Trabajo Fin de Grado
Ingeniería de Tecnologías Industriales

Modelo de Simulación para la Operación Óptima de Centros de Procesos de Datos

Autor: Celia Regidor Pérez

Tutor: Daniel Limón Marruedo

Dpto. Ingeniería de Sistemas y Automática
Escuela Técnica Superior de Ingeniería
Sevilla, 2019

Trabajo Fin de Grado
Ingeniería de Tecnologías Industriales

Modelo de Simulación para la Operación Óptima de Centros de Procesos de Datos

Autor:

Celia Regidor Pérez

Tutor:

Daniel Limón Marruedo

Catedrático

Dpto. Ingeniería de Sistemas y Automática

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2019

Proyecto Fin de Carrera:

Modelo de Simulación para la Operación Óptima de Centros de Procesos de Datos

Autor: Celia Regidor Pérez

Tutor: Daniel Limón Marruedo

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2019

El Secretario del Tribunal

A mi familia

A mis profesores

Agradecimientos

Quiero aprovechar este espacio para agradecer a mis padres la educación, los valores, el cariño y el apoyo incondicional que me han transmitido siempre y que me han convertido en la persona que soy.

Mi agradecimiento lo hago extensivo a toda mi familia por todo el cariño que me ofrecen en todo momento y a mis amigos por acompañarme en este camino.

Este trabajo no habría sido posible sin la ayuda, el ánimo y la dirección de mi tutor, Daniel Limón, al que expreso mi admiración y gratitud.

¡Gracias a todos de corazón!

Celia Regidor Pérez

Sevilla, 2019

Resumen

Los Centros de datos (CPD) o Data Centres son infraestructuras tecnológicas que dan servicios de procesamiento y almacenamiento de datos a gran escala basados en Internet.

El crecimiento exponencial de los servicios en red, de los sistemas Cloud Computing y más recientemente del IoT, está haciendo crecer estas infraestructuras hasta escalas extremas.

Estos centros son grandes consumidores de energía, tanto para satisfacer las necesidades de operación de los equipos, como para los sistemas de refrigeración necesarios para evitar el sobrecalentamiento de los mismos.

Los modelos de consumo de energía son fundamentales en el diseño y la optimización de las operaciones de eficiencia energética para frenar el consumo excesivo de energía en los centros de datos.

Existe toda una diversidad de modelos utilizados para predecir el consumo de energía para los centros de datos y sus componentes, de forma que se pueda gestionar la eficiencia en las operaciones de los mismos.

Casi todos los modelos existentes abordan la simulación desde el punto de vista de la potencia consumida por los elementos hardware y software en servicio.

En este documento abordaremos el diseño de un modelo en el que incluiremos la variable de la temperatura de los elementos, a fin de analizar posibles estrategias de optimización del consumo energético global en estas instalaciones y de su funcionamiento.

Abstract

Data Centers are technological infrastructures that provide large-scale data processing and storage services based on the Internet.

The exponential growth of network services, Cloud Computing systems and more recently IoT, are growing these infrastructures to extreme scales.

These centers are large consumers of energy, both to meet the operational needs of the equipment, and for the cooling systems necessary to avoid overheating.

Energy consumption models are fundamental in the design and optimization of energy efficiency operations to curb excessive energy consumption in Data Centers.

There is a variety of models used to predict energy consumption for data centers and their components, so that they can manage efficiency in their operations.

Almost all existing models address the simulation from the point of view of the power consumption of hardware and software elements in service.

In this document we propose the design of a model in which we will include the variable of the temperature of the elements, in order to analyze possible strategies for optimizing global energy consumption in these facilities.

Índice

1 INTRODUCCIÓN Y OBJETO DEL TRABAJO

Como ya hemos señalado, los modelos de consumo de energía son fundamentales en el diseño y la optimización de las operaciones de eficiencia energética para frenar el consumo excesivo de energía en los centros de datos.

Existen diversos modelos que se utilizan para predecir el consumo de energía en los centros de datos, la mayoría de los cuales abordan la simulación desde el punto de vista de la potencia eléctrica consumida por los elementos hardware y software en servicio.

El objetivo de este documento es presentar un modelo de simulación en el que se incluya la variable de la temperatura de los elementos, de forma que sea posible analizar distintas estrategias de optimización del consumo energético global en estas instalaciones y de otros aspectos de su funcionamiento.

Para la realización de este modelo nos hemos basado en las últimas tendencias y estudios de la eficiencia energética en centros de datos, en los que hemos basado el modelo teórico.

Para el desarrollo de la simulación hemos utilizado la herramienta informática Arena® Simulation, de la que expondremos algunas nociones de manejo básicas.

Como resultado de este trabajo, presentamos la implementación del modelo en dicha herramienta Arena® Simulation junto con algunos resultados preliminares.

2 CENTRO DE DATOS

¿Qué es un centro de proceso de datos?

Los centros de proceso de datos o centros de datos, son instalaciones centralizadas donde los equipos de procesamiento, almacenamiento y redes se concentran con el propósito de recopilar, almacenar, procesar, distribuir o permitir el acceso a grandes cantidades de datos.

Desde la llegada de los ordenadores, los centros de datos siempre han estado presentes de una u otra forma.

En los comienzos de la informática, los ordenadores tenían el tamaño de una habitación y los centros de datos eran los lugares que las albergaban.

A medida que los equipos se fueron haciendo más pequeños y baratos y las necesidades de procesamiento de datos comenzaron a aumentar exponencialmente, comenzamos a conectar en red múltiples servidores para aumentar la potencia de procesamiento. Los conectamos a redes de comunicación para facilitar el acceso a la información de forma remota. Un gran número de estos servidores agrupados y equipos relacionados se pueden alojar en una habitación, un edificio completo o grupos de edificios.

Los centros de datos actuales albergan miles de servidores muy potentes y muy pequeños que se ejecutan 24 horas al día durante 365 días al año.

Debido a sus altas concentraciones de servidores, a menudo apilados en bastidores que se colocan en filas, los centros de datos a veces se denominan granjas de servidores. Proporcionan servicios importantes como almacenamiento de datos, respaldo y recuperación, gestión de datos y redes. Estos centros pueden almacenar y servir sitios web, ejecutar servicios de correo electrónico y mensajería instantánea (IM), proporcionar almacenamiento y aplicaciones en la nube, permitir transacciones de comercio electrónico, potenciar las comunidades de juegos en línea, etc.

Casi todas las empresas y entidades gubernamentales necesitan su propio centro de datos o necesitan acceso al de otra persona. Algunos los construyen y mantienen internamente, algunos alquilan servidores en instalaciones de uso compartido (también llamados *colos*) y otros usan servicios públicos basados en la nube en hosts como Amazon, Microsoft, Sony o Google.

Los *colos* y los otros grandes centros de datos comenzaron a surgir a fines de la década de 1990 y principios de la década de 2000, en algún momento después de que el uso de Internet se generalizó. Los centros de datos de algunas grandes empresas están espaciados en todo el planeta para satisfacer la necesidad constante de acceso a grandes cantidades de información. Según los informes, actualmente hay más de 3 millones de centros de datos de diversas formas y tamaños en el mundo.

¿Por qué necesitamos centros de datos?

A pesar de que el hardware se está volviendo cada vez más pequeño, más rápido y más potente, somos una especie cada vez más hambrienta de datos, y la demanda de poder de procesamiento, espacio de almacenamiento e información en general está creciendo y constantemente amenazando con superar las capacidades que las empresas pueden asumir.

Cualquier entidad que genere o use datos necesita centros de datos en algún nivel, incluidas agencias gubernamentales, organismos educativos, compañías de telecomunicaciones, instituciones financieras, minoristas de todos los tamaños y proveedores de información en línea y servicios de redes sociales como Google y Facebook. La falta de acceso rápido y confiable a los datos puede significar la incapacidad de proporcionar servicios vitales o la pérdida de la satisfacción del cliente y los ingresos.

Todos estos medios deben almacenarse en algún lugar. Y hoy en día, cada vez más cosas también se están moviendo hacia la nube, lo que significa que, en lugar de ejecutarlas o almacenarlas en nuestras propias computadoras de casa o del trabajo, estamos accediendo a ellas a través de los servidores host de los proveedores de la nube. Muchas compañías también están trasladando sus aplicaciones profesionales a servicios en la nube para reducir el costo de ejecutar sus propias redes y servidores informáticos centralizados.

La nube no significa que las aplicaciones y los datos no se alojen en el hardware informático. Simplemente significa que alguien más mantiene el hardware y el software en ubicaciones remotas donde sus clientes pueden acceder a ellos a través de Internet. Esas ubicaciones son centros de datos. [1]

Escala y diseño de un centro de datos

Loscentros de datos existen en todas las formas, tamaños y configuraciones. Van desde unos pocos servidores en una habitación hasta enormes estructuras independientes que miden decenas de miles de metros cuadrados con decenas de miles de servidores y otro hardware que lo acompaña. Sus tamaños y los tipos de equipos que contienen varían según las necesidades de la entidad o entidades a las que dan soporte.

Hay varios tipos, incluidos proveedores de nube privada como los *colos*, proveedores de nube pública como Amazon y Google, centros de datos privados de empresas y centros de datos gubernamentales como los de la NSA o instalaciones de investigación científica.

No cuentan con personal como oficinas con una persona por computadora, sino con un número menor de personas que monitorean una gran cantidad de computadoras y dispositivos de red, así como energía, refrigeración y otras instalaciones necesarias del edificio. Algunos son tan grandes que los empleados se desplazan en scooters o bicicletas. Los pisos tienen que soportar más peso que un edificio de oficinas típico porque el equipo puede ser pesado. También deben tener techos altos para acomodar cosas como bastidores altos, pisos elevados y cableado suspendido en el techo, entre otras cosas.

Google tiene trece grandes centros de datos, incluidas ubicaciones en el condado de Douglas, Ga .; Lenoir, N.C .; Condado de Berkeley, S.C .; Council Bluffs, Iowa; Condado de Mayes, Okla .; The Dalles, Ore .; Quilicura, Chile; Hamina, Finlandia; St. Ghislain, Bélgica; Dublín, Irlanda; Hong Kong, Singapur y Taiwán; así como muchos mini centros de datos, algunos incluso en sitios de ubicación conjunta. El gigante tecnológico también es propenso a experimentar con el diseño. Por ejemplo, alrededor de 2005, Google utilizó contenedores de envío que contenían equipos de servidor en sus centros de datos, y desde entonces se ha pasado a otros diseños personalizados.

La configuración de los servidores, la topología de la red y el equipo de soporte pueden variar mucho según la empresa, el propósito, la ubicación, la tasa de crecimiento y el concepto de diseño inicial del centro de datos. Su diseño puede afectar en gran medida la eficiencia del flujo de datos y las condiciones ambientales dentro del centro. Algunos sitios pueden dividir sus servidores en grupos por funciones, como la separación de servidores web, servidores de aplicaciones y servidores de bases de datos, y algunos pueden hacer que cada uno de sus servidores realice múltiples tareas. No hay reglas estrictas y rápidas, y no hay muchos estándares oficiales.

Algunas organizaciones están tratando de crear estándares. La Asociación de la Industria de Telecomunicaciones desarrolló un estándar de clasificación de nivel de centro de datos en 2005 llamado proyecto TIA-942, que identificó cuatro categorías de centro de datos, clasificadas por métricas como redundancia y nivel de tolerancia a fallas.

Este estándar clasifica cuatro niveles:

- Nivel 1: infraestructura básica del sitio con una única ruta de distribución que no tiene redundancia incorporada.
- Nivel 2: infraestructura de sitio redundante con una única ruta de distribución que incluye componentes redundantes.
- Nivel 3: infraestructura de sitio que se puede mantener de manera concurrente que tiene múltiples rutas, solo una de las cuales está activa a la vez.
- Nivel 4: infraestructura de sitio tolerante a fallas que tiene múltiples rutas de distribución activas para mucha redundancia.

En teoría, los sitios que caen en las categorías de nivel 1 y 2 deben cerrarse ocasionalmente por mantenimiento, mientras que los sitios de nivel 3 y 4 deben poder mantenerse activos durante el mantenimiento y otras interrupciones. Cuanto mayor es el nivel, mayor es el nivel de fiabilidad (lo que significa menos tiempo de inactividad potencial) como también mayor es el coste de operación.

No todos los centros de datos siguen estos estándares. Y los centros de datos de hoy son un fenómeno tan nuevo que no hay códigos de construcción específicos para ellos en la mayoría de las áreas en este momento. Generalmente se agrupan en algún otro tipo genérico.

Sus diseños, equipos y necesidades evolucionan continuamente, pero hay algunos elementos comunes que encontrará en muchos centros de datos.

El hardware

Aunque los diseños físicos varían, cada centro de datos alberga agrupaciones de servidores (clusters)

Una característica física común de los centros de datos son los grupos de servidores interconectados. Todos pueden ser muy similares, apilados de forma ordenada en bastidores abiertos o armarios cerrados de igual altura, ancho y profundidad, o podría haber un montón de diferentes tipos, tamaños y edades de máquinas que coexisten, como pequeños servidores modernos planos junto a viejos y voluminosos. Cajas Unix y mainframes gigantes (una raza que desaparece rápidamente, pero que aún no ha desaparecido por completo).

Cada servidor es una computadora de alto rendimiento, con memoria, espacio de

almacenamiento, un procesador o procesadores y capacidad de entrada / salida, algo así como una versión mejorada de una computadora personal, pero con un procesador más rápido y potente y mucha más memoria y, por lo general, sin monitor, teclado u otros periféricos que usaría en casa. Los monitores pueden existir en una ubicación centralizada, cercana o en una sala de control separada, para monitorear grupos de servidores y equipos relacionados.

Un servidor o servidores en particular pueden estar dedicados a una sola tarea o ejecutar muchas aplicaciones diferentes. Algunos servidores en centros de datos de ubicación conjunta están dedicados a clientes particulares. Algunos incluso son virtuales en lugar de físicos (una nueva tendencia que reduce la cantidad necesaria de servidores físicos). También es probable, cuando solicita algo a través de Internet, que varios servidores estén trabajando juntos para resolver la solicitud de servicio.

Redes, software y control

Los equipos de redes y comunicación son absolutamente necesarios en un centro de datos para mantener una red de gran ancho de banda para la comunicación con el mundo exterior, y entre los servidores y otros equipos dentro del centro de datos. Esto incluye componentes como routers, switches, controladores de interfaz de red (NIC) de los servidores y, potencialmente, kilómetros y kilómetros de cableado. El cableado viene en varias formas, incluyendo par trenzado (cobre), coaxial (también cobre) y fibra óptica (vidrio o plástico). Los tipos de cable y sus diversos subtipos afectarán la velocidad a la que fluye la información a través del centro de datos.

Todo ese cableado también debe estar organizado. Se despliega sobre la zona alta de las salas en bandejas colgadas del techo o unidas a la parte superior de los bastidores, o bien debajo de un suelo técnico elevado. La codificación de color y el etiquetado meticuloso se utilizan para identificar las diversas líneas de cableado. Los suelos elevados de los centros de datos generalmente tienen paneles o mosaicos que se pueden levantar para acceder al cableado y otros equipos. Las unidades de refrigeración y los equipos de energía a veces también se encuentran debajo del suelo.

Otros equipos importantes del centro de datos incluyen dispositivos de almacenamiento (como unidades de disco duro, unidades de estado sólido y unidades de cinta robótica), fuentes de alimentación ininterrumpida (UPS), baterías de respaldo, generadores de respaldo y otros equipos relacionados con la energía.

Los centros de datos también tienen muchos equipos para controlar la temperatura y el control de calidad del aire, aunque los métodos y tipos de equipos varían de un sitio a otro. Pueden incluir ventiladores, controladores de aire, filtros, sensores, aires acondicionados de sala de computadoras (CRAC), enfriadores, tuberías de agua y tanques de agua. Algunos sitios también utilizan barreras de plástico o metal o utilizan elementos como gabinetes de servidores de chimenea para controlar el flujo de aire caliente y frío para evitar el sobrecalentamiento de los equipos informáticos.

Y, por supuesto, se necesita software para ejecutar todo este hardware, incluidos los diversos sistemas operativos y aplicaciones que se ejecutan en los servidores, software de agrupación en clúster como MapReduce o Hadoop de Google para permitir que el trabajo se distribuya en cientos o más máquinas, programas de conexión a Internet para controle las redes, las aplicaciones de monitoreo del sistema y el software de virtualización como VMware para ayudar a reducir la cantidad de servidores físicos.

Algunos problemas a los que se enfrentan los centros de datos

Los centros de datos se esfuerzan por proporcionar un servicio rápido e ininterrumpido. Los fallos de los equipos, las interrupciones de la comunicación o la alimentación, la congestión de la red y otros problemas que impiden que las personas accedan a sus datos y aplicaciones deben resolverse de inmediato. Debido a la constante demanda de acceso instantáneo, se espera que los centros de datos funcionen 24 horas al día los 365 días del año, lo que crea una gran cantidad de problemas.

Las necesidades de red de un centro de datos son muy diferentes de las de, por ejemplo, un edificio de oficinas lleno de trabajadores. Las redes de centros de datos son potentísimas. Las redes de fibra óptica de Google envían datos mucho más rápido que el servicio de Internet de su hogar.

Casi nadie tiene tanto tráfico como Google, pero todos los centros de datos probablemente verán más y más uso. Necesitan la capacidad de escalar sus redes para aumentar el ancho de banda y mantener la fiabilidad. Lo mismo ocurre con los servidores, que se pueden ampliar para aumentar la capacidad del centro de datos. La red existente necesita poder manejar la congestión controlando el flujo adecuadamente. Y todo lo que está deteniendo el flujo debe ser eliminado. Una red solo será tan rápida como su componente más lento. Los acuerdos de nivel de servicio (SLA) con los clientes también deben cumplirse, y a menudo incluyen cosas como el rendimiento y el tiempo de respuesta.

Hay varios puntos de posible falla. Los servidores o equipos de red pueden apagarse, los cables pueden fallar o los servicios se suministran desde el exterior, como la alimentación y la comunicación, pueden verse afectados. Los sistemas deben estar en su lugar para monitorear, responder y notificar al personal sobre cualquier problema que surja. La planificación de la recuperación ante desastres es de vital importancia en caso de fallas importantes, pero los problemas menores también deben ser resueltos.

Enfriamiento y eficiencia energética

Los centros de datos deben tener controles ambientales estrictos y absorber o generar grandes cantidades de energía para mantener las cosas en funcionamiento. Y estos son costosos.

Dado que los servidores y otros equipos no funcionan muy bien en temperaturas extremas, la mayoría de los centros de datos tienen enormes sistemas de enfriamiento y flujo de aire que consumen grandes cantidades de energía y, a veces, agua. Los sensores deben estar en su lugar para monitorear las condiciones ambientales para que se puedan hacer ajustes.

No es solo la temperatura lo que es un problema. Factores como la humedad deben mantenerse bajo control.

Los racks de servidores a menudo se organizan en filas que crean pasillos donde los servidores están uno frente al otro o todos están uno frente al otro para controlar el flujo de aire y la temperatura de manera más eficiente. El pasillo donde se enfrentan es el pasillo frío (Figura 1), y el aire en el pasillo caliente se canaliza convenientemente.

El consumo de energía es otra preocupación importante. Es absolutamente necesario que estas instalaciones tengan acceso constante a la energía adecuada, algunas incluso tienen sus propias subestaciones de energía. Una medida utilizada para juzgar la eficiencia energética del centro de datos es la eficacia del uso de energía (PUE). Es un cálculo de la Energía de los servidores dividida por la Energía total (servidores y sistema de refrigeración).

Se están haciendo muchas cosas para reducir el consumo energético de los centros de datos y otras necesidades de recursos. Las salas de servidores solían mantenerse alrededor de los 15º C, pero la tendencia en los centros de datos más eficientes energéticamente es mantenerlos a unos 26º C, al menos en el pasillo frío, aunque no todos han adoptado esta práctica. Aparentemente, los servidores funcionan bien a esta temperatura, y requieren menos energía de refrigeración.

Existe una tendencia creciente a usar refrigeración al aire libre, que extrae aire del exterior en lugar de utilizar muchas unidades de aire acondicionado y enfriadores que requieren mucha energía. Otra tendencia es ubicar centros de datos cerca de fuentes de agua listas que puedan reciclarse para su uso en refrigeración, como el centro de datos de Google en Finlandia, que utiliza agua de mar. Otra es ubicar centros de datos en climas fríos.

Los cambios en el equipo informático real también pueden ayudar. Muchos componentes en los centros de datos pierden energía, lo que significa que parte de la energía que usan nunca llega al procesamiento real: se desperdicia. Reemplazar servidores antiguos con modelos más nuevos y más eficientes energéticamente obviamente ayuda. Pero el equipo también se puede rediseñar para requerir menos energía. La mayoría de los centros de datos usan servidores tradicionales y otros equipos, pero Google y Facebook usan servidores personalizados. Google fue diseñado para eliminar componentes innecesarios como tarjetas gráficas y minimizar la pérdida de energía en la fuente de alimentación y el regulador de voltaje. Los paneles que contienen el logotipo del fabricante se omiten para permitir un mejor flujo de aire hacia y desde los componentes, y la compañía fabrica algunos de sus propios equipos de red.

Además, los procesadores y los ventiladores también pueden reducir la velocidad cuando no son necesarios. Los servidores más eficientes también tienden a generar menos calor, reduciendo aún más el consumo de energía necesario para la refrigeración. Los servidores ARM de baja potencia, creados originalmente para dispositivos móviles pero rediseñados para usos del servidor, también están llegando a los centros de datos.

El uso de aplicaciones fluctúa dependiendo de lo que se está haciendo y a qué hora en varios software y aplicaciones web, cualquiera de los cuales tiene diferentes necesidades de recursos. La gestión de recursos de aplicaciones es importante para aumentar la eficiencia y reducir el consumo. El software se puede escribir a medida para que funcione de manera más eficiente con la arquitectura del sistema. La virtualización del servidor también puede reducir el consumo de energía al reducir el número de servidores en ejecución.

Sistemas de refrigeración en centros de datos

El consumo energético destinado a la refrigeración de un centro de datos puede superar el 40% del consumo total de energía.

La optimización de los sistemas de refrigeración en grandes centros de datos es esencial para reducir los costes de operación.

Los centros de datos modernos son sistemas complejos que incluyen instalaciones de TI, sistemas de energía, sistemas de refrigeración y ventilación.

En la actualidad, existen muchas vías de optimización de los sistemas de enfriamiento de centros de datos, entre las que cabe citar las siguientes:

- **Refrigeración por aire acondicionado.**

  Es la técnica más utilizada para refrigerar centros de datos. La sala donde están instalados los servidores cuenta con climatizadores que, bien por el techo o bien por el suelo, generan corrientes de aire que mueven el calor hacia unos extractores, los cuales llevan el aire caliente a un sistema de enfriamiento, y lo devuelven mediante ventiladores de gran potencia a las instalaciones.

  Este sistema permite una refrigeración muy constante y potente para los centros de datos, pero su mayor inconveniente lo encontramos en su consumo energético: tanto en el movimiento del aire (aspirado e insuflado) como en su propio enfriamiento se está consumiendo una gran cantidad de electricidad.

  La optimización de este sistema pasa por ajustar de forma inteligente los ciclos de refrigeración (reduciendo el movimiento de aire cuando la temperatura es adecuada, modificando la velocidad a la que se mueven los ventiladores) y mejorando la circulación natural del aire (creando configuraciones especiales de pasillos fríos y calientes que eviten que se mezcle el aire caliente con el frío, por ejemplo).

- **Refrigeración por aire exterior.**

  Una fórmula que se está imponiendo en los últimos tiempos es la utilización del aire exterior como forma de refrigerar los centros de datos, también conocido como 'free cooling'.

  Esta tendencia está llevando a la localización de grandes centros de datos en zonas de climas fríos para aprovechar las condiciones ambientales naturales para "enfriar" el aire que entra al CPD, evitando el consumo energético de la partida de refrigeración.

  Un caso paradigmático de esta tendencia se está produciendo en Islandia como foco de atracción para la localización de grandes centros de datos.

- **Refrigeración por agua**

  La principal alternativa a la refrigeración mediante aire es la refrigeración líquida. En estos casos, se utiliza el agua como mecanismos para enfriar no el centro de datos, sino los propios servidores desde su interior. Lo que se hace es introducir agua fría (bien enfriada de forma artificial o aprovechando el agua fría de un río o manantial cercano) para que pase por una serie de tuberías dispuestas en torno o dentro de los racks, enfriando por contacto los equipos en cuestión.

- **Refrigeración con CO_2**

  La refrigeración utilizando CO_2 consiste en utilizar dióxido de carbono como refrigerante. A grandes rasgos en las salas de CPD hay unos evaporadores donde se gasifica el CO2, absorbiendo el calor desprendido en los racks que será descargado en el exterior a través de condensadores.

- **Refrigeración por Inmersión**

  Esta nueva técnica de enfriamiento se está imponiendo en los últimos tiempos y consiste en sumergir los servidores líquido dieléctrico. El líquido en contacto directo con componentes calientes absorbe el calor que, a continuación, a través de intercambiadores de calor lo disipa en el exterior.

  Esta técnica permite prescindir de los ventiladores de refrigeración y es altamente eficiente.

  Presenta el inconveniente de la aparición de fallos en los circuitos como consecuencia del medio al que están expuestos.

El centro de datos está compuesto por unidades aisladas térmicamente en las que se disponen los racks de servidores de forma perimetral. Estas unidades aisladas pueden ser aisladas en frío (Cold Aisle) o en calor (Hot Aisle). En las unidades tipo Cold Aisle, entra el aire frío del CRAC por el suelo, circula por un pasillo aislado de donde toman el aire los racks gracias a unos ventiladores que hacen pasar el aire frío por los servidores y salen al exterior, donde se unen con el aire de otras unidades que circulan en retorno al CRAC.

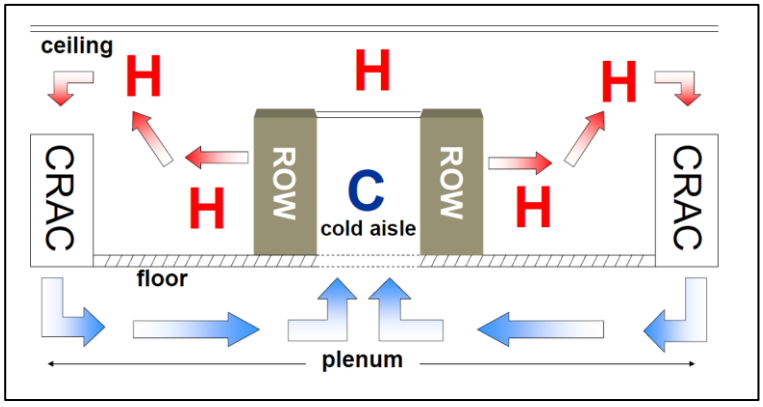


Figura 1. Unidad asilada en frío

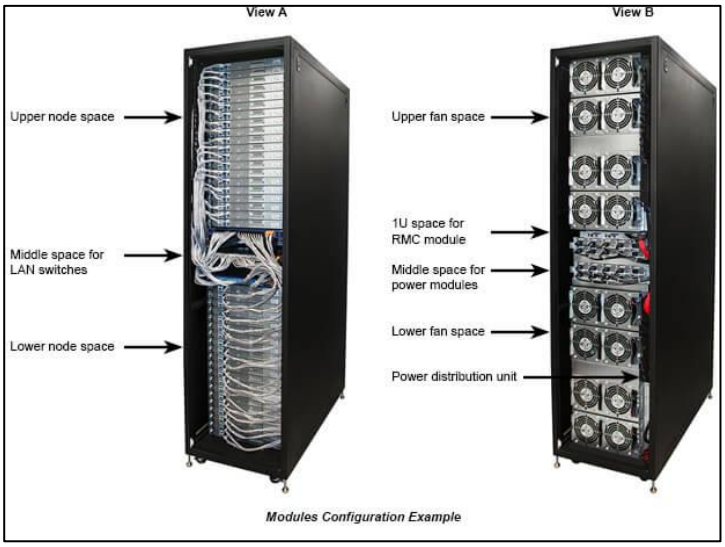


Figura 2. Racks

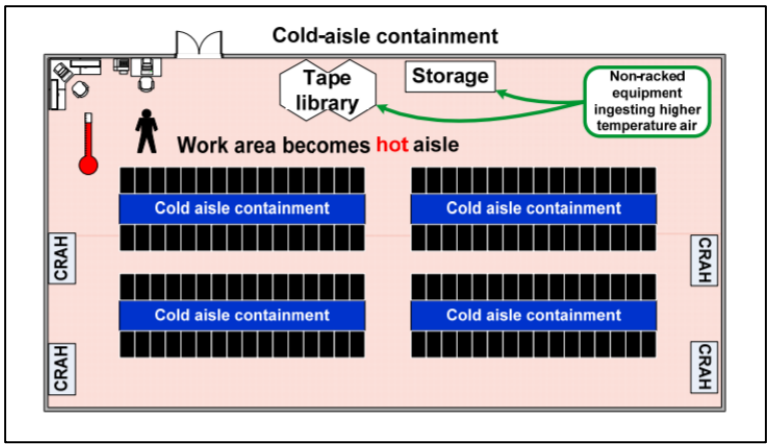


Figura 3. Cold aisle containment

3 HERRAMIENTA DE SIMULACIÓN ARENA®

Conceptos generales

La simulación nos permite comprender el comportamiento de un sistema real a través de un modelo con la finalidad de analizarlo y tomar decisiones sobre cómo mejorarlo. A través del software de modelado de procesos, podemos obtener resultados mucho antes de que surjan repercusiones de las decisiones tomadas.

En este proyecto se utilizará el software de modelado Arena® Simulation para modelar y simular el funcionamiento de un centro de datos.

Podemos diferenciar dos tipos de sistemas según el valor que toman las variables que caracterizan a los componentes del sistema:

- Discretos: el valor de las variables cambia en instantes determinados y separados en el tiempo.
- Contínuos: el valor de las variables cambia contínuamente con respecto al tiempo.

A la hora de diferenciar el tipo de sistema lo hacemos teniendo en cuenta el tipo de cambio en el valor de las variables que predomina, ya que un sistema no suele ser completamente contínuo ni completamente discreto.

El modelado de eventos discretos es el proceso de representar el comportamiento de un sistema complejo como una serie de eventos bien definidos y ordenados, y funciona bien en prácticamente cualquier proceso donde haya variabilidad, recursos limitados o interacciones de sistemas complejos [2].

La simulación discreta de eventos permite analizar el comportamiento de un proceso o sistema a lo largo del tiempo, y diseñar procesos nuevos o cambiar los ya existentes sin ninguna cosecuencia.

Gracias a ello podemos encontrar respuesta a cuestiones como cómo alcanzar nuestros objetivos de desempeño, cúando aumentar o reducir los recursos o el impacto que tendrán los cambios.

El software de simulación de procesos de Arena permite crear estos modelos de procesos, proporcionándo la inteligencia necesaria para reducir costos, medir el rendimiento y optimizar sus operaciones. Con Arena es posible simular el rendimiento futuro del sistema y de esta manera identificar oportunidades de mejora, visualizar operaciones con gráficos dinámicos de operación, analizar el comportamiento del sistema en su estado "as-is" (su estado actual) y poder elegir la mejor alternativa para el "to-be" (posible estado futuro) [2]. Arena utiliza el lenguaje de programación SIMAN.

Algunas de las características del software de simulación de eventos discretos de Arena son [2]:

- La metodología de modelado de diagrama de flujo incluye una gran biblioteca de bloques de construcción predefinidos para modelar su proceso sin la necesidad de una programación personalizada.
- Gama completa de opciones de distribución estadística para modelar con precisión la variabilidad del proceso.
- Análisis estadístico y generación de informes.
- Métricas de rendimiento y paneles de control.
- Capacidades realistas de animación 2D y 3D para visualizar resultados más allá de los números.

Algunas de las ventajas que brinda [2]:

- Mejorar la visibilidad del efecto de un cambio de sistema o proceso
- Explorar nuevos procedimientos o métodos sin interrumpir el sistema actual
- Diagnosticar y solucionar problemas.
- Reducir o eliminar los cuello de botella
- Reducir los costes operativos
- Mejorar el pronóstico financiero
- Evaluar mejor los requisitos de hardware y software.
- Reducir los tiempos de entrega
- Administrar mejor los niveles de inventario, personal, sistemas de comunicación y equipos.
- Aumentar la rentabilidad a través de mejoras en las operaciones generales

Arena® Simulation está considerada como una de las mejores herramientas de simulación, teniendo en cuenta sus características técnicas, capacidades funcionales y la alta valorción que tiene por sus usuarios.

Nociones básicas de Arena.

Se muestra a continuación la ventana principal de Arena (*Figura 4*) en la que podemos diferenciar tres zonas.

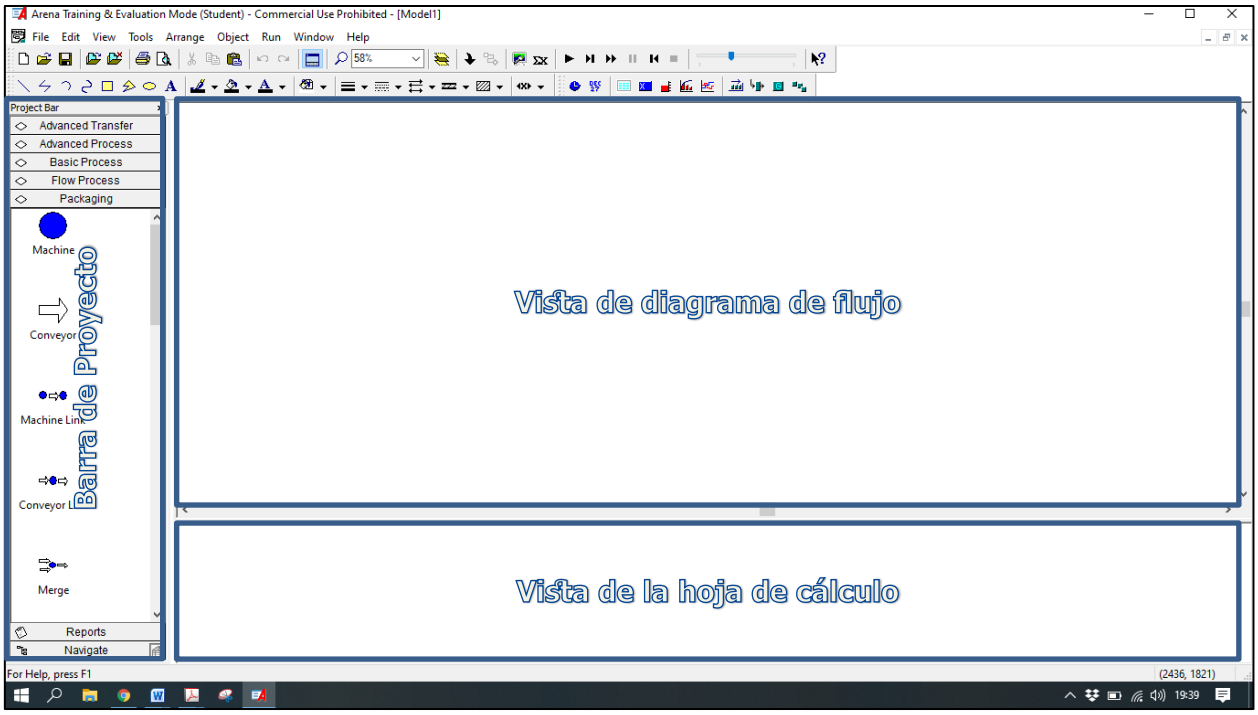


Figura 4. Ventana principal de Arena

- Vista de diagrama de flujo: Contiene el diagrama y los elementos gráficos de la simulación.
- Vista de la hoja de cálculo: Contiene la información del modelo. Datos como tiempos, costes y otros parámetros. Estos parámetros se pueden ver y editar también en la vista del diagrama pero en sistemas como el nuestro resulta más cómodo en la vista de la hoja de cálculo ya que nos da el acceso a muchos de los parámetros a la vez.
- Barra de Proyecto: Contiene los paneles con los elementos necesarios para diseñar los modelos. Estos bloques que nos permiten configurar el modelo se llaman módulos.

Además de estas tres regiones podemos diferenciar la Barra de Herramientas (en la parte superior de la ventana) y la Barra de Estado (situada en la parte inferior de la ventana) que nos facilita información sobre el estado de la simulación.

Podemos diferenciar dos tipos de módulos: los módulos de datos y los módulos de diagrama.

Los módulos de datos (Figura 5 y Figura 6) nos permiten la descripción estática del modelo. Definen las características de varios elementos del proceso como:

- Entidades: objetos dinámicos que circulan a través del Sistema cambiando su estado.
- Atributos: características específicas de las entidades.
- Recursos: elementos característicos del Sistema utilizados por las entidades. Existen atributos asignados automáticamente por Arena como *Entity Create Time* o *Entity Type*
- Colas: las entidades deben esperar una cola para acceder a los recursos si estos están ocupados.
- Variables: características del Sistema que pueden cambiarse durante el periodo de simulación. Al igual que los atributos, hay algunas variables que ya vienen definidas por Arena como pueden ser el estado de un recurso *STATE(recurso)* o el tiempo de simulación *TNOW.*

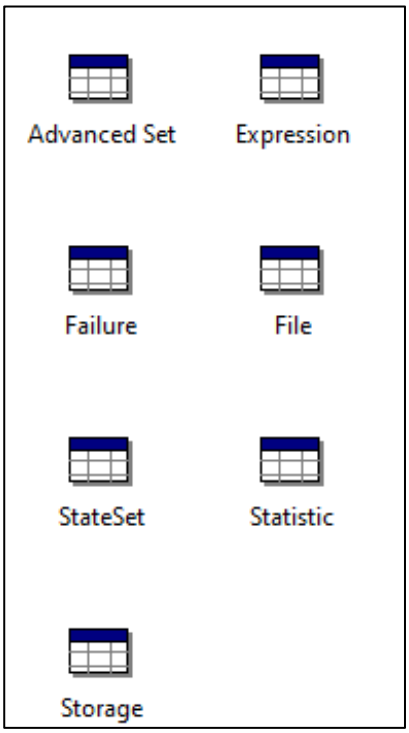


Figura 5. Módulos avanzados de datos

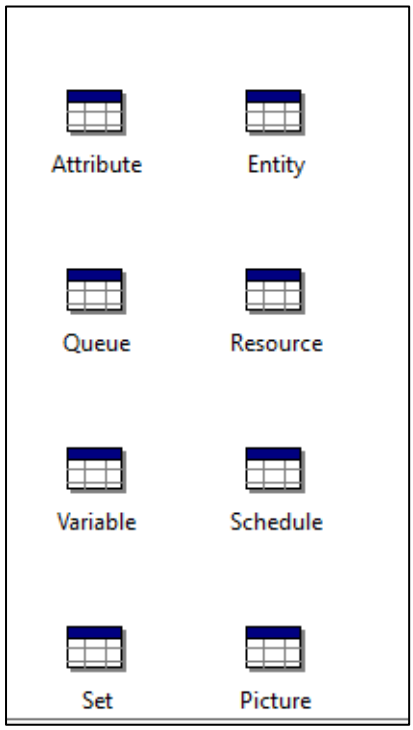


Figura 6. Módulos básicos de datos

Los módulos de los diagramas (*Figura 8* y *Figura 7*) describen el Sistema dinámicamente. Estos módulos se suelen conectar unos con otros y a través de ellos fluyen o se destruyen las entidades. Podemos encontrar distintos tipos.

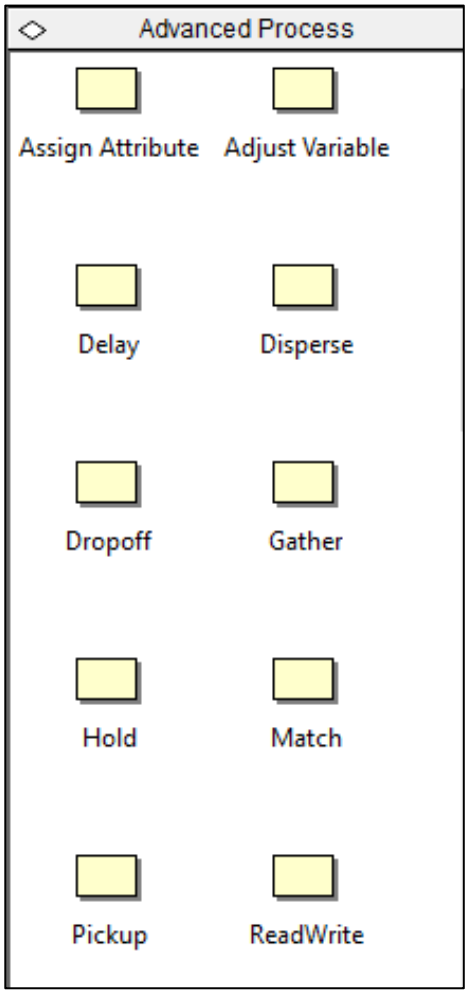


Figura 7. Módulos avanzados de diagramas

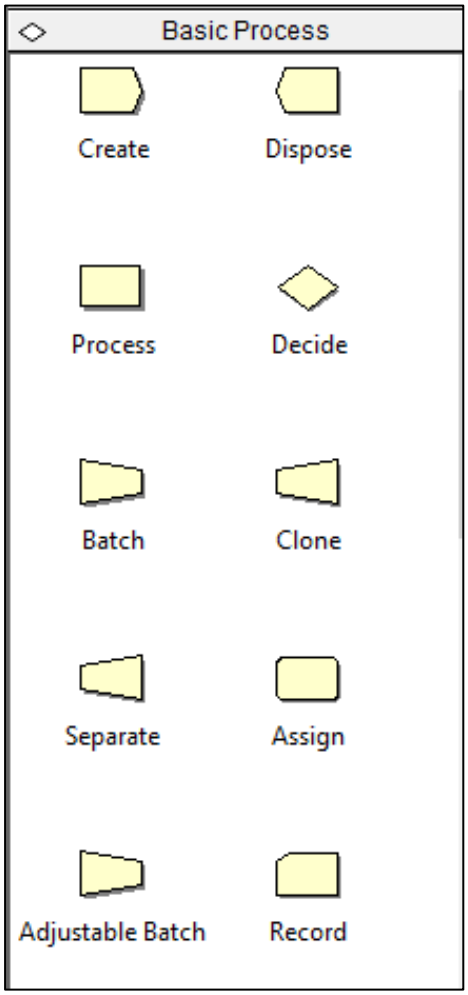


Figura 8.Módulos básicos de diagramas

4 MODELADO DINÁMICO DE CENTRO DE DATOS

Conceptos preliminares

Un modelo es una representación abstracta de un sistema real. Un sistema informático puede representarse con ecuaciones, árboles decisionales, modelos gráficos, etc. La elección de la representación afectará a la exactitud del modelo.

Para gestionar el consumo energético de un centro de datos existen 4 pasos: extracción de características, construcción del modelo, validación del modelo, aplicación del modelo [3].

La fase de extracción de características consiste en identificar cuáles son los componentes que más energía consumen y utilizarlos como inputs para la construcción del modelo de consumo energético. Este sería el output de la segunda fase. Una vez construido el modelo, necesita ser validado para utilizarlo como predictor de consumo energético. Estas predicciones nos sirven para optimizar la eficiencia energética de los CPD [4].

En este apartado nos centraremos en la construcción del modelo dinámico de un centro de datos.

Modelar el comportamiento energético de un centro de datos con exactitud, ya sea de todo el sistema o sólo de uno de sus componentes, no es una tarea sencilla.

El modelo de consumo energético de un centro de datos depende de diversos factores como las especificaciones hardware, las cargas de trabajo, los tipos de aplicaciones, los requisitos de enfriamiento, etc. Debido al sobrecalentamiento que causan las infraestructuras que se utilizan para medir el consumo energético de todos los componentes del sistema se han desarrollado distintas técnicas que pueden estimar el consumo de energía de un sistema para una carga de trabajo dada [4].

Para la realización del modelo nos hemos basado en las últimas tendencias y estudios de la eficiencia energética en centros de datos.

Tendremos en cuenta un modelo típico de centro de datos con una configuración de aisamiento en frío, o Cold Aisle y patron de flujo de aire. En este tipo de unidades (*Figura 1*), como se ha señalado anteriormente, entra el aire frío del CRAC por el suelo, circula por un pasillo aislado, los racks gracias a los ventiladores absorven el aire frío de este pasillo y lo expulsan a través de los servidores al exterior, donde se unen con el aire de otras unidades que circulan por el CRAC.

El centro de datos da servicio a un número de usuarios de la nube (CUs) que alquilan servidores a un proveedor de la nube (CP) para obtener la capacidad informática requerida.

El Sistema de gestión es el que asignará y gestionará las tareas manipulando el nivel de tension y la escala de frecuencia (DVFS) del servidor con objeto de mejorar su eficiencia energética en

función de la carga de trabajo de los servicios que se gestionan por máquinas virtuales (VM).

El proveedor se compromete a una calidad de servicio (QoS) que consiste en garantizar un tiempo de respuesta máximo y a una fiabilidad del sistema en forma de restricciones blandas de temperatura.

Modelado del consumo de energía en el centro de datos.

El consumo de energía en el centro de datos deriva de la operación de los servidores y del Sistema de refrigeración. En este Proyecto no vamos a tener en cuenta infraestructuras fundamentales. Podemos aproximar el consumo total de energía como la suma de energía consumida por TI y por los subsistemas de enfriamiento.

Modelado del consumo de los servidores

Suponemos un total de J usuarios que solicitan servicio al centro de datos. Para un usuario j de los J existentes tenemos:

- L_j como el número de peticones del usuario que se tienen que gestionar. Es una perturbación, es decir no es algo que podamos manipular en nuestro modelo.
- m_j como número de servidores que dan servicio al usuario de los M contratados. Es una variable manipulable con la que podemos mejorar la eficiencia del Sistema.
- p_j como el consumo de los servidores del usuario.

Para una cantidad de peticiones L_j y un número de servidores en uso m_j el Sistema de gestión de los servidores y de virtualización reparten las tareas entre los servidores, tratando de mejorar la eficiencia regulando la temperature de servicio, esto se consigue ajustando la frecuencia y el voltaje del servidor. De manera que cuanto mayor sea la carga por servidor, es decir el número de tareas entre el número de servidores en uso, menor será la frecuencia de funcionamiento y en consecuencia al aumento del tiempo de servicio, empeorará la calidad QoS.

Así, podríamos decir que cuanto mayor sea el número de servidores mejor será la calidad del servicio. Sin embargo, esto aumentaría el consumo energético.

El consumo de energía de un solo servidor se puede escribir de la siguiente manera [5]:

$$p_j\,(t) = a_1 \times ma_j(t) + a_2 \times m_j(t) \qquad (1)$$

Siendo ma_j el número de servidores ocupados para j en el instante t, m_j el número de servidores en activos para el usuario j, a1 el consumo marginal del servidor que dependerá de su frecuencia s y a2 el consumo del servidor sin carga de trabajo, generado por componentes como la fuente de alimentación y dispositivos de almacenamiento [6].

$$a_1 = C_f \times s_{ij} \qquad (2)$$

De este modo, podemos diferenciar tres diferentes estados de un servidor: apagado en el que no consume potencia, ocioso en el que tan solo debemos tener en cuenta el consumo del servidor sin carga de trabajo, y ocupado en el que además tenemos que contar con el consumo marginal del servidor.

Cuando un servidor se enciende (pasa del estado inactivo a activo), trascurre un periodo de tiempo en el que está consumiendo potencia pero no puede ser ocupado, su capacidad es cero. Denominamos este periodo de tiempo como tiempo de arranque T_{arr}.

Así, el consumo del servidor i asociado al usuario j vendrá dado por

$$p_{ij} = \begin{cases} a_2 & si\ está\ ocioso\ o\ arrancando \\ a1 + a2 & si\ está\ ocupado \end{cases}$$

El consumo medio asociado al usuario j será:

$$p_j = a_1 \times L_j + a_2 \times m_j \qquad (3)$$

Por tanto, el consumo total de energía de los servidores en el centro de datos será:

$$P(t) = \sum_{j=1}^{j=J} p_j(t) \qquad (4)$$

Modelado de la cola de los servidores

La gestión de los servicios del usuario se modela mediante una cola $M/M/m_j$ (Figura 9) [5], donde M_j es el número de servidores disponibles para el usuario j. Por lo tanto habrá una única cola para las peticiones de cada usuario aunque el número de recursos disponibles sea M_j.

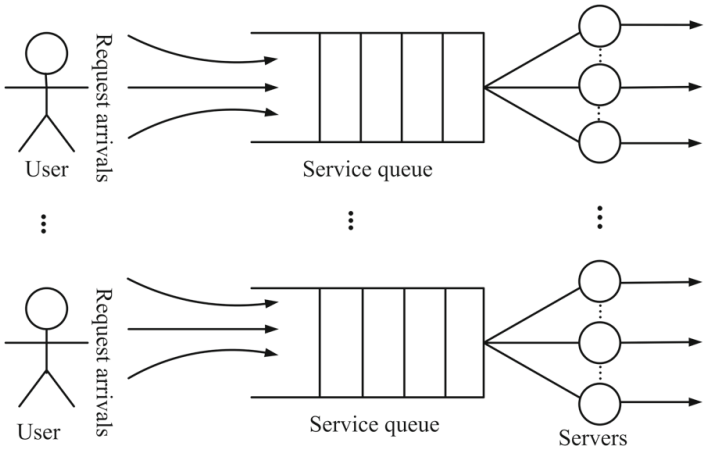


Figura 9. Cola $M/M/m_j$

Tanto el tiempo entre llegadas de peticiones como el tiempo de servicio (tiempo que tarda el servidor en procesar la petición) siguen una distribución exponencial de probabilidad.

La función de densidad del tiempo entre llegadas de peticiones cuyo tiempo medio es T_{lleg} viene dada por:

$$f(t) = \frac{1}{T_{lleg}} \times e^{-\frac{t}{T_{lleg}}} \qquad (5)$$

También podemos modelar la llegada de las peticones mediante el numero de peticiones que llegan en un intervalo de tiempo ΔT_{lleg}. En este caso sigue una distribución de probabilidad de poisson cuya función de densidad viene dada por:

$$f(n) = \frac{\left(\frac{\Delta T_{lleg}}{T_{lleg}}\right)^n \times e^{-\frac{\Delta T_{lleg}}{T_{lleg}}}}{n!} \qquad (6)$$

Por otro lado, el tiempo de servicio del servidor i es otra variable aleatoria $T_{ser\ ij}$ que sigue una distribución de probabilidad exponencial cuya función de densidad viene dada por:

$$f(t) = \frac{1}{T_{ser\ ij}} \times e^{-\frac{t}{T_{ser\ ij}}} \qquad (7)$$

La velocidad de procesado de cada servidor i viene regulada por su frecuencia s_{ij} (medido en términos de instrucciones por segundo). Así, el tiempo medio de servicio será $T_{serij} = Kj/s_{ij}$, [5] siendo K_j una media de la complejidad (en número de instrucciones) del servicio que solicita el usuario.

La disciplina de la cola representa la modalidad en la que se procesan las tareas que van llegando. Nosotros asumimos una cola FIFO (first-in/first-out) aunque podríamos tomar diversas modalidades como LIFO (last-in/first-out), random o con alguna preferencia a especificar.

Siendo λ el número medio de llegadas previstas y μ el número medio de peticiones que se procesan, en el caso en el que λ > μ la cola tendería a crecer indefinidamente (cola infinita). Esto crea una situación de inestabilidad de la cola, y por tanto del Sistema. Como consecuencia impondremos la condición de estacionariedad:

$$M \times \frac{1}{Tser} > \frac{1}{Tlleg} \qquad (8)$$

Modelado térmico de los servidores

Para esto, realizaremos un balance térmico de cada servidor y asumiremos que la temperatura de la unidad aislada en frío es θ_c. La dinámica de la temperatura del servidor i asociada al usuario j (θ_{ij}) vendrá dada por:

$$K_t \times \frac{d\theta_{ij}}{dt} = C_p \times q_a \times \left(\theta_c - \theta_{ij}\right) + p_{ij} \qquad (9)$$

Siendo K_t la capacidad térmica del servidor, C_p la capacidad térmica del aire y q_a el caudal de aire.

En la práctica la temperatura puede superar durante un intervalo de tiempo la temperature máxima, es por eso que consideramos las restricciones de temperatura restricciones blandas aunque también pueden considerarse restricciones duras [5].

Con el fin de no superar la temperatura máxima podemos manipular la frecuencia del servidor,que afecta al consumo marginal a_1 y por tanto, a la potencia p_{ij}. Al disminuir la frecuencia, disminuye la potencia marginal, y a su vez la temperatura. Esta disminución de frecuencia aumentaría el tiempo de cálculo del servidor y a su vez el tiempo de retraso del Sistema.

Modelado de la máquina de refrigeración

Las máquinas de refrigeración se suelen diseñar con un sistema de control que manipulala velocidad del compresor refrigerante para regular la temperatura del aire de salida, con un caudal de aire constante. El controlador podrá garantizar que la temperature de salida del aire sea la de consigna siempre que el salto térmico que deba aportar la máquina sea menor que el máximo de diseño. Esta situación suele darse en las máquinas de frío, ya que estas se suelen sobredimensionar.

Podríamos de este modo, modelar la temperatura de la máquina como un Sistema de primer orden de ganancia estática unidad. Teniendo la temperatura de consigna del aire θ_r, como la variable manipulable.

$$\tau \times \frac{d\theta_c}{dt} = \theta_r - \theta_c \qquad (10)$$

Podemos definir el consumo de un CRAC con la siguiente expression [7]:

$$P_{CRAC} = \frac{P}{CoP(\theta_c(t))} \qquad (11)$$

Siendo P la potencia demandada por el centro de datos, y CoP describe la eficiencia de refrigeración de la máquina. Es una función parabólica y podemos definirla como [7]:

$$CoP(\theta) = b \times \theta^2 + c \times \theta + d \qquad (12)$$

Donde a, b, y c dependen de la máquina y los facilita el fabricante.

Restricciones del modelo

El tiempo de respuesta del sistema lo podemos calcular de la siguiente manera [5]:

$$T_{resj} = \frac{1}{\sum_i^{m_j} \mu_{ij} - \lambda_j} \qquad (13)$$

Como bien hemos señalado anteriormente, μ_{ij} es el número medio de servicios, lo podemos calcular como el ratio entre la frecuencia s y la media de la complejidad en número de instrucciones K_j:

$$\mu_{ij} = \frac{s_{ij}}{K_j} \qquad (14)$$

Cada usuario permite un tiempo medio de servicio máximo que denominamos D_j. Podemos representar la restricción de QoS como:

$$T_{resj} \leq D_j \qquad (15)$$

Otra de las restricciones del sistema es el número de servidores activos m_j, depende del número de servidores que el usuario haya contratado M_j.

$$0 \leq m_j \leq M_j \qquad (16)$$

El número de servidores apagados será por tanto $M_j - m_j$.

Como referimos anteriormente otra de las restriccionesde este Sistema es la temperatura máxima. Utilizamos la temperatura media del servidor en vez de utilizar la restricción dura. Esta puede superarse en determinados intervalos de tiempo.

Índice de desempeño

Se suele tomar como índice de desempeño Performance Usage Efficiency (PUE) que es la razón entre la energía consumida por los servidores y la energía total consumida (servidores y sistema de refrigeración).

$$PUE(t) = \frac{E_c(t)}{E_c(t) + E_{CRAC}(t)} \qquad (17)$$

Siendo $E_c(t) = \int_0^t P(\tau)d\tau$ y $E_{CRAC}(t) = \int_0^t \frac{P(\tau)}{CoP(\theta_c(\tau))} d\tau$

En el índice de desempeño global, además del PUE también se pueden incluir términos que penalicen el exceso de temperatura de los servidores o la calidad del servicio del centro de datos.

Definición de variables del modelo

Las variables manipulables son las que cambiamos para influir en nuestro sistema y estudiar los distintos comportamientos que causan en él . Se definen a continuación las variables manipuables de nuestro sistema:

- El número de servidores activos m_j
- La frecuencia de trabajo de los servidores s_{ij}
- La temperatura de consigna del aire θ_r

Existen distintas variables internas en el sistema:

- Vcap: esta variable indica si elservidor está apagado o encendido.

$$Vcap = \begin{cases} 1 & si\ el\ servidor\ está\ encendido\ o\ en\ arranque \\ 0 & en\ casocontrario \end{cases}$$

- Vcap_ant: la capacidad del servidor en el instante de simulación anterior.
- VcapR: la capacidad real del servidor.

$$VcapR = \begin{cases} 0 & si\ el\ servidor\ está\ apagado\ o\ en\ arranque \\ 1 & en\ caso\ contrario \end{cases}$$

- TVcap: el tiempo en el que se debe actualizar VcapR.
- Tant: el instante anterior de simulación.
- La temperatura de refrigeración θ_c

Las variables de salida son las que nos interesa estudiar el cambio que han sufrido al variar las variables manipulables. En nuestro sistema son:

- La temperatura de los servidores θ_{ij}
- La potencia de los servidores p_{ij}
- La potencia del CRAC P_{CRAC}

Dimensionamiento de parámetros

Nos basamos en el artículo de Fu [5], donde se muestran datos experimentales de un centro de datos, para deducir un posible valor de los parámetros físicos de nuestro modelo:

En el artículo se consideran 1700 servidores y 4 aplicaciones que solicitan tareas que varían entre 1000 y 150000 tareas por segundo cada aplicación. La media se estima en 100000. Escalando estos valores a nuestro modelo, con una simple reglade tres obtenemos que la media de llegadas para nuestro modelo será 580, tomaremos 600.

Como se explica anteriormente para evitar el problema de la cola infinita la media de llegadas debe ser menor que la media de procesadas:

$$M \times \frac{1}{Tser} > \frac{1}{Tlleg}$$

Tenemos $M = 10$ servidores y $\frac{1}{Tlleg} = 600$, por tanto el tiempo de servicio T_{ser} se estima 0,01 y el T_{lleg} 0,0016. Sin embargo, para un mejor desarrollo de la simulación en Arena hemos multiplicado ambos tiempos por diez, resultando $T_{ser} = 0{,}1$ y $T_{lleg} = 0{,}016$. Se necesitará un mínimo de 7 servidores activos. Sin embargo, esta restricción puede violarse durante un período determinado de tiempo.

Hemos considerado una frecuencia $s_{ij} = 1$ para todos los servidores. Por tanto, sabiendo que el tiempo de servicio es la media de la complejidad de las instrucciones K dividida por la frecuencia, tendremos $K = 0{,}01$.

En el artículo se considera un consumo energético $a1 = a2 = 40W$. Por tanto, teniendo $a_1 = C_f \times s_{ij}$, se obtiene $C_f = 40$.

Para el dimensionamiento de $C_p q_a$ se utiliza el balance térmico en condición de máxima potencia (80W) para este caso la temperatura máxima de servicio es $\theta_c + 50$ ºC. Despejando obtenemos de $C_p q_a = 1{,}6$. La constante de tiempo K_t se estima de 1 segundo.

La temperatura de refrigeración $15 \leq \theta_c \leq 25\ ºC$, en nuestro modelo $\theta_c = 20ºC$ como valor inicial. La constante de tiempo τ se estima del orden de 10 segundos.

Consideramos los coeficientes de la máquina $a = \ b = \ c = 1$.

Consideramos el valor inicial de la temperatura de los servidores $\theta_{ij} = 20ºC$ y la temperatura de consigna del aire $\theta_r = 20ºC$.

Asumimos un tiempo de arranque T_{arr} =1 segundo y un tiempo medio de retraso máximo $D_j = 0{,}05$ segundos.

5 IMPLEMENTACIÓN EN *ARENA*®

Una vez adquiridas las nociones básicas de Arena y definido el modelo, podemos pasar a describir la implementación.

Implementación del modelo

En este modelo tendremos en cuenta un usuario con diez servidores contratados.

En la siguiente imagen se muestra una vista general del diseño.

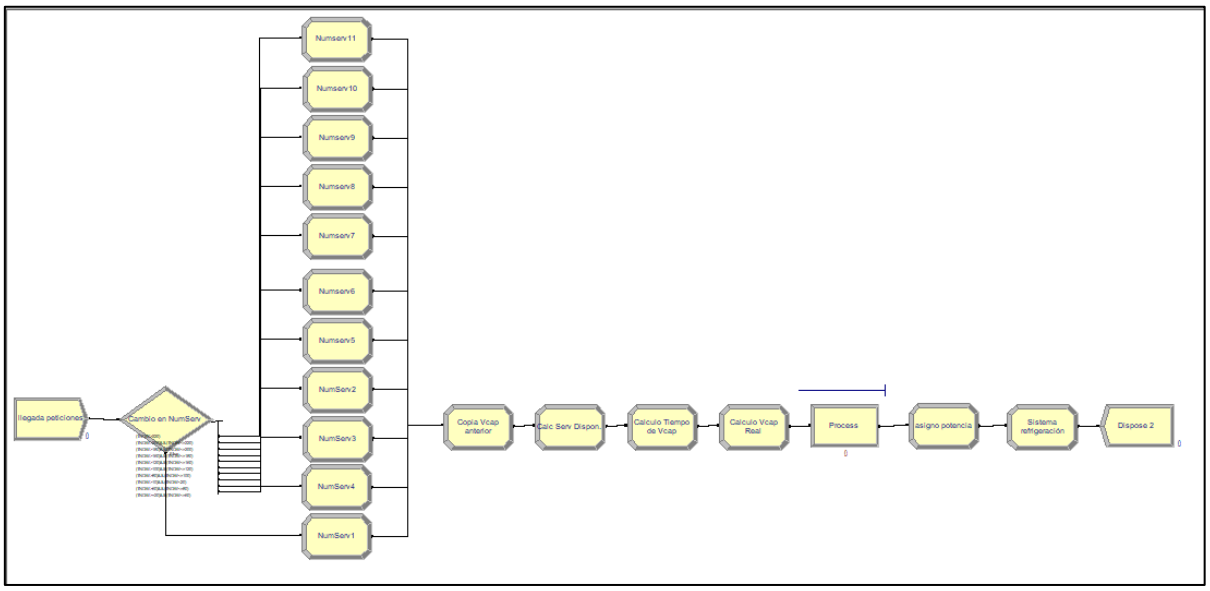


Figura 10. Vista general del modelo

En el módulo *llegadas peticiones* utilizamos la expresión EXPO con media T_{lleg} ya que como se muestra en el apartado anterior las llegadas siguen una distribución exponencial de probabilidad. Llamamos peticiones al tipo de entidad (*Entity type*), como las peticiones llegarán de una en una el valor de entidades por llegada (*Entities per Arrivals*) será uno.

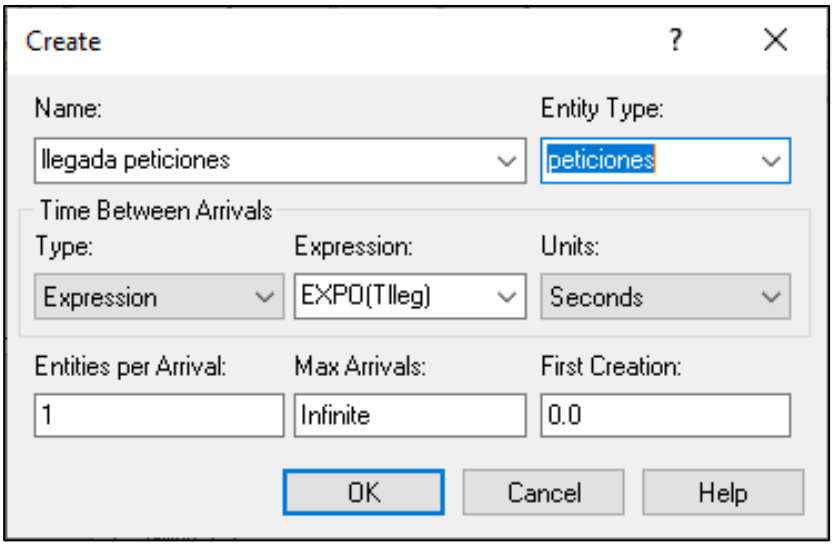


Figura 11. Ajustes del módulo create

Una de las variables manipulables del modelo es el número de servidores activos *NumServ*. En el siguiente módulo decisional se expresan las condiciones para que esta variable tome un valor u otro.

Para el módulo *cambio en NumServ (Figura 13*) elegimos el tipo *N-way by condition* ya que tenemos once posibles caminos dependientes de una condición impuesta, y no de una probabilidad. Las condiciones dependen del tiempo que lleve la simulación, el tiempo *TNOW* (una variable que viene definida por Arena).

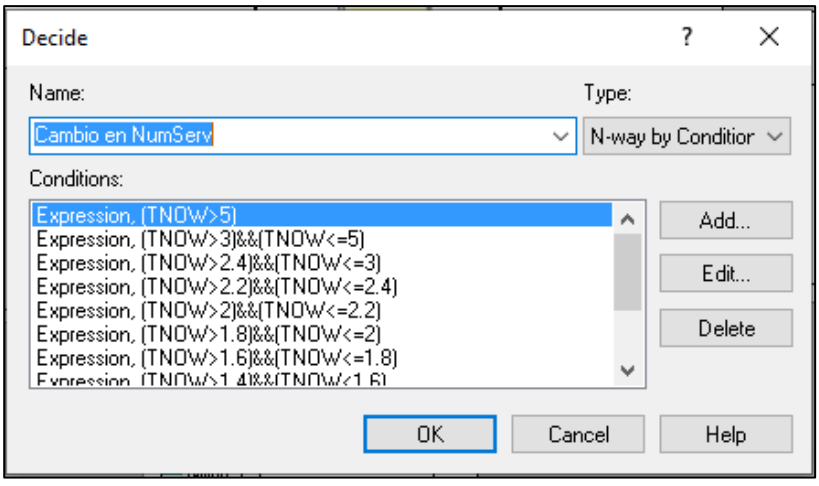


Figura 12. Ajustes del módulo decide

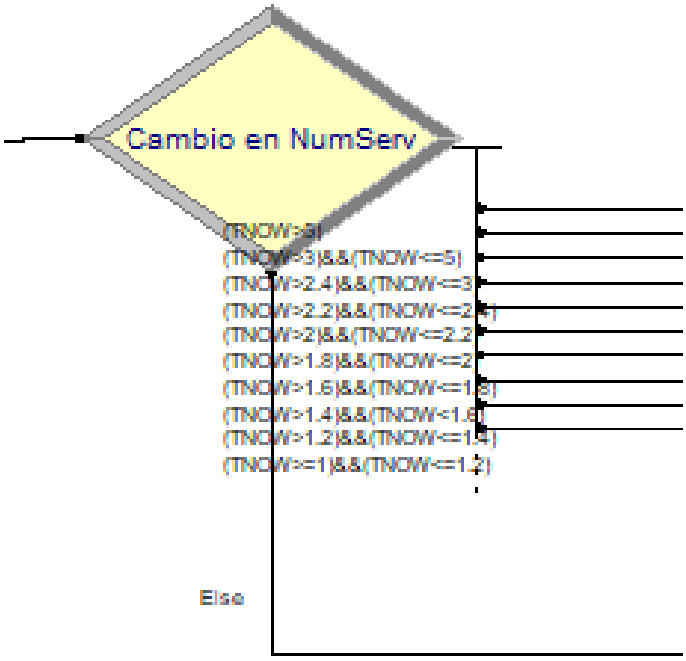


Figura 13. Módulo Cambio en NumServ

Estos once caminos llevan a once módulos diferentes de asignación donde se da valor a la variable *NumServ* dependiendo del tiempo que lleve la simulación. Según estas condiciones (*Figura 12*) se irá aumentando el número de servidores activos de uno en uno hasta activar los diez servidores.

Además de los caminos de las diez condiciones tenemos un último camino, el de *else.* Se da cuándo no se cumple ninguna de las condiciones especificadas en el módulo decisional. Conlleva al modulo de asignación *NumServ1 (Figura 14)*, donde se asigna a la variable *NumServ* el valor de diez.

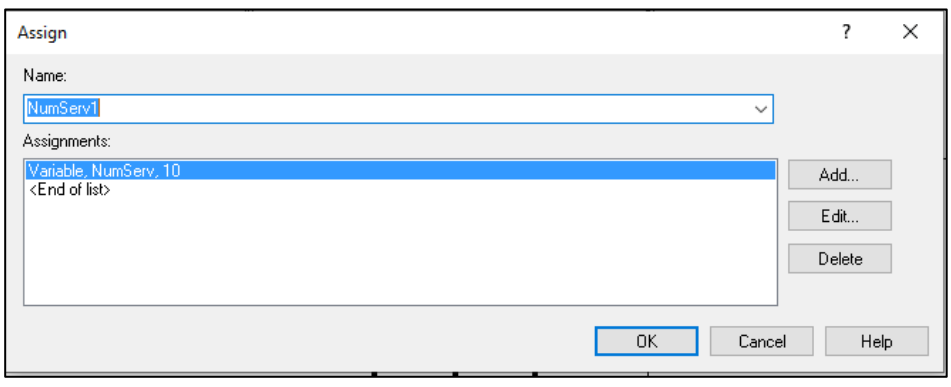


Figura 14. Asignación NumServ1

Si en vez de ir activando de uno en uno todos los servidores quisiéramos estudiar el comportamiento del sistema al pasar de un número concreto de servidores a otro, bastaría con utilizar el tipo del bloque *decide* de *two-way-by-condition* y asignar en los bloques de *assign* el número de *NumServ* que queramos en cada caso. Esto es aplicable con cualquiera de las condiciones y de las variables manipulables del sistema.

Para el modelado del procesamiento de las peticiones se utiliza un módulo *process* de tipo *Standard.* A las peticiones que lleguen se les asignará un recurso para que las procese con un tiempo de retraso y después lo liberen. Por tanto, elegimos la acción *seize delay release.* El tiempo de servicio sigue una distribución exponencial de media $1/\mu_j$, y de unidad segundos *Figura 16*).

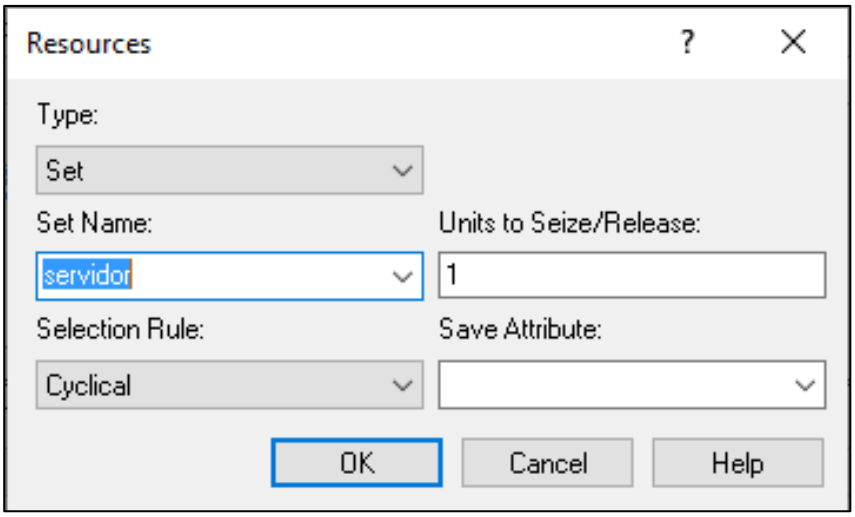


Figura 15. Ajustes del set

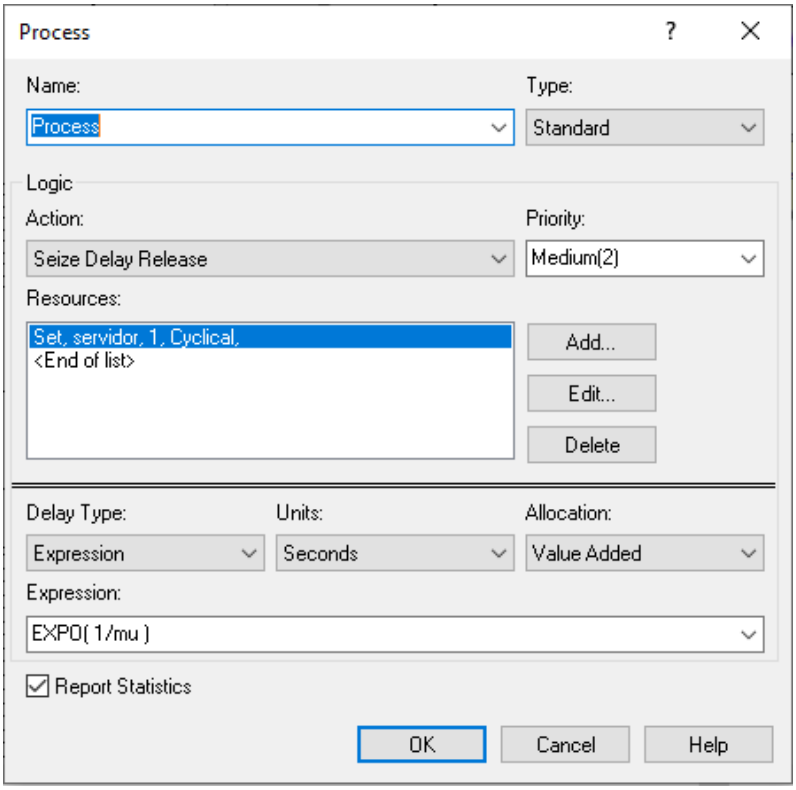


Figura 16. Ajustes del módulo process

Existe un único módulo de proceso ya que la cola es única para todos los servidores. Queremos conseguir que cada servidor tenga una capacidad variable e independiente a la de los demás servidores. Es por esto que no utilizamos un recurso con capacidad 10, si no un set de recursos con capacidad programable.

Añadimos el set de recursos que llamaremos *servidor (Figura 15),* formado por diez recursos. Cada recurso procesa las peticiones de una en una y las suelta, de manera que el valor que debemos poner en *unites to seize-release* es uno. La norma de asignación del recurso la hemos supuesto cíclica, aunque podríamos asignarlos de manera aleatoria o con algún orden de preferencia, por ejemplo se que se ocupe antes el servidor cuya temperatura sea menor.

Los diez recursos del set *servidor* se definen en el módulo de datos *set (Figura 17).* El tipo del set será recursos (*Resources)*. Los recursos se añaden manualmente *(R1, R2, R3, R4, R5, R6, R7, R8, R9 y R10).*

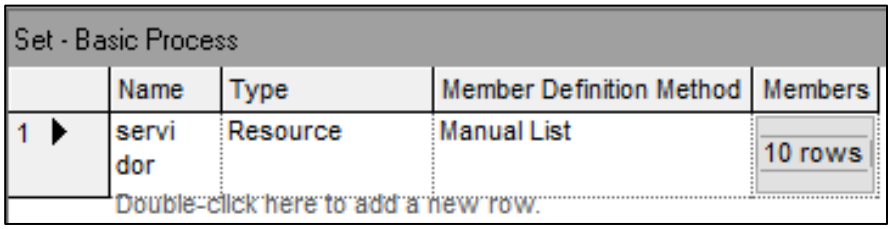


Figura 17. Módulo de datos Set

Cada recurso tiene una capacidad. La capacidad es el número de peticiones que puede procesar a la vez. En este caso la capacidad del recurso no es constante ya que será uno si el recurso está activo y cero si no lo está. Para ello vamos al módulo de datos *Resource (Figura 18*) y seleccionamos para los diez servidores el tipo *based on Schedule,* como se expresa anteriormente queremos que la capacidad sea variable, eso lo conseguimos basándola en una programación. Existen tres tipos diferentes de *Schedule rule:* Wait (acaba la tarea que esté realizando para cambiar la capacidad), ignore (cambia la capacidad aunque esté realizando una tarea y si no está acabada la deja sin acabar), preempt (cambia la capacidad aunque esté realizando una tarea, pero cuando vuelva a tener capacidad uno la acaba). En este caso se elegirá wait.

Resource - Basic Process

	Name	Type	Schedule Name	Schedule Rule	Busy / Hour	Idle / Hour	Per Use	StateSet Name	Failures	Report Statistics
1	R1	Based on Schedule	InR1	Wait	0.0	0.0	0.0		0 rows	☑
2	R2	Based on Schedule	InR2	Wait	0.0	0.0	0.0		0 rows	☑
3	R3	Based on Schedule	InR3	Wait	0.0	0.0	0.0		0 rows	☑
4	R4	Based on Schedule	InR4	Wait	0.0	0.0	0.0		0 rows	☑
5	R5	Based on Schedule	InR5	Wait	0.0	0.0	0.0		0 rows	☑
6	R6	Based on Schedule	InR6	Wait	0.0	0.0	0.0		0 rows	☑
7	R7	Based on Schedule	InR7	Wait	0.0	0.0	0.0		0 rows	☑
8	R8	Based on Schedule	InR8	Wait	0.0	0.0	0.0		0 rows	☑
9	R9	Based on Schedule	InR9	Wait	0.0	0.0	0.0		0 rows	☑
10	R10	Based on Schedule	InR10	Wait	0.0	0.0	0.0		0 rows	☑

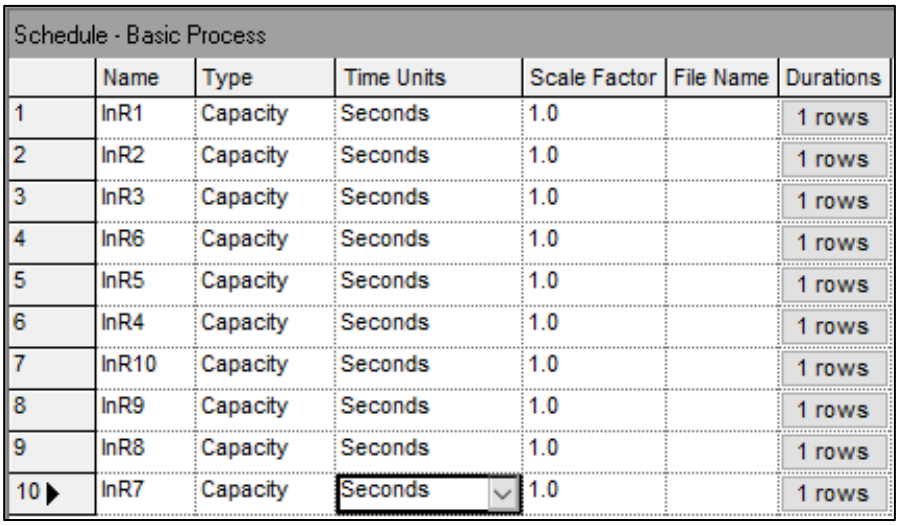
Figura 18. Módulo de datos Resource

El módulo de datos *Schedule (Figura 20)* para los diez recursos será de tipo *capacity* ya que lo que queremos programar es la capacidad del recurso. La duración será 0,01 en segundos y el valor será la variable VcapR. Esta variable es un vector de diez posiciones, se le asignará por tanto el valor de la posición uno a R1 y así sucesivamente.

Durations

	Value	Duration
1	VcapR(1)	0.01
	Double-click here to add a new row.	

Figura 19. Ajustes de Durations



Schedule - Basic Process

	Name	Type	Time Units	Scale Factor	File Name	Durations
1	InR1	Capacity	Seconds	1.0		1 rows
2	InR2	Capacity	Seconds	1.0		1 rows
3	InR3	Capacity	Seconds	1.0		1 rows
4	InR6	Capacity	Seconds	1.0		1 rows
5	InR5	Capacity	Seconds	1.0		1 rows
6	InR4	Capacity	Seconds	1.0		1 rows
7	InR10	Capacity	Seconds	1.0		1 rows
8	InR9	Capacity	Seconds	1.0		1 rows
9	InR8	Capacity	Seconds	1.0		1 rows
10	InR7	Capacity	Seconds	1.0		1 rows

Figura 20. Módulo de datos Schedule

Como se explica en el apartado anterior, diferenciamos tres estados diferentes para los servidores: inactivo, ocioso y ocupado. Cuando un servidor está apagado y se da la orden de activarlo pasa un tiempo, llamado tiempo de arranque, hasta que este puede ser ocupado. Durante este período, la capacidad real del servidor es cero, debido a que no puede ser ocupado. No obstante, consume la misma potencia que consumiría si la capacidad fuera uno y su estado fuera ocioso. Es por ello que creamos tres variables diferentes: *VcapR*, *Vcap* y *Vcap_ant*.

Las tres son vectores de diez posiciones, el valor de la primera posición va asociado al servidor *R1*, el de la segunda al servidor *R2*, y así sucesivamente para los diez servidores.

- *VcapR*: es la capacidad real del servidor, es decir, la que indica si el servidor puede ser ocupado o no.
- *Vcap*: es la variable que indicará si el servidor está ecendido o apagado, independientemente de que éste pueda ser ocupado o aún no haya transcurrido el tiempo de arranque. La utilizaremos para el cálculo de la potencia.
- *Vcap_ant*: esta variable indica la capacidad anterior del servidor. Gracias a ella se conocerá cuándo un servidor pasa de estada inactivo a activo. En este caso deberá transcurrir el tiempo de arranque para que el servidor pueda ocuparse.

En el módulo de asignación *Copia Vcap anterior (Figura 21)* se asigna el valor de Vcap a Vcap_ant. Como Vcap_ant es un vector, el tipo de la variable será *Variable Array(1D)*. En la casilla de *row* se pondrá la posición del vector que se está actualizando. Para la posición uno, respectiva al servidor R1, se asignará el valor de *Vcap(1).* Esto se repetirá para todos los servidores.

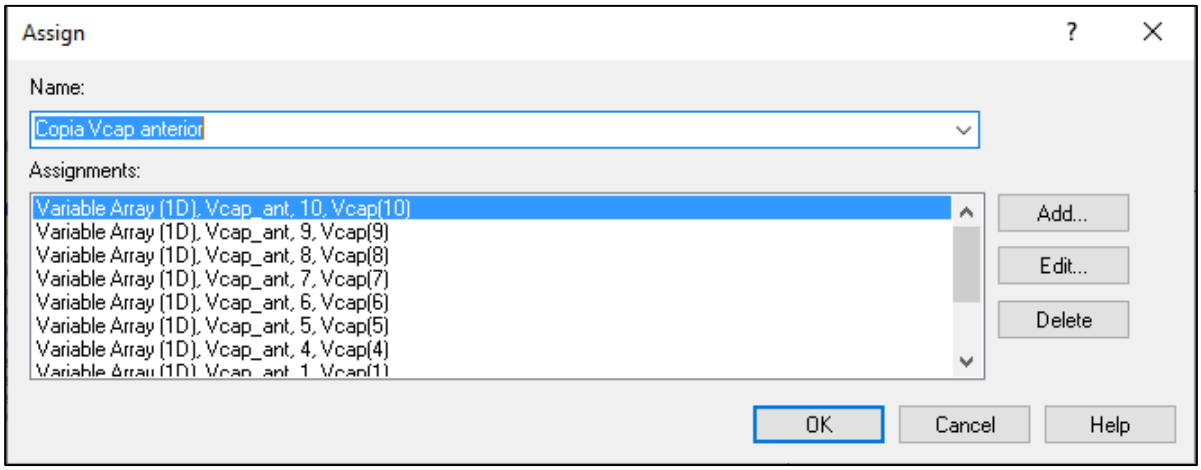


Figura 21. Ajustes del módulo de asignación de Vcap_ant

Una vez copiado el valor de *Vcap* en la variable *Vcap_ant*, podemos cambiar el valor de la capacidad. Esto se hace en el módulo de asignación *Calc Serv Dispon (*Figura 22*).* Se asignará valor uno a la primera posición del vector *Vcap* si el numero de servidores activos es mayor o igual que uno (*NumServ*>=1), a la segunda si el número de servidores activos es mayor o igual que dos y respectivamente hasta la posición diez, que tendrá valor uno cuando el número de servidores activos sea diez.

Se utiliza esta condición en el nuevo valor de la variable *Vcap*, ya que al ser booleana devolverá un uno cuando se cumpla y un cero cuando no.

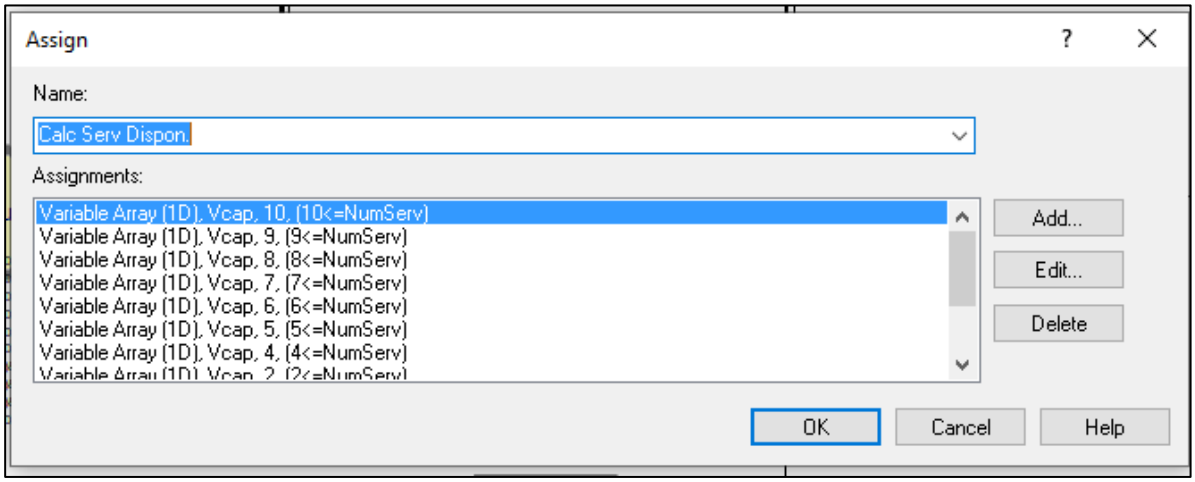


Figura 22. Ajustes del módulo de asignación Vcap

A continuación, se ecuentra el módulo de asignación *Cálculo tiempo de Vcap (Figura 23)*, donde se actualizará el valor de la variable *TVcap*. Esta variable es también un vector de diez posiciones, cada una respectiva a un servidor. Se utiliza para indicar el momento en el que debe actualizarse la capacidad real del servidor, ya que si este estaba inactivo tiene que transcurrir el tiempo de arranque antes de que el servidor pueda ser ocupado.

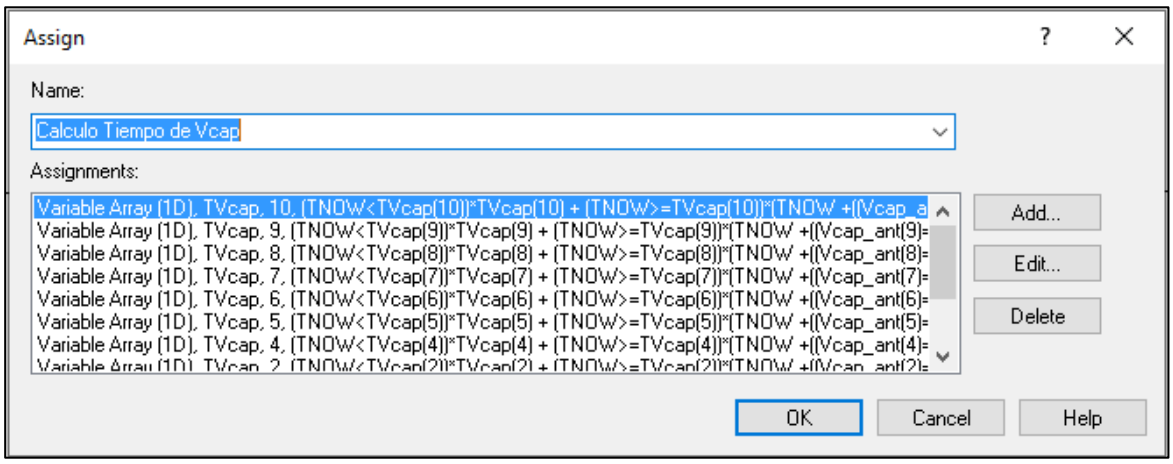


Figura 23. Ajustes del módulo de asignación de TVcap

TVcap no se actualizará si el tiempo de la simulación *TNOW* todavía es menor. Sin embargo, si *TNOW* es mayor o igual que *TVcap*, la variable tomará el valor del tiempo de la simulación más el tiempo de arranque si el servidor pasa de apagado a encendido, es decir, si *Vcap_ant* es cero y *Vcap* es uno. Si no se da esta situación, *TVcap* tomará el valor de *TNOW.*

	Type	Variable Name	Row	New Value
1	Variable Array (1D)	TVcap	10	(TNOW<TVcap(10))*TVcap(10) + (TNOW>=TVcap(10))*(TNOW
2	Variable Array (1D)	TVcap	9	(TNOW<TVcap(9))*TVcap(9) + (TNOW>=TVcap(9))*(TNOW
3	Variable Array (1D)	TVcap	8	(TNOW<TVcap(8))*TVcap(8) + (TNOW>=TVcap(8))*(TNOW
4	Variable Array (1D)	TVcap	7	(TNOW<TVcap(7))*TVcap(7) + (TNOW>=TVcap(7))*(TNOW
5	Variable Array (1D)	TVcap	6	(TNOW<TVcap(6))*TVcap(6) + (TNOW>=TVcap(6))*(TNOW
6	Variable Array (1D)	TVcap	5	(TNOW<TVcap(5))*TVcap(5) + (TNOW>=TVcap(5))*(TNOW
7	Variable Array (1D)	TVcap	4	(TNOW<TVcap(4))*TVcap(4) + (TNOW>=TVcap(4))*(TNOW
8	Variable Array (1D)	TVcap	2	(TNOW<TVcap(2))*TVcap(2) + (TNOW>=TVcap(2))*(TNOW
9	Variable Array (1D)	TVcap	3	(TNOW<TVcap(3))*TVcap(3) + (TNOW>=TVcap(3))*(TNOW
10	Variable Array (1D)	TVcap	1	(TNOW<TVcap(1))*TVcap(1) + (TNOW>=TVcap(1))*(TNOW +((Vcap_ant(1)==0)&&(Vcap(1)==1))*Tarr)

Figura 24. Asignación de TVcap

Antes de calcular la potencia de cada servidor y su temperatura, se explica la actualización de la variable *VcapR*. Esta se actualiza en el módulo de asignación *Cálculo Vcap Real (Figura 25).*

VcapR tomará el valor de *Vcap* en instante *TVcap*. Esto se calcula de la misma manera para los diez servidores. Por tanto, si se cumple que el tiempo de la simulación es mayor o igual que el tiempo *TVcap* el valor de *VcapR* será el mismo que *Vcap* . Si no se cumple, el valor de la capacidad real será cero. Esto ocurre sólo cuando está transcurriendo el tiempo de arranque ya que en cualquier otro caso el valor de *TVcap* será igual al de *TNOW*.

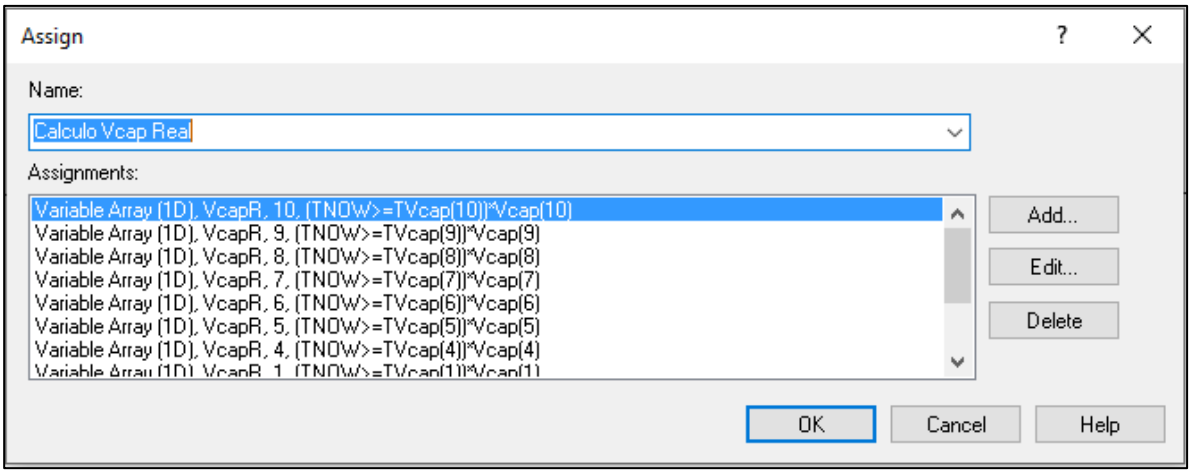


Figura 25. Ajustes del módulo de asignación VcapR

En el módulo de asignación *Asigno potencia* se calcula tanto la potencia (*Figura 26*) cómo la temperatura de cada servidor *(Figura 27*). La potencia de cada servidor dependerá de si el servidor está inactivo, ocioso u ocupado.

STATE(recurso) es una variable que indica el estado del recurso. El valor que devuelve si el servidor está ocupado es -2. Por tanto, *(STATE(recurso)==-2)* devolverá un uno si el servidor está ocupado, y un cero si no lo está. El consumo marginal del servidor si está realizando una tarea es a2. a1 sin embargo, es el consumo que hace el servidor por estar encendido. Por eso toda la expresión se multiplica por *Vcap* , que será uno si el servidor está encendido y cero si está apagado.

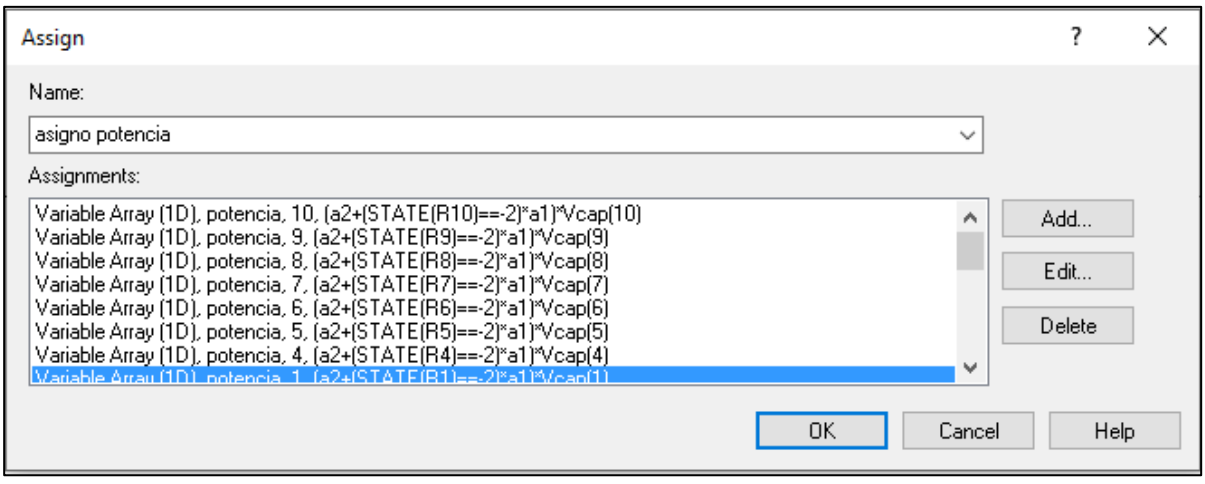


Figura 26. Asignación de la potencia

Para calcular la temperatura de cada servidor se utiliza el vector *Temp*, de diez posiciones. Creamos la variable *Tant*, que se va actualizando, para poder calcular el incremento de tiempo. Utilizamos la expresión de la temperatura anteriormente mostrada:

$$K_t \times \frac{d\theta_{ij}}{dt} = C_p \times q_a \times (\theta_c - \theta_{ij}) + p_{ij};$$

$$\theta_{ij}2 = \theta_{ij}1 + \Delta t \times \left(\frac{C_p \times q_a}{K_t} \times (\theta_c - \theta_{ij}1) + p_{ij} \times \frac{1}{K_t}\right)$$

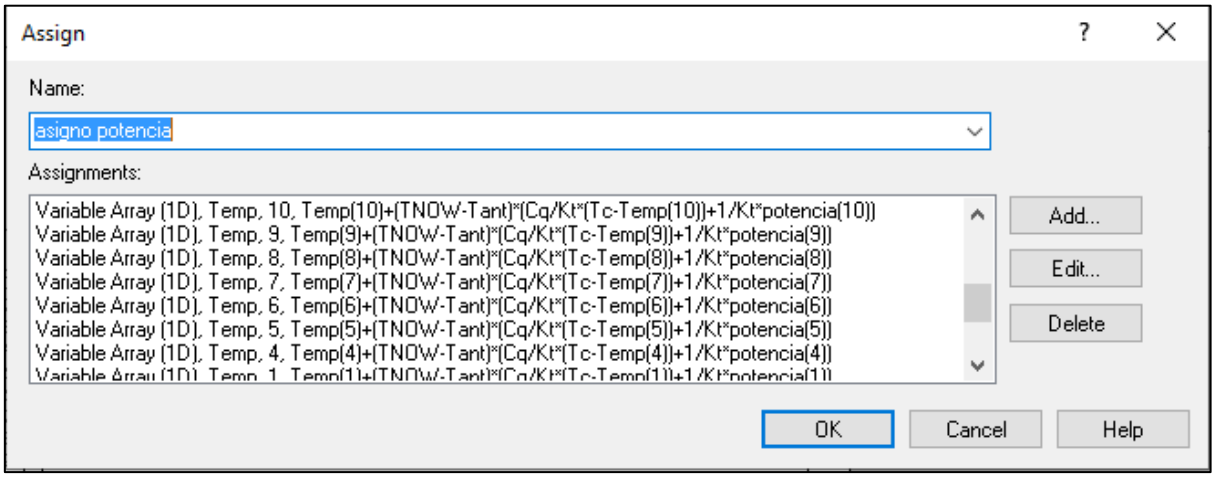


Figura 27. Asignación de la temperatura

Siendo *(TNOW-Tant)* el incremento de tiempo, Tc la temperatura de la unidad aislada en frío θ_c, K_t la capacidad térmica del servidor y Cq la multiplicación de la capacidad térmica del aire C_p y del caudal de aire q_a.

Por último, antes del bloque de *dispose* dónde salen del sistema las entidades, se encuentra el módulo del sistema de refrigeración (*Figura 28*). La temperatura de refrigeración es otra de las variables manipulables del sistema. En este módulo se asignan los valores tanto de la temperatura, como de la potencia del CRAC. La temperatura θ_c y de la potencia del CRAC se utilizan las expresiónes mostradas anteriormente:

$$\tau \times \frac{d\theta_c}{dt} = \theta_r - \theta_c;$$

$$\theta_c 2 = \theta_c 1 + \frac{\Delta t}{\tau} \times (\theta_r - \theta_c)$$

$$P_{CRAC} = \frac{P}{CoP(\theta_c(t))}$$

$$CoP(\theta) = b \times \theta^2 + c \times \theta + d$$

Assignments

	Type	Variable Name	New Value
1	Variable	COP	coef1*Tc*Tc+coef2*Tc+coef3
2	Variable	Tant	TNOW
3	Variable	Tc	Tc+(TNOW-Tant)/tau*(Tempr-Tc)
4	Variable	PCRAC	(potencia(1)+potencia(2)+potencia(3)+potencia(4)+potencia(5)+potencia(6)+potencia(7)+potencia(8)+potencia(9)+potencia(10))/COP

Figura 28. Asignación de variables del módulo de refrigeración

Donde T_c es la temperatura de la unidad aislada en frio θ_c y $Temp_r$ la temperatura de consigna del aire θ_r. Los coeficientes de la función de desempeño coef1, coef2 y coef3 son b, c y d respectivamente.

Antes de comenzar la simulación es necesario dar los valores iniciales a las variables y alos parámetros. Para ello tenemos el módulo de datos *variable (Figura 29)*. El valor se introduce en *Initial Values.* En la columna *Rows* se introduce el número de filas que tienen las variables que son vectores.

Variable - Basic Process

	Name	Comment	Rows	Columns	Data Type	Clear Option	File Name	Initial Values
1	Kt	Capacidad termica de servidor			Real	System		1 rows
2	potencia	potencia del servidor	10		Real	System		0 rows
3	Tant	Tiempo de calculo anterior			Real	System		1 rows
4	Temp	Temperatura del servidor	10		Real	System		1 rows
5	a2	consumo del servidor sin carga de trabajo			Real	System		1 rows
6	Tlleg	Tiempo medio entre llegadas			Real	System		1 rows
7	k2	media de complejidad de servicio			Real	System		1 rows
8	frec	frecuencia de trabajo			Real	System		1 rows
9	NumServ	Num servid. disponibles			Real	System		1 rows
10	Vcap	Vector de capacidades de servidores	10		Real	System		1 rows
11	VcapR	Vec. Cap Serv Reales	10		Real	System		1 rows
12	Vcap_ant	Vec Cap Serv. Anterior	10		Real	System		1 rows
13	TVcap	Vector tiempos cambio Vcap	10		Real	System		1 rows
14	Tarr	Tiempo de arranque del servidor			Real	System		1 rows
15	Cq	caudal de aire*capacidad termica			Real	System		1 rows
16	Cf				Real	System		1 rows
17	Tc	Temperatura unidad aislada enfrio			Real	System		1 rows
18	tau				Real	System		1 rows
19	Tempr	temperatura de consigna del aire			Real	System		1 rows
20	COP	funcion de desempeno de la maquina			Real	System		0 rows
21	PCRAC	potencia de CRAC			Real	System		0 rows
22	coef1	coeficiente depende maquina			Real	System		1 rows
23	coef2	coeficiente depende maquina			Real	System		1 rows
24	coef3	coeficiente depende maquina			Real	System		1 rows

Figura 29. Módulo de datos variable

La frecuencia es también una variable manipulable del sistema. De ella dependen tanto μ_j (*Figura 31*), como a1 *(Figura 30*). Utilizamos el módulo de datos avanzados *expression* para definirlas (*Figura 32*).

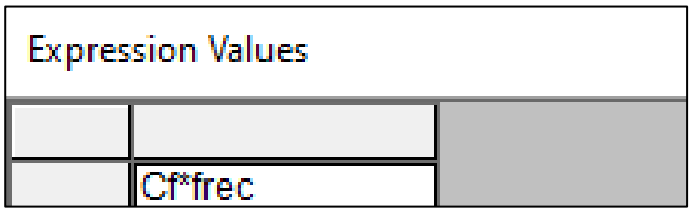


Figura 30. Expresión de a1

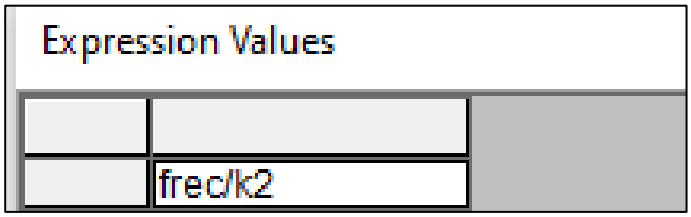


Figura 31. Expresión de mu

Expression - Advanced Process

	Name	Comment	Rows	Columns	Data Type	File Name	Expression Values
1	a1				Native		1 rows
2	mu				Native		1 rows

Figura 32. Módulo de datos Expression

Arena, por defecto, actualiza las variables de los módulos de asignación cada vez que llega una nueva entidad. Sin embargo, a nosotros nos interesa que se actualicen con respecto al tiempo. Para ello, en la barra de herramientas cambiamos las opciones *Run Setup (Figura 33)* de manera que actualice cada unidad de tiempo.

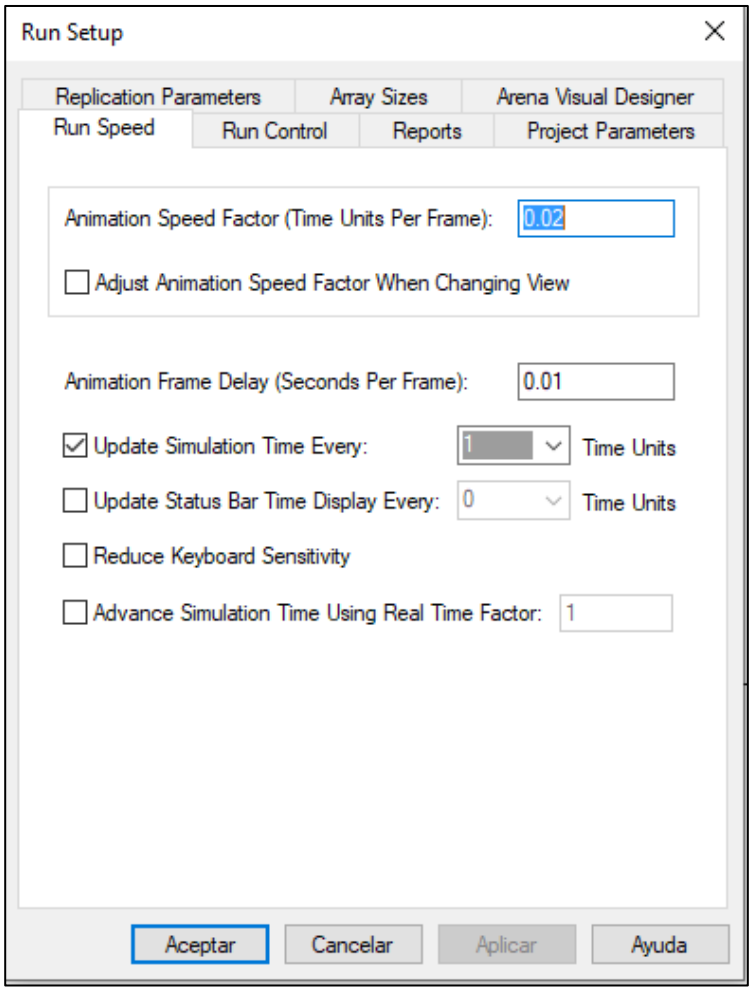


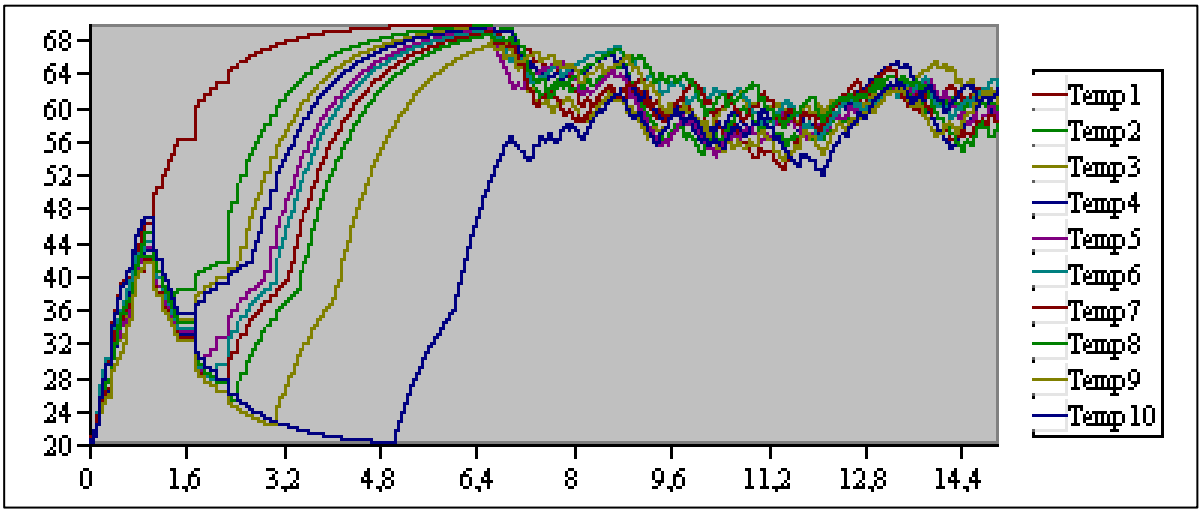
Figura 33. Setup de Simulación

Dentro de *Run Setup* se ajustan todas las condiciones de la simulación, en *Replication Parameters* podamos defenir el tiempo de simulación, o si queremos que se simule hasta una condición dada, por ejemplo el número de peticiones simuladas.

6 RESULTADOS Y CONCLUSIONES

Respuesta a la activación de servidores

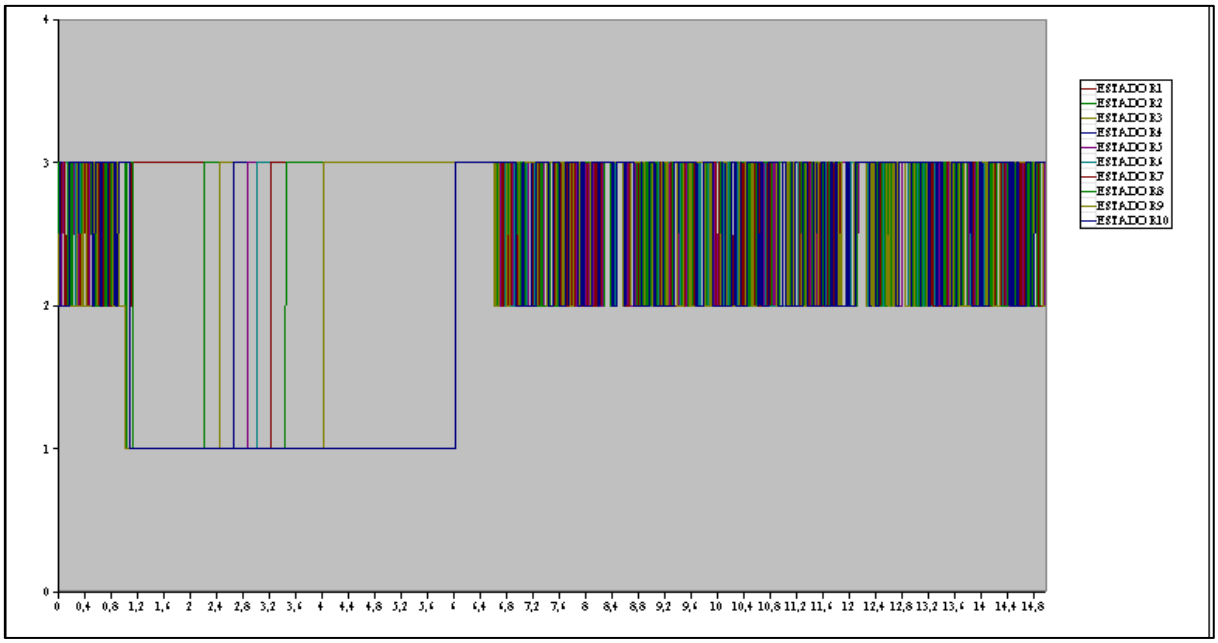
A partir del modelo realizamos una simulación para ver como afecta al sistema la activación y desactivación de los distintos servidores. De la simulación de 15 segundos se obtienen los siguientes resultados:



Gráfica 1. Temperatura servidores caso 1

Como se observa en la gráfica de temperaturas respecto al tiempo (Gráfica 1), hasta el segundo 1 todas las temperaturas van incrementando. Llegados a este punto se apagan todos los servidores menos el primero. Es por eso que la única temperatura que sigue aumentando es la del primer servidor. Las demás gracias al sistema de refrigeración van disminuyendo hasta que se de la condición que active los servidores. El servidor 10 es el último que vuelve a encenderse, este consigue volver a alcanzar la temperatura inicial.

Se cumple la restricción de temperatura máxima de servicio, que como se menciona anteriormente es $\theta_c + 50$ ºC, 70ºC en nuestro modelo donde $\theta_c = 20ºC$. A lo largo de la simulación el tiempo de la unidad aislada en frío se mantiene en el valor estipulado ($15 \leq \theta_c \leq 25\ ºC$).

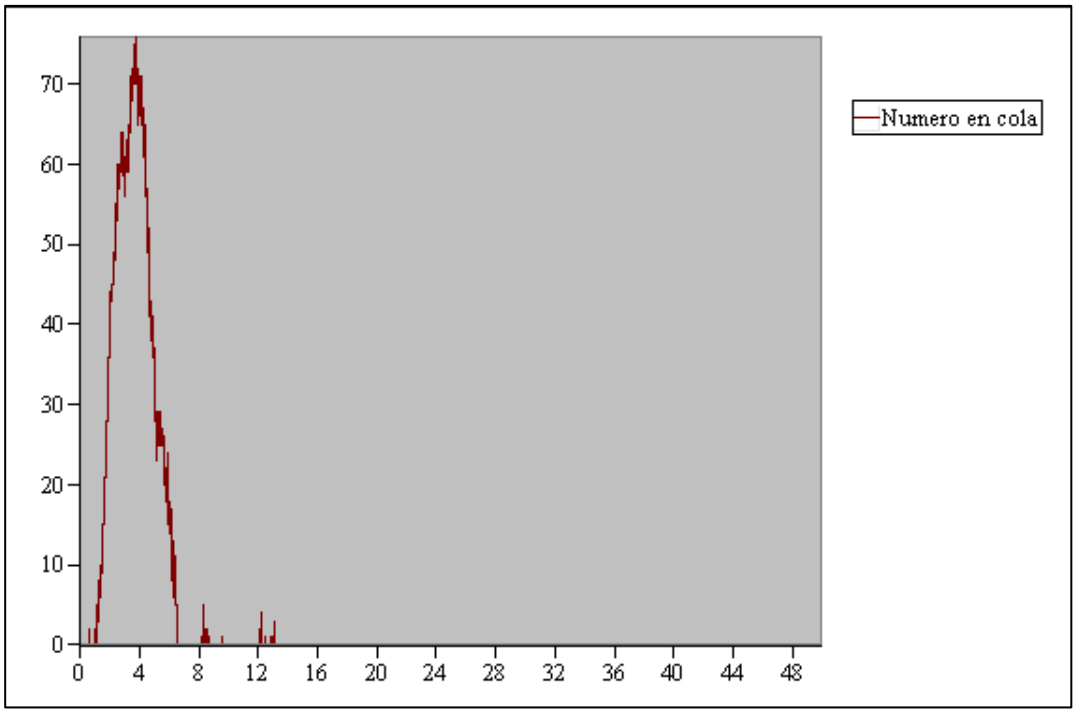


Gráfica 2. Estados servidores caso 1

En esta gráfica (Gráfica 2) se muestran los estados de los servidores siendo inactivo, ocioso y ocupado 1, 2 y 3 respectivamente. Se observa que hasta el segundo 1 los estados de los diez servidores oscilan entre ocioso y ocupado. Esto se debe a que el número mínimo de servidores necesarios para abastecer la cola es 7 y hasta ahora se dispone de 10 servidores activos, por tanto algunos servidores podrán estar ociosos. Sin embargo, en el momento que se apagan los demás servidores y hasta que vuelven a estar activos los diez, se trabaja a máxima potencia y no dejan de estar ocupados.

Podemos ver el efecto del tiempo de arranque en ambas gráficas. En la de las temperaturas se ve como la temperatura de cada servidor comienza a incrementar cuando este se enciende y sufre un crecimiento mas rápido pasado un segundo de su activación. Esto se debe a que durante ese segundo la única potencia que se consume es la potencia sin carga, una vez ha trascurrido el tiempo de arranque y el servidor admite la carga la potencia que consume es el doble. En la gráfica del estado de los servidores se puede ver que permanecen "inactivos" hasta que trascurre el segundo de arranque.

La condición para que no se de una cola infinata se cumple con un mínimo de 7 servidores activos. Esta restricción, sin embargo, puede violarse en determinados períodos de tiempo. Estos períodos de tiempo no pueden ser demasiado grandes, ya que Arena da un error cuando ve que los recursos no pueden abastecer la cola que se ha formado.



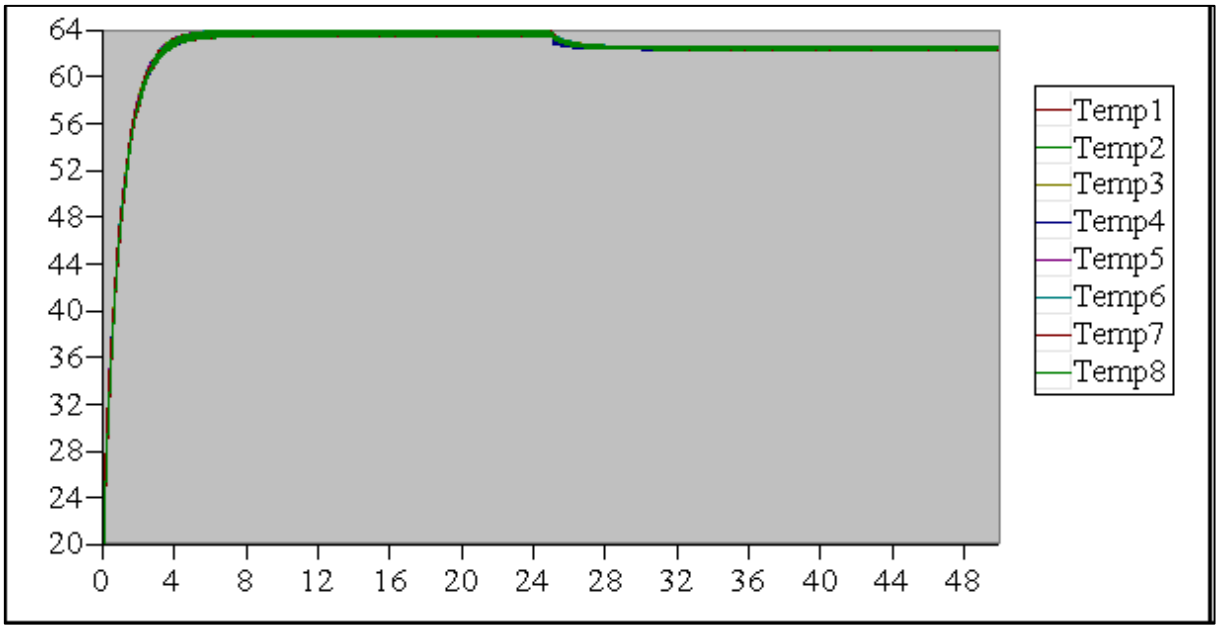
Gráfica 3. Número de peticiones en cola caso1

Se muestra la evolución del número de peticiones en cola respecto al tiempo (*Gráfica 3*). Se ve como hasta el segundo 1 que están los 10 servidores activos el número de peticiones en cola es nulo. Aunque el número de servidores vaya aumentando también lo hace el número de peticiones en cola, esto se debe a que hasta que no haya un mínimo de 7 servidores activos sigarán llegando nuevas peticiones antes de que se procesen las que están en el sistema. Si esta situación se hubiera prolongado en el tiempo, Arena hubiera impedido que siguiera corriendo la simulación, ya que el número de entidades en la cola superaría el límite. Una vez están activos mas de 7 servidores el número de peticiones en cola comienza a bajar hasta llegar de nuevo a cero. Como las llegadas y el procesamiento dependen de una función de probabilidad habrá momentos en los que lleguen mas peticiones que las que se pueden procesar, por eso aparecen picos aún teniendo 10 servidores encendidos.

Respuesta al cambio de frecuencia

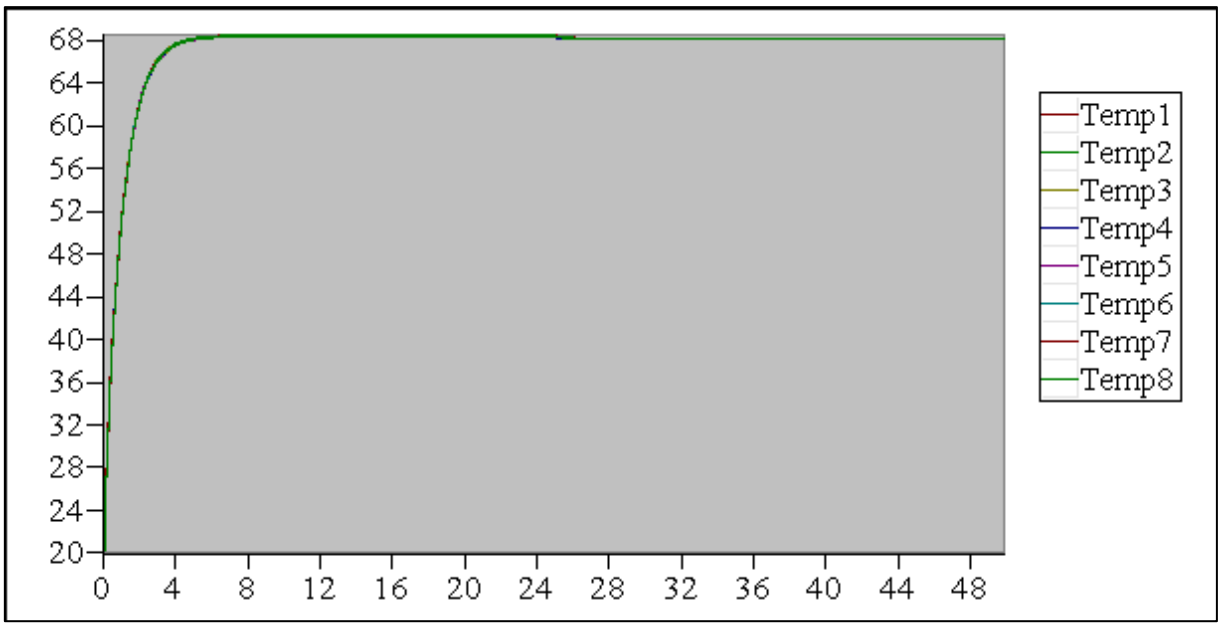
Resulta interesante ver como afecta al sistema un cambio en la frecuencia. Tanto el tiempo de llegada de las peticiones como el tiempo de servicio siguen una distribución exponencial de probabilidad. Sin embargo, se asumen constantes para evitar el ruido y obtener unos resultados más claros.

En el segundo 25 de la simulación se disminuye un 20% la frecuencia de trabajo. Consideramos el número de servidores activos constante e igual a 8.



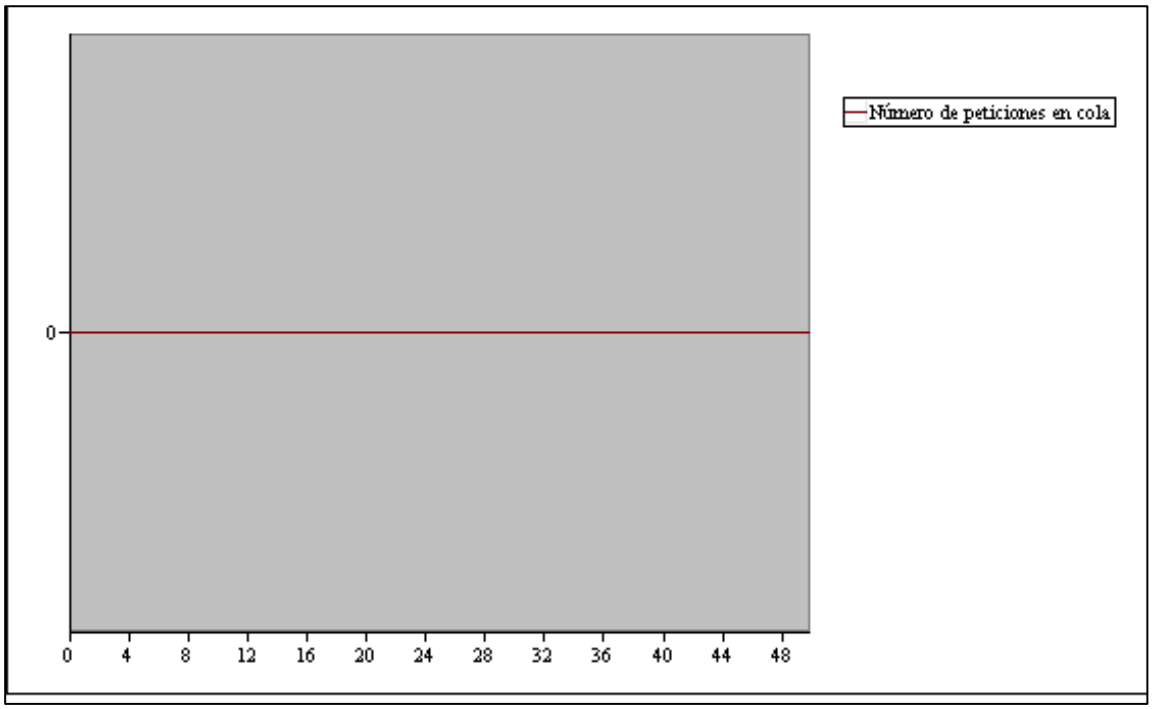
Gráfica 4. Temperatura servidores caso 2

Se observa (*Gráfica 4*) como la temperatura alcanza la estabilad hasta que se cambia la frecuencia, en ese momento la temperatura disminuye. Esto se debe a que la frecuencia influye en el consumo marginal, por tanto si disminuye la frecuencia, disminuye el consumo marginal y a su vez la temperatura. Este cambio influirá máso menos dependiendo del valor de C_f y del consumo sin carga. Si por ejemplo tomamos $C_f = 10$ y $\alpha 2 = 70$ el cambio de frecuencia no afectaría a penas a la temperatura (*Gráfica 5*) ya que la aportación que depende de la frecuencia es mucho menor.



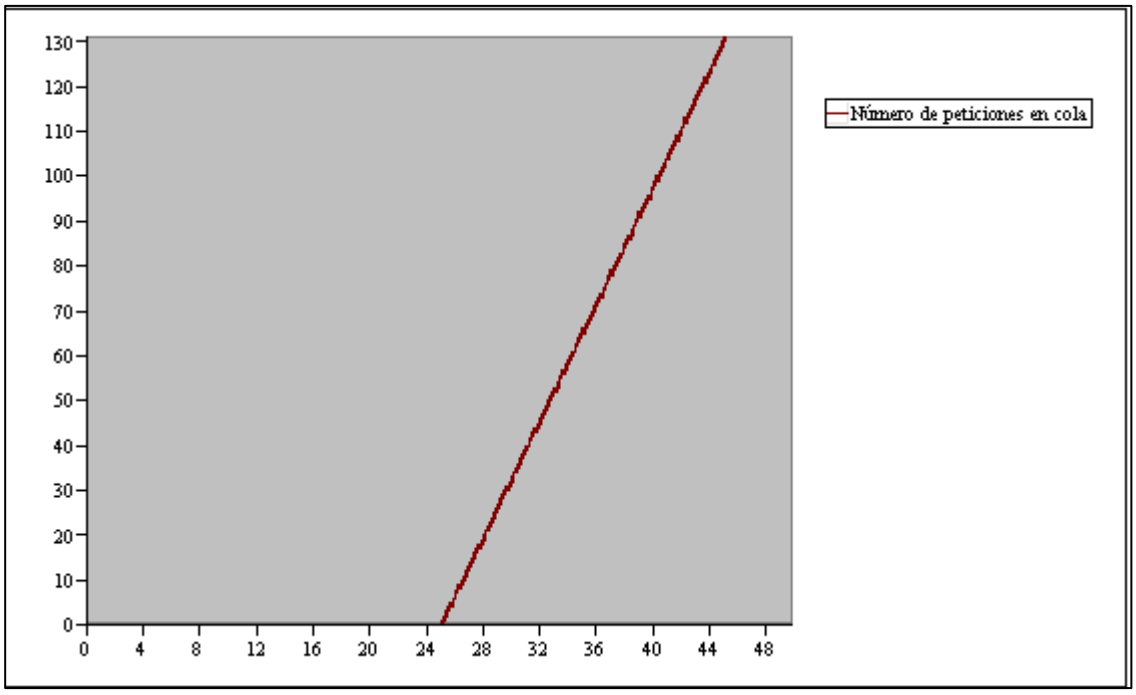
Gráfica 5.Temperatura servidores caso 3

En los resultados de la simulación también se observa que a frecuencia=1 los servidores están ocupados entorno a un 78% del tiempo, de diferente manera, a frecuencia=0,8 los servidores están alrededor del 97% del tiempo trabajando. En ambos casos se respeta el límite de temperatura. El número de peticiones en la cola es cero durante toda la simulación (*Gráfica 6*), en el caso de haber cola los servidores estarían ocupados el 100% del tiempo, y aún asi tendrían que seguir llegando peticiones.



Gráfica 6. Número depeticiones en cola caso2

Estudiamos este mismo caso pero con 7 servidores, que sería el número mínimo de servidores para frecuencia 1, en vez de 8. Se observa como desde el segundo que disminuimos la frecuencia la cola comienza a crecer indifinidamente (*Gráfica 7*), Arena para la simulación antes de acabar los 50 segundos porque la considera una cola infinita.



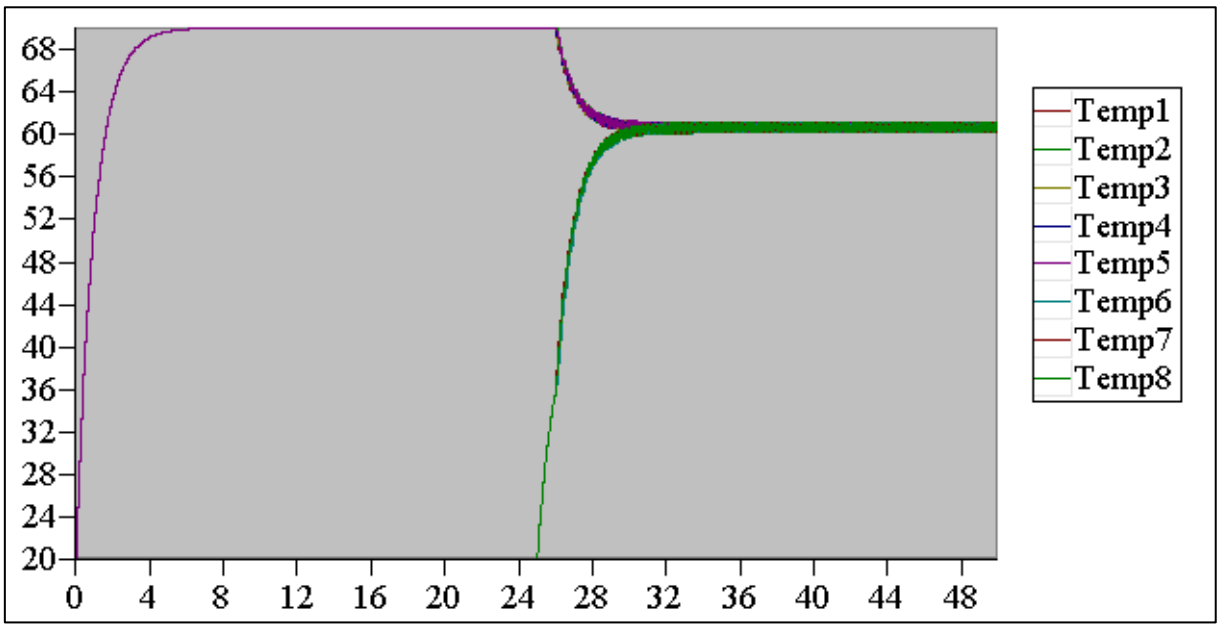
Gráfica 7. Número de peticiones en cola caso4

Si quisieramos reducir la frecuencia más, habría que activar más servidores ya que la cola se colapsaría con tan solo 8 activos.

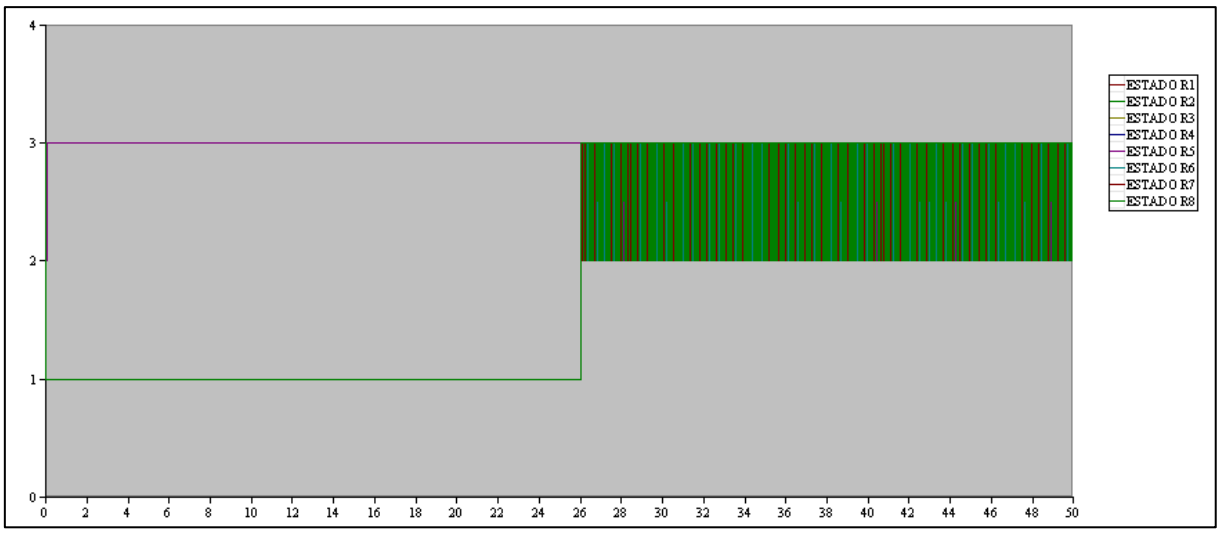
Respuesta al cambio en escalón del número de servidores activos

Esta vez se quiere estudiar la respuesta al cambio en escalón del número de servidores activos. Para ello utilizaremos tiempos de llegada y servicio constantes, como anteriormente se explicó, para evitar ruidos que pueden causar los cambios de la demanda y del tiempo de procesado. El tiempo medio de servicio que consideraremos será 0,1 y el de llegadas 0,02. El número mínimo de servidores activos es 5 exactamente. Se cambia de 5 a 8 servidores en el segundo 25 de la simulación.

Se ve como hata el segundo 25 los servidores no dejan de estar ocupados (*Gráfica 9*), trabajando a máxima potencia todo el rato y llegando a la temperatura máxima (se comprueba que no se supera) (*Gráfica 8*). En el momento en el que el número de servidores aumenta la temperatura comienza a disminuir hasta que alcanzala estabilad de nuevo. El estado de los servidores pasa a oscilar de ocioso a ocupado, ya no es necesario que trabajen todo el rato a máxima potencia.

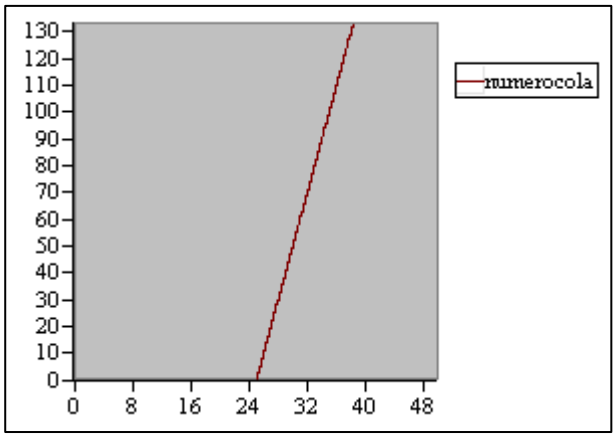


Gráfica 8. Temperatura de servidores caso5

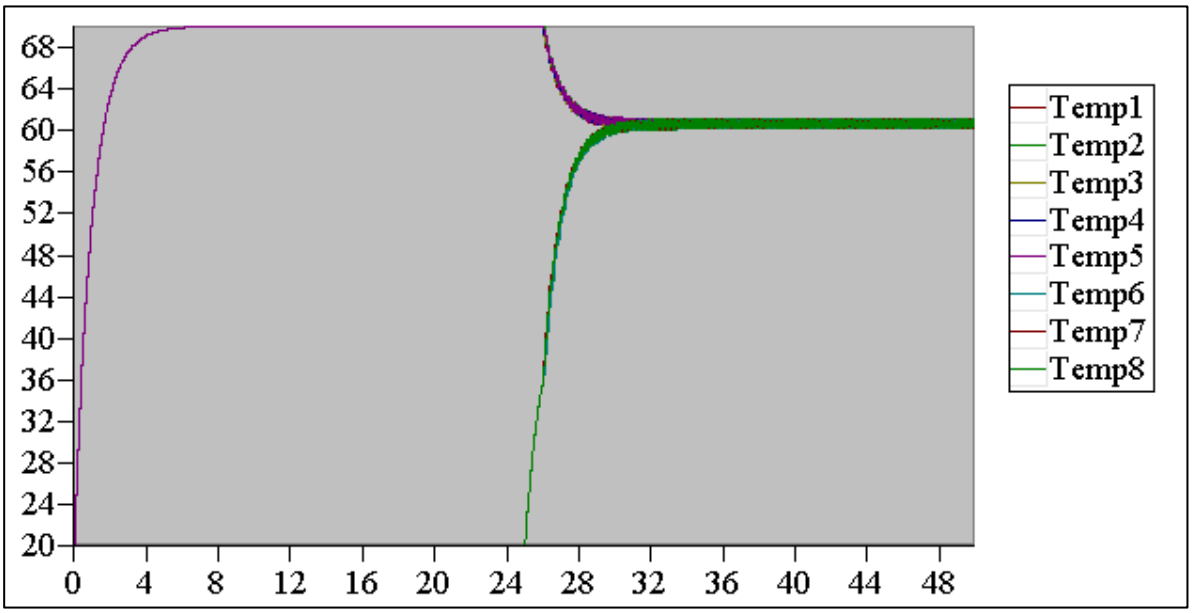


Gráfica 9. Estado de servidores caso5

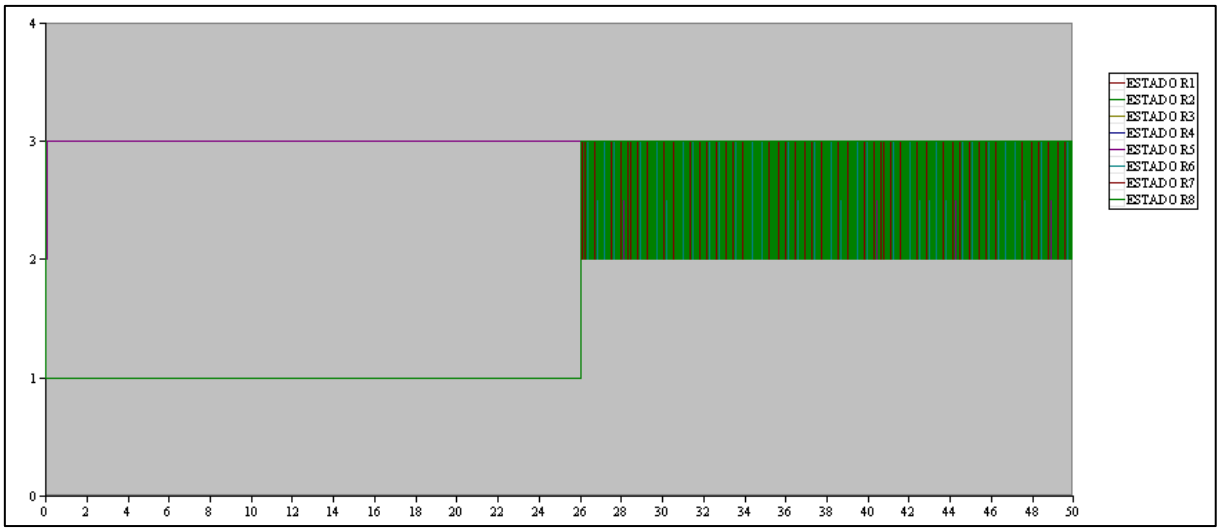
El número de elementos en la cola es siempre nulo ya que 5 servidores son suficientes para abastecer las peticiones. Sin embargo, si pasaramos de tener 8 servidores activos a 4, observamos como la cola comienza a crecer hasta no dar abasto (*Gráfica 10*). La temperatura de los servidores que se desactivan disminuiría hasta alcanzar la temperatura inicial, los 4 servidores que siguen activos, incrementarían su temperatura al estar trabajando a máxima potencia (*Gráfica 11*) y los servidores pasarían a estar ocupados todo el rato (*Gráfica 12*).



Gráfica 10. Número de peticiones en cola caso 6



Gráfica 11. Temperatura de servidores caso 6



Gráfica 12. Estado de servidores caso 6

Conclusiones

A partir del modelo en Arena® Simulation, se han obtenido unos resultados que nos permiten sacar algunas conclusiones. Se concluye que un aumento en el número de servidores activos disminuye la temperatura de cada servidor, los cuales trabajan con menos carga. El incremento en el número de servidores también ayuda a disminuir el número de peticiones en cola, lo que mejoraría la calidad de servicio QoS, sin embargo, tambiénpodría afectar a un aumento en la potencia total que se consume. Por otro lado, la disminución de frecuencia afecta tanto al consumo marginal de cada servidor como al tiempo de servicio de los servidores, la temperatura de cada servidor disminuye, pero el número de peticiones en cola aumenta, empeorando la calidad de servicio.

Estos son solo los resultados de algunos de los múltiples casos que se pueden estudiar con este modelo. El resultado más interesante de este proyecto es el diseño e implementación del modelo que podrá ser utilizado para simular diferentes estrategias de optimización del funcionamiento de un centro de datos.

ÍNDICE DE FIGURAS

REFERENCIAS

[1] B. Johnson. [En línea]. Available: https://computer.howstuffworks.com/data-centers1.htm.

[2] A. Simulation, «Arena Simulation,» [En línea]. Available: https://www.arenasimulation.com/.

[3] D. J. B. a. C. Reams, «Toward energy-efficient computing,,» *Queue,* p. 30:30–30:43, Febrero 2010.

[4] Y. S. M. I. a. R. F. Miyuru Dayarathna, «Data Center Energy Consumption,» *IEEE COMMUNICATIONS SURVEYS & TUTORIALS,* 2016.

[5] «Dynamic thermal and IT resource management strategies for data center energy minimization,» *Journal of Cloud Computing:Advances, Systems and Applications,* 2017.

[6] C. A. G. P. a. L. Heath T, «Temperature Emulation and Management from server Systems,» *International Conference on Architectural Support for Programming Languages and Operating Systems,* pp. 106-116, 2006.

[7] G. S. V. G. Tang Q, «Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach.,» *IEEE Trans Parallel and distributed systems,* p. 1458–1472, 2008.

[8] T. L. N. X. V. V. A. M. X. Zhanga*, «Cooling Energy Consumption Investigation of Data Center IT Room with Vertical Placed Server,» *Energy Procedia,* 2017.

[9] P. S. M. A. T. d. A. Thiago L. Vasques, «Energy efficiency insight into small and medium data centres: A comparative analysis based on a survey».*ECEEE SUMMER STUDY PROCEEDINGS.*