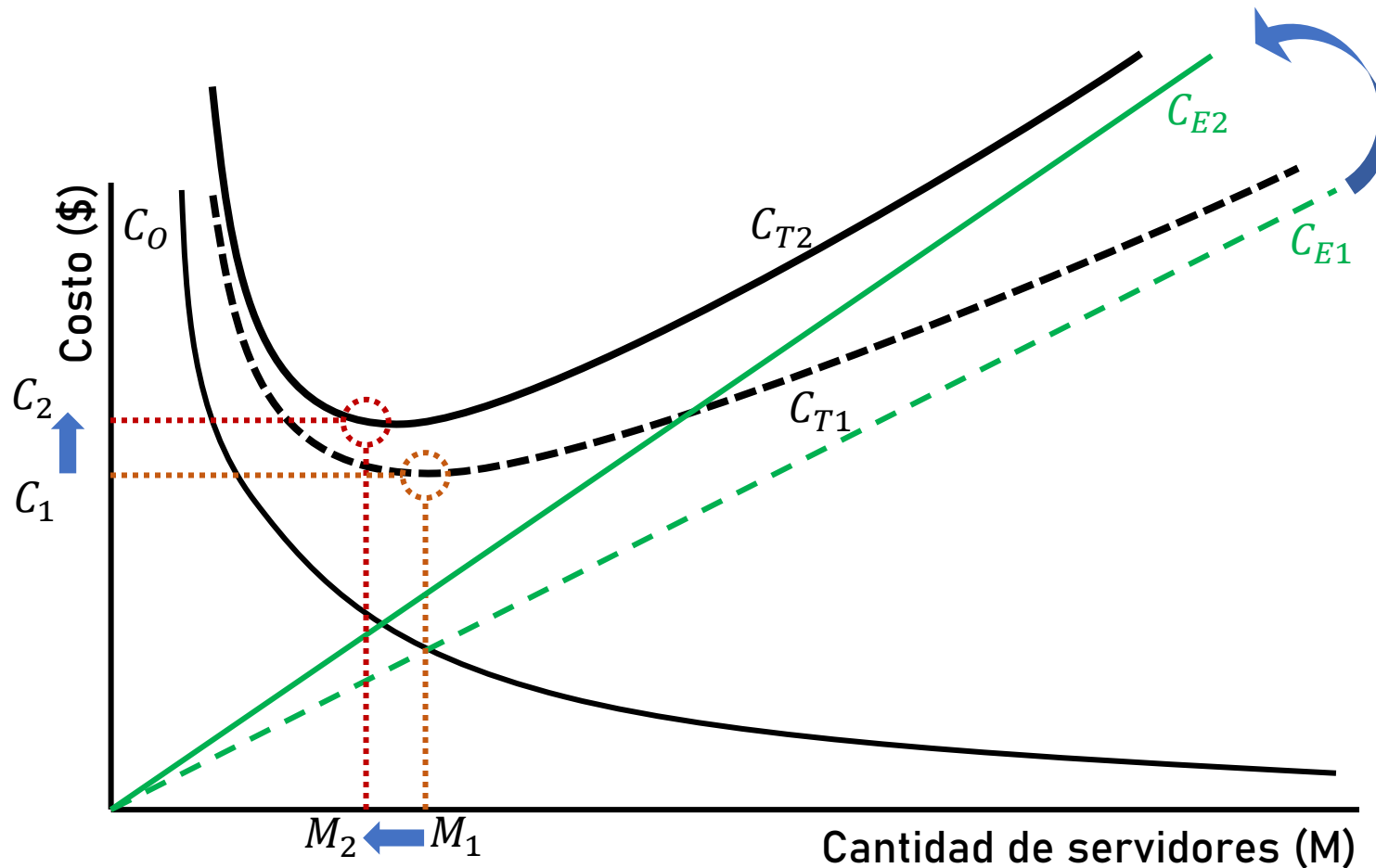
$$C_E = M * C_m$$



Cambio de costo operativo : ¿Qué pasa si aumenta?

El **costo operativo** es la pendiente de la recta C_E .

$$C_E = M * C_m$$



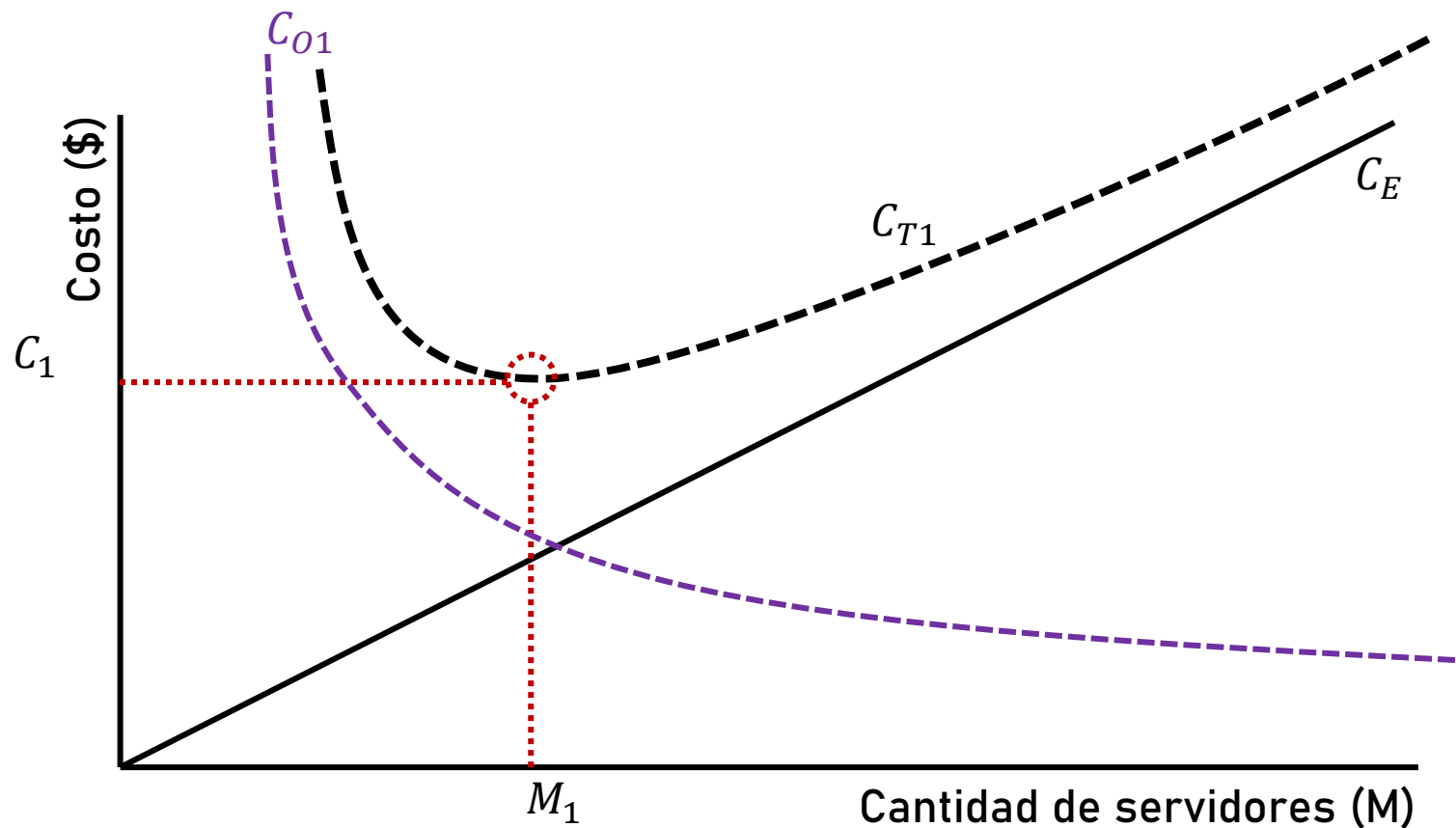
Si aumenta costo operativo.

El óptimo lo encontramos:

- Aumentando el costo total.
- Bajando la cantidad de servidores.

Cambio de tasa de servicio: ¿Qué pasa si aumenta?

$$C_o = \lambda * W_s * e$$



Tasa de servicio en fórmulas M/M/S

$$C_o = \lambda * W_s * e$$

$$W_s = \frac{L_q}{\lambda} + \frac{1}{\mu}$$

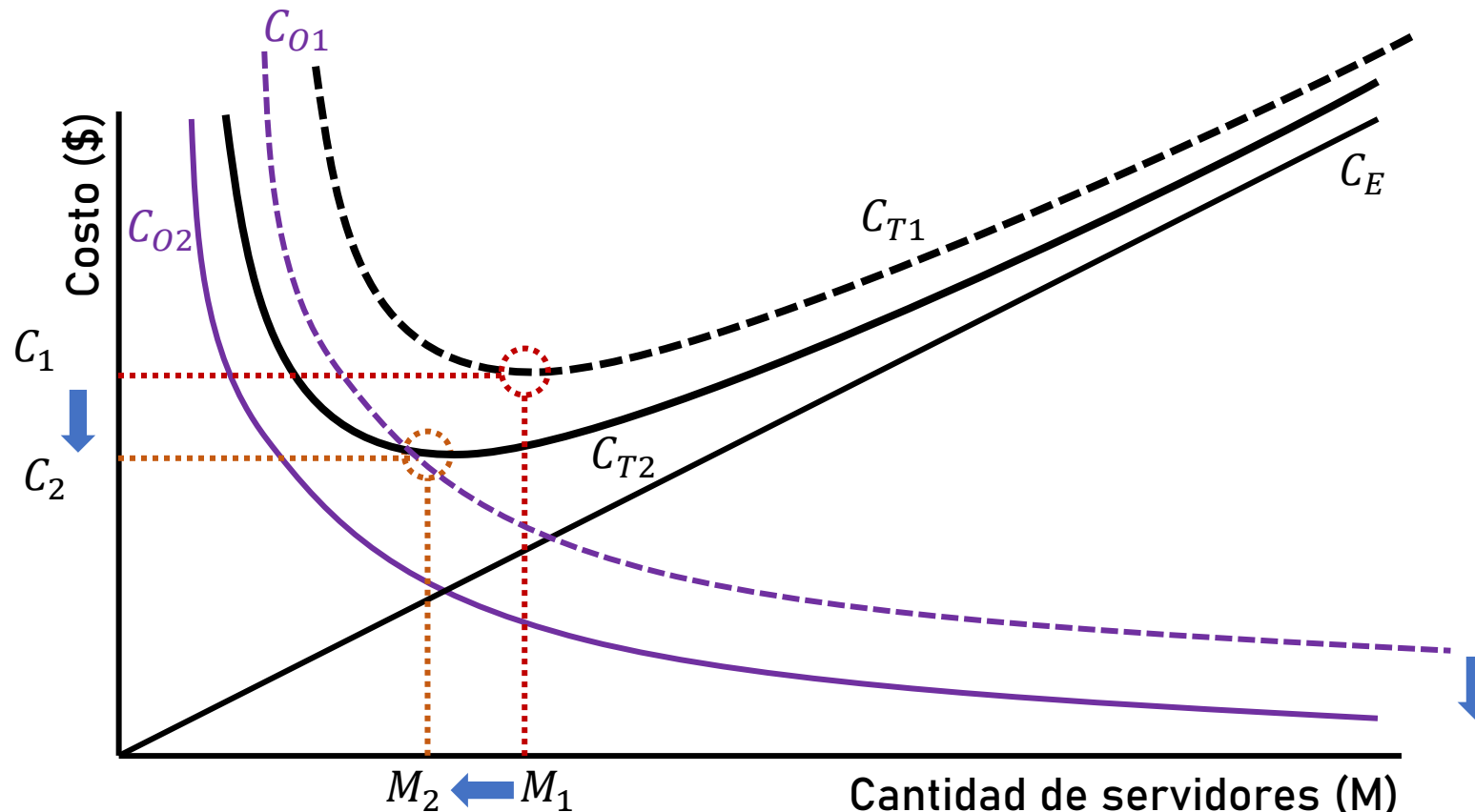
$$\text{Siendo } L_q = \frac{P_0 \left(\frac{\lambda}{\mu} \right)^M \frac{\lambda}{\mu M}}{M! \left(1 - \frac{\lambda}{\mu M} \right)^2}$$

El aumento de la tasa de servicio produce la caída del tiempo de espera en el sistema.

Cambio de tasa de servicio : ¿Qué pasa si aumenta?

La **tasa de servicio** afecta inversamente al costo de oportunidad C_O .

$$C_O = \lambda * W_s * e$$



Si aumenta la tasa de servicio.

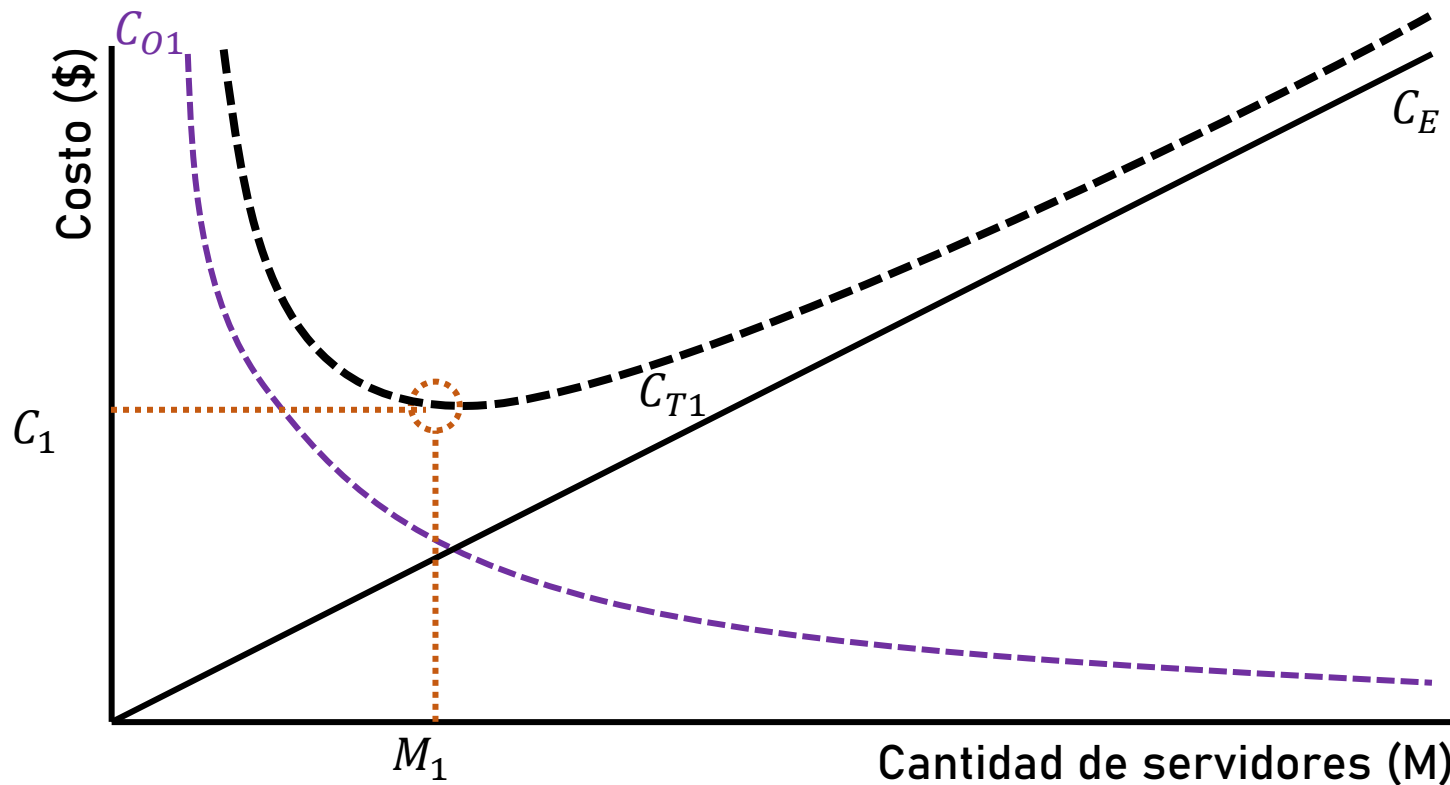
El óptimo lo encontramos:

- Disminuyendo el costo total.
- Disminuyendo la cantidad de servidores.

IMPORTANTE: este caso tiene en cuenta que el costo operativo no se ve afectado por μ . ¡No siempre es así!

Cambio de tasa de servicio : ¿Qué pasa si aumenta?

$$C_O = \lambda * W_s * e$$



Tasa de arribos en fórmulas M/M/S

$$C_o = \lambda * W_s * e$$

$$C_o = L_q + \frac{\lambda}{\mu}$$

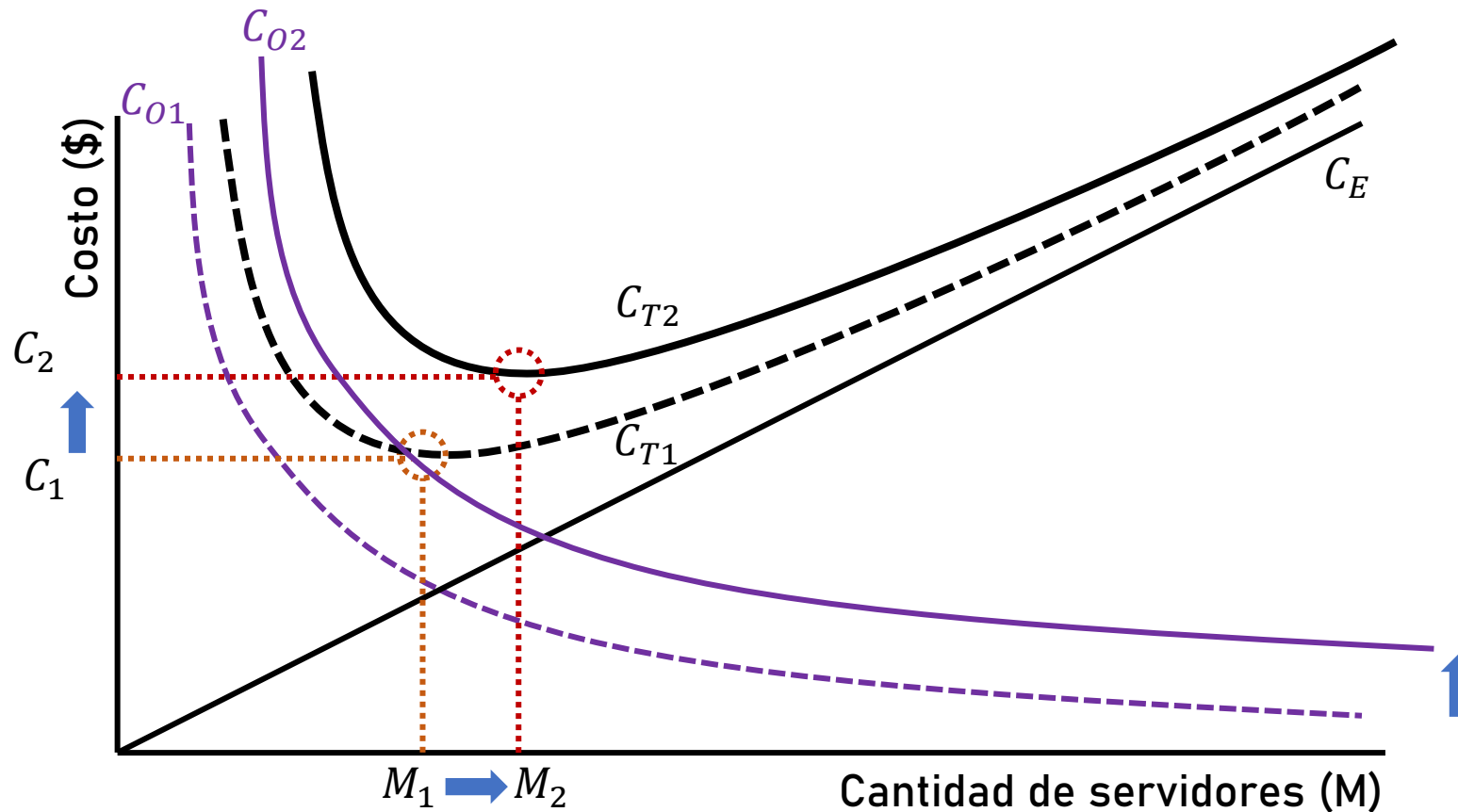
$$\text{Siendo } L_q = \frac{P_0 \left(\frac{\lambda}{\mu} \right)^M \frac{\lambda}{\mu M}}{M! \left(1 - \frac{\lambda}{\mu M} \right)^2}$$

El aumento de la tasa de arribos produce el aumento del tiempo de espera en el sistema.

Cambio de tasa de servicio

La **tasa de arribos** afecta directamente al costo de oportunidad C_E .

$$C_O = \lambda * W_s * e$$



Si aumenta la tasa de arribos.

El óptimo lo encontramos:

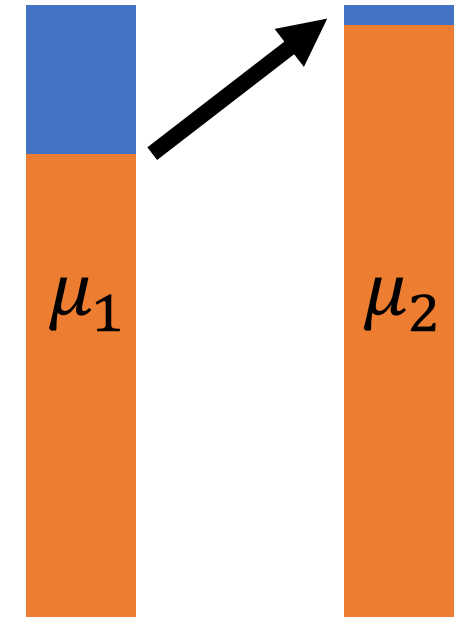
- Aumentando el costo total.
- Aumentando la cantidad de servidores.

Casos particulares: costo operativo piecewise

Las fórmulas de costos deben ser adaptadas al caso de estudio.

Por ejemplo: En el caso de modificación de tasa de servicio μ

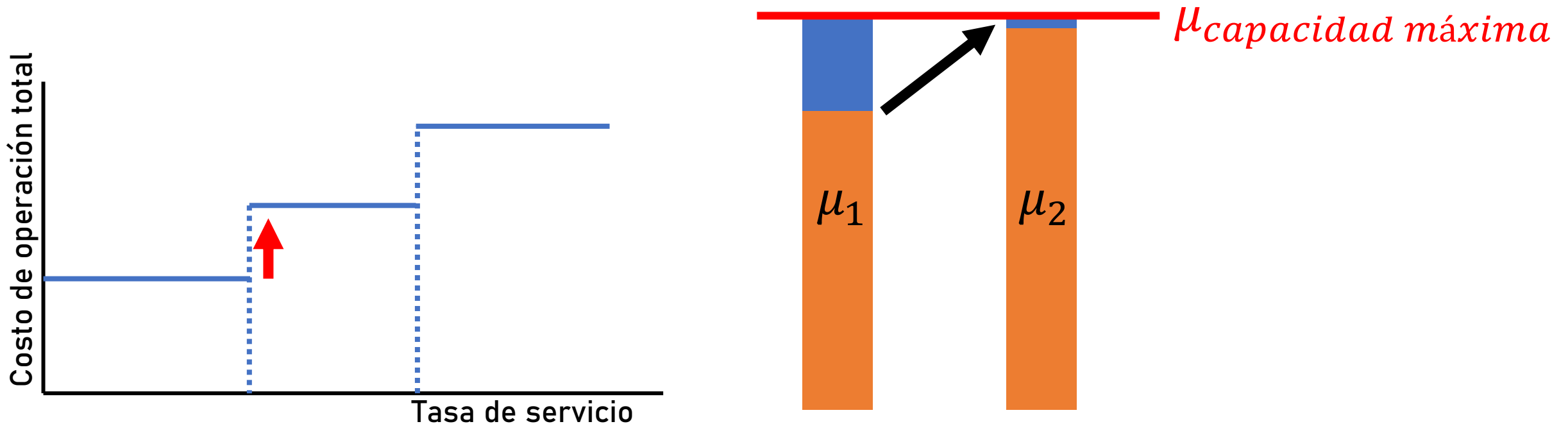
- En una implementación real, el costo operativo se mueve dentro de un margen.
- Significa “aumentar la cadencia del sistema actual”
- ¿Cuanto se puede aumentar?



Casos particulares: costo operativo piecewise

*Se puede aumentar hasta **una capacidad máxima**.*

- *Una vez alcanzada, una mayor tasa se alcanza con una mejora del sistema, e impacta en C_m*



Casos particulares: costo operativo variable

Existen sistemas donde el costo operativo no es solo variable por “M”, sino también por μ .

Casos particulares: costo operativo variable

Existen sistemas donde el costo operativo no es solo variable por “M”, sino también por μ .

Ej: Una web proyecta ventas de vendedores independientes. Cada vez que un vendedor utiliza la plataforma, se envía el cálculo a un servidor que ejecuta un modelo de forecast y regresa la respuesta.

Existe una fila de espera en las peticiones, para mantener controlado el gasto de cómputo en la nube.

Se utiliza *Google Cloud* para el cálculo, particularmente el módulo *Vertex AI Forecast*.

Google Cloud



Casos particulares: costo operativo variable

En este caso el costo es por punto calculado del forecast.

Por lo tanto, la función de C_E debe afectarse por μ .

Dentro de cada rango de “puntos”, la función es lineal.

Vertex AI Forecast	
AutoML	ARIMA+
Precios	
Etapa	
Predicción	<ul style="list-style-type: none">\$0.2 por 1,000 puntos de datos* (de 0 a 1,000,000 de puntos)\$0.1 por 1,000 puntos de datos* (de 1 millón a 50 millones de puntos)\$0.02 por 1,000 puntos de datos* (50 millones de puntos)

Fuente: <https://cloud.google.com/vertex-ai/pricing?hl=es-419#automl>