UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE INGENIERIA

METODOLOGIA PARA LA PLANEACION DEL CRECIMIENTO DEL CENTRO DE PROCESAMIENTO DE DATOS APLICADO A UN CASO PRACTICO

TESIS PROFESIONAL

QUE PARA OBTENER EL TITULO DE
INGENIERO EN COMPUTACION
PRESENTAN:

SALVADOR MORENO VAREA
JOAQUIN PEREZ VILLEGAS
JULIO RAYMUNDO AMBROSIO
FRANCISCO ALEJO RIOS GASCON
JOSE VAZQUEZ ORTEGA

ASESOR DE TESIS: M. EN I. LAURO SANTIAGO CRUZ

MEXICO, D. F. 1992

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

INTRODUCCIÓN

El reto planteado por el advenimiento del siglo XXI y la férrea competencia que se ha manifestado en mercados cada vez más exigentes, implican la necesidad de alcanzar niveles de calidad y productividad superiores. Es por esto que en los últimos años, tanto en empresas de participación estatal como privadas, se ha observado un creciente interés en el campo de los procesos tecnológicos y de negocios. Este fenómeno implica un crecimiento en las tendencias de inversión realizadas. Estas inversiones deben alimentar las mejoras en la empresa de manera contínua, con lo cuál se pueden establecer mecanismos precisos de autocontrol sin perderse en acciones innecesarias, acudiendo a los puntos neurálgicos del negocio, y edificando la capacidad de competir que el mismo exige.

Una de las inversiones más importantes en el campo de los procesos tecnológicos y de negocios, está orientada a la adquisición de equipos de cómputo, con la finalidad de contar con la tecnología que les permita incrementar los niveles de servicio, elaborando sistemas que lleven a la acción y no sólo al registro histórico de información.

Generalmente una buena inversión depende en gran medida de la correcta planeación con que ésta haya sido realizada, por lo cual es conveniente tomar en consideración aspectos como los siguientes:

¿Cuáles son las necesidades de la empresa?

¿Cuál es el crecimiento esperado a corto, mediano y largo plazo?

¿Qué tipo de equipo de cómputo nos conviene adquirir?

¿Cuáles son los efectos en el esquema operativo de la empresa?

¿Qué tipo de aplicaciones vamos a correr en el equipo?

De esta manera se tiene una panorámica general de las expectativas que se desea cubrir, y por consecuencia, la toma de decisiones es más sencilla.

Sin embargo, cuando se invierte en el crecimiento de los recursos de un sistema de cómputo, se ha observado que la mayoría de las empresas no cuentan con un equipo técnico lo suficientemente capacitado para llevar a cabo toda una serie de estudios de estimación y planeación del crecimiento de los recursos en forma satisfactoria. Por ende las adquisiciones realizadas, en la mayoría de los casos, apenas satisfacen las necesidades actuales y como una consecuencia directa, en el futuro próximo se deberá realizar una nueva inversión para cubrir los nuevos requerimientos.

Por estas razones, la tesis del presente trabajo es ofrecer una metodología para lograr una correcta planeación de la capacidad de los recursos del sistema de procesamiento de datos, que permita entender, predecir y definir las alternativas para llegar a una toma de decisiones adecuada.

Se presenta la metodología que permite estructurar una serie de análisis necesarios para la planeación de la capacidad del equipo que es usado en el centro de cómputo. Como complemento de las ideas plasmadas en este trabajo se desarrolla un caso en particular con la finalidad de mostrar el uso y conveniencia de ésta metodología en su aplicación práctica sobre problemas reales.

CAPÍTULO 1
ANTECEDENTES

A medida que el procesamiento de datos es una parte importante en una empresa, su dependencia de él es mayor. Así, el departamento de procesamiento de datos se convierte en una renglón estratégico y la planeación de la capacidad se transforma en un punto de la más alta prioridad.

La capacidad de un centro de cómputo no es más que su poder de procesamiento, tener capacidad excesiva resulta generalmente costoso, tener muy poca capacidad puede ser desastroso. La planeación de la capacidad busca el óptimo balance entre ambas situaciones, esto con base en las demandas y acuerdos de servicio que se tengan con el usuario.

Se puede definir a la planeación de la capacidad como un procedimiento desarrollado que provee un enfoque ordenado para el entendimiento y predicción de la capacidad de los sistemas de procesamiento de datos. Se trata de una metodología que integra los conceptos de la administración del rendimiento con los de la moderna tecnología de medición. Por medio de este procedimiento es posible visualizar los problemas involucrados en la administración de los recursos computacionales, ayudando, entre otras cosas, a resolver preguntas como las siguientes:

¿Qué parámetros deberán recolectarse para caracterizar la carga de trabajo?

¿Qué parámetros deberán recolectarse para caracterizar los componentes de *hardware* y *software*?

¿Qué parámetros son requeridos para pronosticar las cargas de trabajo y el rendimiento del sistema a futuro?

¿Qué herramientas se requieren para recolectar, analizar y reportar los datos descritos anteriormente?

¿Cuáles son los parámetros para la elección de la mejor alternativa?

A través de la planeación de la capacidad, son monitoreados y estudiados, la carga, utilización, y respuesta de los recursos, controlando el flujo de trabajo actual y futuro, para mantener o mejorar el servicio.

Actividades de la Planeación de la Capacidad

Las etapas fundamentales para lograr una planeación óptima son:

a) Niveles de servicio y requerimientos.

En estos niveles se establecen todos los acuerdos entre el usuario y el departamento de proceso de datos.

En los requerimientos se contemplan las necesidades actuales, a mediano y largo plazo por parte de los usuarios.

b) Análisis de rendimiento

Con las mediciones y parámetros clave del rendimiento, así como históricos y estándares del sistema; se evalúa la situación actual comparándola con los acuerdos de nivel de servicio que se pactaron con el usuario.

c) Modelado

Con la ayuda de técnicas de modelado, se proyecta, y evalúa la capacidad del sistema. Obteniendo diversas alternativas de solución.

d) Toma de decisiones

Corresponde a los ejecutivos el decidir sobre alguna de las alternativas propuestas por el equipo de planeación de la capacidad.

Finalmente, diremos que la planeación de la capacidad dentro de la organización debe ser reconocida y apoyada ampliamente para ser efectiva.

A continuación mencionaremos algunos aspectos importantes involucrados en el proceso de planeación de la capacidad.

El sistema de computación

La metodología de la planeación de la capacidad fue desarrollada para su implantación en un sistema de computación como el descrito en la figura 1. Los subsistemas básicos de *hardware* son : La unidad central de procesamiento (UCP), la memoria principal, los subsistemas de Entrada/Salida que incluyen: controladores de comunicaciones locales, unidades de control, manejadores de discos, manejadores de cintas, impresoras y equipo de telecomunicaciones.

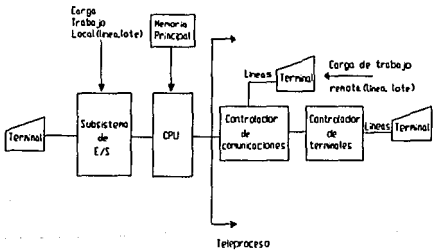


Fig.1 Sistema de Computación

La capacidad de un sistema de computación

Para propósitos de planeación, una instalación debe ser vista como un sistema de recursos, idealmente su análisis deberá abarcar la capacidad total del sistema en conjunto.

Es necesario entonces, diferenciar los factores que afectan a la capacidad de un recurso de aquellos que afectan a la capacidad global de todo el sistema.

Dentro de los factores que afectan a la capacidad del recurso tenemos:

Porcentaje de ocupación

Tamaño promedio de colas

Características técnicas, etc.

Dentro de los factores que afectan la capacidad global del sistema tenemos:

Disponibilidad: Horario disponible del *hardware* y *software*, horarios para mantenimiento, interrupciones, etc.

Requerimientos de servicio del usuario: Tiempo de respuesta, hora de inicio y terminación de *job*, servicio extraordinario, prioridades, etc.

Cargas de trabajo.

Disponibilidad

Comprender la disponibilidad de un sistema de cómputo es un aspecto muy importante de planeación de la capacidad. Por ello son tres las áreas de disponibilidad vigiladas, principalmente:

1) *Hardware*

2) *Software*

3) Las percibidas por el usuario

Esta última se refiere a la posible eventualidad de que una parte del sistema no se encuentre disponible a pesar de que el *software* y *hardware* lo estén.

Organización de la Información

La figura 2 nos muestra diferentes herramientas para medir el rendimiento y proveer información a:

Nivel directivo

Nivel corporativo

Nivel usuario

Gerencia de proceso de datos

Departamento de planeación de la capacidad

Producción

Soporte técnico

Programación y desarrollo de aplicaciones

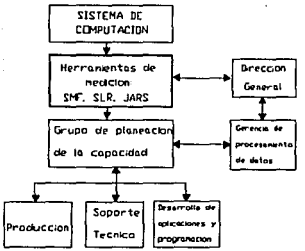


Fig.2 Herramientas de Medición SLR,SMF,JARS

Cada área tiene requerimientos únicos, pero existe un intercambio de información. El reporte apropiado de información de las operaciones nos proporciona dos cosas:

Primero: Una mejor comprensión del comportamiento de las cargas de trabajo con respecto a los usuarios en operación y la proporción en que están siendo utilizados los recursos.

Segundo: Valida la medición de los datos y la visión que el grupo de planeación de la capacidad tiene sobre la eficiencia del sistema (disponibilidad, utilización de recursos, cargas de trabajo, tendencias, entre otros)

Como parte del proceso de la planeación de la capacidad es muy importante que el administrador del sistema reciba información del rendimiento de éste, para validar la efectividad cuando se realicen cambios en los parámetros del sistema, pues una selección incorrecta puede crear los llamados *cuellos de botella* comunes en estos casos, de ahí la importancia de tener reportes continuos.

La información reportada a nivel directivo deberá ser clara y concisa, presentando los factores más impactantes sobre el rendimiento del sistema. Es común observar que los niveles directivos de una empresa son las últimos en enterarse cuando el sistema de cómputo ya no tiene la suficiente capacidad para atender a todas las demandas de procesamiento. Una de las razones para mantener a la dirección continuamente informada sobre los acontecimientos, es la de contar con la aprobación para asignar nuevos recursos al sistema en el momento oportuno, evitando con esto, decisiones apresuradas o soluciones a medias.

Como iniciar la planeación de la capacidad en la organización

En principio, se debe seleccionar el personal adecuado con la experiencia técnica y administrativa necesaria para mantener una estrecha relación entre las diversas áreas que requieren del servicio.

No existen reglas para determinar el número de personas que deban iniciar el proyecto. En algunos casos se establece un nuevo departamento llamado *CAPACITY PLANNING*.

Es conveniente entonces recurrir a las siguientes áreas de la organización para lograr un grupo homogéneo:

El departamento de producción

El departamento de programación y desarrollo de sistemas

El departamento de soporte técnico

A su vez éstas aportan información básica y demandan el servicio.

A continuación se deben establecer los siguientes puntos:

Obtener el reconocimiento de la dirección.

Detectar las áreas de oportunidad dentro de la organización y el impacto en éstas del desarrollo de la planeación de la capacidad.

Establecer con detalle los objetivos y funciones del departamento.

Iniciar los análisis previos sobre rendimiento.

Definir los parámetros para cuantificar las cargas de trabajo por aplicación

Definir una guía de procedimientos para el cumplimiento continuo de los lineamientos establecidos.

Definir el tipo de reportes y formatos necesarios para la implantación.

Investigar las herramientas de medición y modelado con las que cuente la organización y las existentes en el mercado.

CAPÍTULO 2
HERRAMIENTAS DE APOYO Y MODELADO

HERRAMIENTAS DE APOYO

Uno de los elementos más importantes dentro del proceso de la planeación de la capacidad lo constituye la información que en forma estadística nos permita conocer el comportamiento del sistema.

Es por esto que la selección de la o las herramientas, por medio de las cuáles obtenemos esta información, es igualmente importante. Estas pueden ser divididas en dos categorías básicas:

a) Las que pueden ser usadas para recolectar la información de los sistemas.

b) Las que pueden ser utilizadas para estimar el comportamiento de los recursos con base en las cargas de trabajo.

Como una manera de apoyar la comprensión de la metodología propuesta, así como de su aplicación práctica, consideramos oportuno en este momento presentar una descripción de algunas herramientas utilizadas para recolectar y explotar la información del rendimiento de los sistemas de cómputo.

Las herramientas utilizadas para la explotación de la información encontramos que pueden ser divididas en dos grupos:

a) Las programables - como el SAS (Statistics Analysis System: Sistema de análisis de estadísticas)

b) Las paramétricas y programables - como el SLR (Service Level Report: Reporte de niveles de servicio), el JARS (Job Accounting Report System : Sistema reporteador del histórico de trabajos) y el SMF (System Management Facilities: Facilidades para la medición del sistema)

Cabe hacer notar que las herramientas descritas en este capítulo son aplicables a sistemas IBM, debido a que el caso práctico del capítulo 4 fue desarrollado para equipos de esta marca, no obstante existen herramientas con funciones similares a las aquí descritas para la mayoría de los equipos de cómputo existentes en el mercado.

2.1 Lenguaje SAS (Statistics Analysis System)

El SAS nos permite procesar una base datos así como resumir y obtener reportes. Es un sistema de análisis estadístico, cuyo objetivo es el proveer los elementos para el estudio de un sistema conociendo sus necesidades de cómputo.

Una de las características principales de SAS es que se pueden agregar módulos para gráficos, proyectar datos de entrada y elaborar interfaces a otras bases de datos. El SAS puede correr en la IBM 370/30xx/43xx y máquinas compatibles; en DEC (Digital Equipment Corporation), series VAX 11/7xx cuyo sistema operativo es VMS; en Data General ECLIPSE series MV corriendo bajo el sistema operativo AOS/VS; en Prime series 50 con sistema operativo PRIMOS; así como en IBM PC AT/370 y XT/370.

El software básico de SAS nos provee facilidades para:

El almacenamiento y recuperación de información

La programación y modificación de datos

Los reportes escritos

El análisis estadístico

El manejo de archivos

Almacenamiento y recuperación de información

El sistema SAS lee datos de tarjetas, discos o cintas, organizándolos posteriormente en archivos. Los datos pueden ser analizados estadísticamente y usados para producir reportes.

Programación y modificación de datos

Un conjunto completo de instrucciones y funciones están disponibles para la modificación y manipulación de los datos. Algunas instrucciones programadas realizan operaciones normales, tales como crear nuevas variables, acumular totales y verificar errores; otras son herramientas más poderosas como las instrucciones DO/END IF/THEN/ELSE.

Reportes escritos

Así como los datos son leídos en cualquier formato lo son también escritos. Adicionalmente a los reportes preformateados en SAS, llevados a cabo por procedimientos ya establecidos, el usuario puede diseñar y producir los reportes impresos en cualquier forma o bien en archivos de salida.

Análisis Estadístico

En SAS existen procedimientos para análisis estadístico, los cuales pueden ser llevado a cabo con sólo mencionar la función, tal es el caso de la frecuencia, media, correlaciones, regresión lineal, entre otras.

Manejo de Archivos

Para el análisis de datos es comunmente necesaria la combinación de valores y observaciones (datos asociados a una entidad simple - un individuo, un registro, un año, una región geográfica - hacen una observación) de varios archivos. SAS cuenta con herramientas para editar, concatenar, mezclar y actualizar archivos. Pueden ser procesados múltiples archivos de entrada en forma simultánea y varios reportes pueden ser producidos con una sola pasada de datos.

2.2 Paquetes Estadísticos: SMF, SLR Y JARS

SMF (System Management Facilities)

El SMF es un grupo de rutinas estandard del sistema operativo MVS (Multiple Virtual Storage) que colecciona y registra información acerca de la configuración, actividad de paginación, cargas de trabajo, tiempos de la Unidad Central de Procesamiento (UCP), actividad de las salidas, los procesos, y sesiones de usuarios. Esta información nos auxilia en la producción de reportes para: análisis en el volumen de transacciones, análisis de cargas de trabajo y monitoreo de uso del sistema.

El SMF permite instalar rutinas para el control de eventos específicos, tal como, definir el tiempo máximo de proceso que permite cumplir con los estándares de la instalación.

La figura número 3 muestra el sistema de registros del SMF. Su uso varía dependiendo de la instalación y de acuerdo a:

a) Parámetros seleccionados

b) Las rutinas agregadas

c) La rutina de análisis del reporte usada para resumir la información

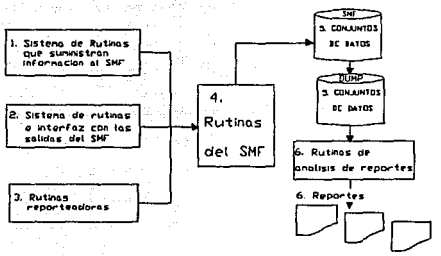


Fig. 3 Sistemas de registro del SMF

Su uso depende también de la mayoría de los componentes del sistema operativo, tales como:

a) La interface para escritura de registros

b) La rutina para escritura de registros

c)Las rutinas de localización de un proceso específico suministran información al SMF

En forma específica el SMF realiza las siguientes actividades:

Formato y colección de estadísticas por proceso

Transfiere registros a los archivos en disco

Envía mensajes al operador sobre eventos específicos en los archivos

Toda esta información se obtiene usando varios parámetros y es posible determinar la cantidad de información que se registra.

Utilidad de la Información del SMF

La gran variedad y volumen de información registrada nos permite producir diversos tipos de reportes tanto para análisis, como para información de control.

Algunos reportes que pueden ser creados son:

Total de transacciones por usuario

Reporte de horario de servicio

Análisis de configuración

Bitácora de procesos

Actividad de volúmenes de acceso directo (discos)

Actividad de los archivos del sistema

Características de cargas de trabajo

Actividad de programación

Reporte de procesos abortados

Tiempo de uso de la UCP

Ejemplo:

El siguiente reporte es un resumen de los diferentes registros utilizados por el SMF, indicando el periodo al cual corresponde:

Registro Tipo 0	Registro Tipo 73
Registro Tipo 4	IPL (Initial Program Load)
Registro Tipo 5	Terminación de los pasos de un proceso
Registro Tipo 6	Terminación de procesos
Registro Tipo 70	Impresión de procesos
Registro Tipo 71	Actividad de la UCP

SLR (Service Level Reporter)

El SLR es un paquete que puede ser usado en forma efectiva para el monitoreo del rendimiento del sistema y constituye uno de los elementos más utilizados para la planeación de la capacidad.

El SLR construye una base de datos integrada coleccionando la información del SMF proveniente de los subsistemas que se ejecutan dentro del sistema. Su importancia radica en su habilidad para combinar la información de varias fuentes y consolidarla dentro de tablas que pueden ser consultadas por los analistas de la capacidad.

La base de datos puede contener información tanto actual como histórica, a partir de la cuál se crean pronósticos a corto y largo plazo, a través de una interfaz que permite al analista adecuar los reportes a sus necesidades específicas.

Los componentes principales del SLR son:

a) El diálogo que invoca a los comandos de SLR

b) El lenguaje de comandos que opera la base datos

c) Las tablas definidas en la base de datos

d) Las funciones provistas por el usuario en base a macros

A continuación se presenta en forma gráfica la relación entre estos componentes.

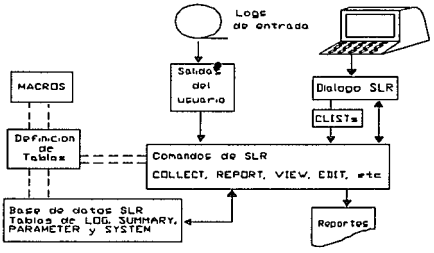


Fig. 4 Relación de componentes de SLR

De los componentes enunciados anteriormente es la base de datos sobre la que SLR apoya su flexibilidad y poderío, es por esto que a continuación describiremos brevemente su contenido y estructura interna.

La base datos SLR es un archivo cuyos registros están lógicamente combinados dentro de tablas. Existen cuatro tipos de tablas:

Tablas de *LOG*

El SLR colecta datos de archivos secuenciales llamados *logs* de entrada. Estos *logs* son creados por diferentes sistemas, subsistemas y programas; la longitud del registro puede ser, variable o indefinida. Los registros de entrada son analizados y escritos en las tablas de *LOG* de la base de datos. Típicamente contienen un renglón por cada registro elegido del archivo de entrada. Ejemplo:

CPULOG contiene información referente a UCP

ADDRLOG contiene información de discos, cintas

WRITERLOG contiene información de impresoras

Tablas de *SUMMARY*

Estas tablas son un resumen de los datos contenidos en las tablas de *log*. Los datos pueden ser resumidos con base en el tiempo, por hora, día, mes año o también pueden ser agrupados como el usuario lo defina, por departamento, por clase de proceso, etc.

Un ejemplo de estas tablas son:

CPUSTAT estadísticas relativas a la UCP

ADDRSTAT estadísticas relativas a discos, cintas

Tablas de *PARAMETERS*

Estas tablas son usadas para tener mediciones objetivas de reportes especiales, trasladar valores por categorías y proveer costos para contabilizar. Un ejemplo de estas tablas son:

WRKLOAD_PER_TAB carga de trabajo

BATCH_BE_TAB procesos realizados en lotes

Tablas de *SYSTEM*

Estas tablas tienen como propósito controlar, referenciar y catalogar la información, se encuentran predefinidas y no se pueden cambiar. Ejemplo de estas tablas son:

MONTHTABLE nombre de los meses

WEEKTABLE número de semana

COLUMNTABLE descripción de columnas por cada tabla

Cada tabla esta formada por renglones y columnas como lo muestra la figura 5. Comparando una tabla con un archivo secuencial diríamos que un renglón en la tabla corresponde a un registro en el archivo y una columna corresponde a un campo en el registro.

El número de renglones en cada tabla es variable, debido a que los renglones se agregan en una función de colectar, o se borran si se esta llevando a cabo mantenimiento a la base de datos. Por lo que respecta al número de columnas, éste es fijo.

Las columnas que contienen datos únicamente identificando los renglones son llamadas columnas clave (*key*), y las columnas que contienen mediciones son llamadas columnas de datos (*data*).

JOB STATISTICS

KEYS				DATA	
YEAR	MONTH	DAY	SYSTEM	NUMBER_JOBS	CPU_TIME
90	APR	8	SYSA	1500	20
90	APR	8	SYSB	700	15
90	APR	8	SYSC	900	13

Fig. 5 Estructura de tablas

Reportes Generados

Además de que el usuario es capaz de adecuar los reportes a sus necesidades, el SLR provee reportes estándar como parte de su configuración inicial. Estos reportes son:

a) Reporte del rendimiento diario. En este reporte la carga de trabajo es impresa a intervalos de tiempo de 10 minutos. Usando este reporte, es fácil determinar los periodos de actividad máxima y mínima durante el día, cuando es usado en combinación con otros reportes del SLR y del SMF. Es posible determinar los valores máximos de carga que causan saturación al sistema.

b) Pronóstico de carga del procesador. Este reporte muestra la variación de la carga de procesos por lote dentro de un periodo largo de tiempo y como puede ser vista al año siguiente. Los datos de años pasados son reducidos y almacenados en la base de datos. Los datos para los pronósticos son creados a través de un conjunto de funciones, las cuáles son una característica estándar del SLR. Este tipo de reportes son muy utilizados para propósitos de la planeación de la capacidad.

JARS (Job Accounting Report System)

El JARS es un sistema por medio del cual se obtienen reportes sobre la contabilidad del sistema. Desarrollado bajo las bases de un programa generador de reportes, y utilizado como herramienta de medición, permite obtener reportes confiables y con un alto grado de detalle.

El JARS interacciona con el sistema SMF, ambos manejan los datos de entrada, y crea una base de datos donde se registran las estadísticas resumidas del tipo de reporte que se haya seleccionado (por ejemplo: por empresa, por departamento, por programador, etc.). Usando la base de datos, es posible producir periódicamente reportes que normalmente requieren demasiado tiempo.

El sistema ofrece una captura de datos externa y simple, por medio de su utilería para la manipulación de datos (*Data Management Utility [DMU]*). Este programa nos da la facilidad de combinar información no propia del sistema, por ejemplo: costos unitarios, facturación, histórico de facturación de clientes, tipos de papelería.

La flexibilidad de JARS está en el programa reporteador, todas las entradas pueden ser introducidas independiente o simultáneamente. Los tipos de reportes son seleccionados y adecuados a los estándares de la instalación a través de instrucciones sencillas de control.

Se pueden obtener en una sola corrida hasta 15 diferentes reportes de usuarios y hasta 30 reportes de utilización del sistema pueden ser formateados y desplegados con un solo acceso al archivo de entrada. El formateado, edición y selección de información son ejecutados por JARS en un solo paso del proceso. Esto hace que personal sin un gran conocimiento técnico pueda utilizarlo.

Componentes del Sistema JARS

JARS Report program (Programa reporteador JARS)

Data Base Capability (Capacidad de la Base de datos)

Data Management Utility (Utilería para el manejo de datos)

Programas de soporte para:

- Sistema operativo **MVS**

- *Time Sharing Operation* **TSO** (Operación de Tiempo Compartido)

Información generada

La información manejada por JARS en su versión para el sistema operativo MVS es:

TCB CPU Time (Tiempo de UCP para el bloque de control de tareas: *Task Control Block*)

Total Pages Printed (Total de páginas impresas)

SRB CPU Time (Tiempo de UCP para el bloque de petición de sistema: *System Request Block*)

Special Pages Printed (Páginas especiales impresas)

Número de Swaps(Conmutaciones)

Swap Pages-in (Conmutación de páginas hacia memoria principal)

Swap Pages-out (Conmutación de páginas fuera de memoria principal)

Average TCB CPU Time (Tiempo promedio de UCP para TCB)

Average SRB CPU Time (Tiempo promedio de UCP para SRB)

Number of Tape Mounts (Número de montajes de cintas)

Number of Disk Mounts (Número de montajes de discos)

Información capturada del SMF:

Run Date (Fecha de ejecución ó corrida)

Step Name (Nombre del paso)

Job Class (Tipo de trabajo)

Program Name (Nombre del programa)

Job Priority (Prioridad del trabajo)

Start Time (Hora de inicio)

Job Number (Número de trabajo)

Storage Used (Tipo de almacenamiento usado)

Termination Indicator (Indicador de terminación)

CPU Time (Tiempo de UCP)

CPU ID (Número de identificación de la UCP)

Active Name (Nombre activo)

Input Device Name (Nombre del dispositivo de entrada)

Page-in Count (Conteo de páginas subidas a memoria)

User - ID (Identificador del usuario)

Page-out Count (Conteo de páginas bajadas de memoria)

Programmer Name (Nombre del programador)

Total page Count (Cuenta total de páginas)

Standard Lines Printed (Líneas impresas de manera estándar)

Total Lines Printed (Total de líneas impresas)

SYSOUT Class

Reportes generados por el *JARS Report Program*:

Utilization Percentages (Porcentajes de Utilización)

Processor Time (Tiempo de procesador)

Elapsed Paging Rate (Razón de tiempo transcurrido)

Elapsed Time (Tiempo transcurrido)

CPU Paging Rate (Tasa de paginación del UCP)

Turnaround Time (Tiempo transcurrido del trabajo combinado con el tiempo de espera en la cola de trabajos)

Tape I/O Count (Conteo de operaciones de Entrada/Salida en cinta)

Reader Queue Time (Tiempo de espera en la cola, para dispositivos de lectura)

Reader I/O Count (Conteo de operaciones de Entrada/Salida para dispositivos de lectura)

Reader Duration (Tiempo transcurrido desde que comienza la lectura hasta que finaliza)

Printer I/O Count (Conteo de operaciones de Entrada/Salida efectuadas sobre dispositivos de impresión)

Writer Duration (Tiempo transcurrido desde que se inicia la operación de escritura hasta que termina)

Total I/O Count (Conteo total de operaciones de Entrada/Salida)

Average Elapsed Time (Tiempo promedio transcurrido)

Processing ID (Número identificador de procesamiento)

Average CPU Time (Tiempo promedio de utilización de la UCP)

I/O Time (Tiempo de Entrada/Salida)

Number of Printers Used (Número de impresoras usadas)

Total Time (Tiempo total)

Number of Disks Used (Número de discos usados)

Number of Tapes Used (Número de cintas usadas)

Number of Jobs (Número de trabajos)

Debit Amount (Cantidad de débito)

Number of Job Steps (Número de pasos por trabajo)

Budget Amount (Cantidad de presupuesto)

Processor Charge (Carga al procesador)

Total Charge (Carga total)

Distributed Charge (Carga distribuida)

Credit Amount (Cantidad de crédito)

La siguiente figura muestra el flujo del sistema general:

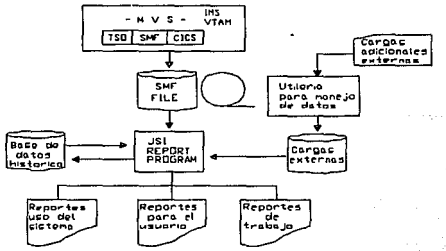


Fig. 6 Flujo del sistema

Resumen de Características Importantes de JARS

Base de Datos Histórico

Reporte por grupos de trabajo

Flexibilidad en la generación de reportes

Generación de Gráficas

Contabilidad de formas especiales

Consolidado de transacciones

Utilería para manejo de información externa

Algoritmos de transacciones múltiples

Selección de reportes

Contabilidad de RJE (*Remote Job Entry*)

Facilidad de Uso y Adaptabilidad.

Ejemplo de un proceso JARS

La siguiente codificación es para obtener el reporte de carga de trabajo de las sesiones en tiempo compartido y el número de procesos que se realizan durante los 7 días de la semana, monitoreados cada hora durante las 24 horas del día

Tarjeta	Codificación
1	//OPERA00Z JOB (2500,$$,,999),'PRUEBA JARS',CLASS = 0,PRTY = 9
2	// MSGCLASS = X,TYPRUN = COPY
3	// USER = OPERA00,PASSWORD =
4	//JARS EXEC PGM = JSIMAIN,TIME = 40,REGION = 1024K
5	//STEPLIB DD DSN = SYS3.JARS.LINKLIB,DISP = SHR
6	// DD DSN = SYS2.SORTVS.LINKLIB,DISP = SHR
7	//SORTLIB DD DSN = SYS2.SORTVS.SORTLIB,DISP = SHR
8	//SYSOUT DD SYSOUT = X
9	//SORTSOUT DD SYSOUT = X
10	//SNAPDUMP DD SYSOUT = X
11	//SMFINP DD DSN = PSIST.SMF.PROT,DISP = OLD,
12	// VOL = SER = PRO800,LABEL = (,SL),UNIT = TAPE6250
13	//ACCOUNT DD UNIT = SYSALLDA,SPACE = (CYL,(150,100))
14	//SORTWK01 DD UNIT = SYSALLDA,SPACE = (CYL,(150,100),,CONTIG)
15	//SORTWK02 DD UNIT = SYSALLDA,SPACE = (CYL,(150,100),,CONTIG)
16	//SORTWK03 DD UNIT = SYSALLDA,SPACE = (CYL,(150,100),,CONTIG)
17	//CONTROL1 DD UNIT = SYSALLDA,SPACE = (CYL,(15,15))
18	//CONTROL2 DD UNIT = SYSALLDA,SPACE = (CYL,(15,15))
19	//PRINTOUT DD SYSOUT = X,DCB = (LRECL = 133,BLKSIZE = 13300,RECFM = FBA)
20	//CARDIN DD *
21	,CONFIG READ00C 01A WRIT00E 380 PUNC01B
22	,CONFIG TAPE920 921 922 924
23	,CONFIG DISK160 161 162 163 164 165 166 167 168 169 16A 16B 320 321
24	,CONFIG 322 323 324 325 328 329 32A 32B 32C 32D 32E 32F 431 432 433
25	,CONFIG 434 435 437 700 701 702 703
26	,CONFIG OTHE080 A80 00E B60 B80
27	,SELECT 1 333 3 (5GHK
28	BHEADER 035WEEKDAY WORKLOAD PROFILE BY H0
29	BSORT 43701A2103702A2 04306A
30	BDISPLAY 0020142B40AB0C50C80661381370191D11D2164
31	BTITLE B4 NBR SESSNS
32	BTITLE 02 DAYS INDAY-OF-WEEK SAMPLE
33	BTITLE 14 NBR JOBS
34	BDESCRIPT11 MONDAY

Descripción de cada tarjeta.

Tarjetas 1 a 3: Identifica al usuario que realiza el proceso, nombre, número de cuenta

Tarjetas 4 a 7: Invocamos la ejecución del proceso con la declaración EXEC, así como a las bibliotecas necesarias

Tarjetas 8 a 10: Definición de salidas (a terminal o impresión)

Tarjetas 11 y 12: Definición de los archivos, en este caso es una cinta magnética

Tarjetas 13 a 18: Definición de espacio temporal en disco, para *sorts* básicamente

Tarjeta 19: Definición de clase de impresión

Tarjeta 20: Declaración de parámetros de entrada al proceso

A partir de la tarjeta 21, se codifican las opciones seleccionadas dependiendo del tipo de información requerida.

JARS, como se vió, ofrece múltiples alternativas. A manera de información describiremos las siguientes tarjetas:

Tarjetas 21 a 26: **CONFIG:** Aquí definimos la configuración del sistema, lectoras, impresoras, discos, cintas, etc; con sus respectivas direcciones. Esta tarjeta es necesaria.

Tarjeta 27: **SELECT:** Define entre otras cosas:

El archivo fuente para proceso (SMF)

Define que necesitamos los reportes de uso del sistema

Define la creación de reportes particulares

Define que se analizará la actividad por proceso

Esta tarjeta es requerida.

Tarjeta 28: **HEADER:** Definición de la posición del título principal del reporte

Tarjeta 29: **SORT:** Definimos el número de niveles de ordenamiento que necesitamos, hasta de cinco diferentes campos

Tarjeta 30: **DISPLAY:** Esta declaración nos permite seleccionar la información a imprimir

Tarjeta 31: **TITLE:** Definimos la contabilización del número de sesiones en tiempo compartido, con el título correspondiente

Tarjeta 32: **TITLE:** Definimos los días y horas de acuerdo a la tarjeta **DESCRIPT**

Tarjeta 33: **TITLE:** Definimos la contabilidad del número de procesos, de acuerdo a las horas y días asignadas en la tarjeta **DESCRIPT**

Tarjeta 34: **DESCRIPT:** Definimos los días y horas de los cuáles necesitamos la contabilización, a saber, los siete días de la semana a cada hora las 24 horas del día

Cada tarjeta de control nos dá una alternativa para obtener nuestro reporte, existe una gran variedad, siendo algunas opcionales y otras requeridas.

Este ejemplo pretende ilustrar la codificación necesaria para el proceso JARS. Su utilización aunque no compleja, pero si laboriosa, requiere de un conocimiento del paquete bastante profundo.

MODELADO

En la metodología de la planeación de la capacidad el modelado juega un papel muy importante, ya que por medio de él podemos simular y predecir la capacidad de los sistemas, básicamente las cargas de trabajo y tiempos de respuesta.

Desde un punto de vista teórico existen varias técnicas de modelado que pueden ser utilizadas para realizar el análisis de la capacidad. Por su baja complejidad y costo, dos de estas técnicas resultan muy utilizadas:

a) La regresión lineal y,

b) La teoría de colas

Aplicando estas técnicas se obtiene un modelo, al cual le son alimentados los valores conocidos de la carga de trabajo como entrada. A continuación describiremos ambas técnicas, ofreciendo un ejemplo de su aplicación a los sistemas de computación.

2.3 Regresión Lineal

En la práctica encontramos que existe una relación entre dos o más variables de nuestro sistema de cómputo, y es necesario expresar esta relación en forma matemática.

Para lograr este objetivo, es necesario apoyarnos en las estadísticas y seleccionar los datos importantes para el análisis de la capacidad del equipo o del sistema en cuestión.

Algunas variables que pueden considerarse son:

Utilización

Cargas de trabajo

Tiempo de respuesta

Hora del día

Día de la semana

Transacciones por segundo

Tiempo de espera

Número de paginaciones, etc.

Como resultado de graficar en un sistema de coordenadas esta información, se obtiene un *diagrama de dispersión*, del cual visualizamos la *curva de aproximación* que es representativa de los datos.

Por ejemplo, en la figura 7 observamos que los datos se aproximan bien por una recta, es decir, existe una *relación lineal* entre las variables. Sin embargo, en la figura 8 aunque existe una relación entre las variables, ésta no es lineal y se denomina como *relación no lineal*.

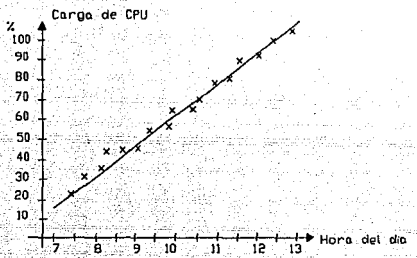


Fig. 7 Relación lineal

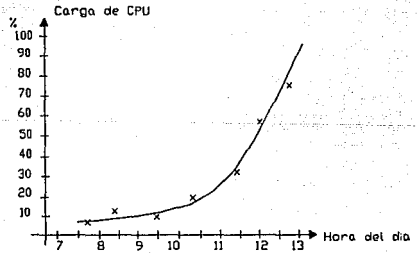


Fig. 8 Relación no Lineal

Otra forma de llegar a obtener la *curva de aproximación* es a través de consideraciones analíticas o teóricas, que nos llevan a obtener la curva que se ajusta a los datos graficados, a esta curva se le denomina *curva de ajuste*.

El propósito de la *curva de ajuste* es estimar una de las variables de la otra, este proceso de estimación se conoce como **regresión**.

Volviendo a la figura 7, debido a la variación del muestreo no todos los puntos caen sobre la línea, más bien están dispersos alrededor. Suponemos que para cada valor de **x** existe una distribución para los valores de **y**. La línea de regresión es una línea que une las medias de las distribuciones correspondientes a los diferentes valores de **x**. Bajo esta suposición, la relación que queremos estimar es:

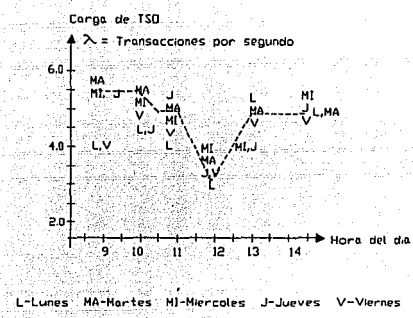
$$\mu x/y = A + Bx$$

Esto significa que el valor medio de **y** para un valor fijo de **x** es igual a A+Bx.

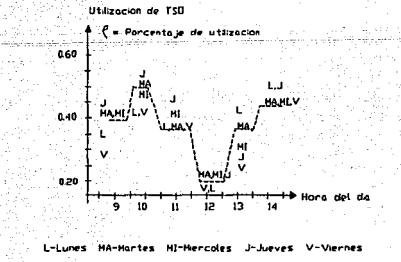
Como ejemplo supongamos que tenemos necesidad de visualizar la curva que nos represente la utilización contra la carga de trabajo de una aplicación denominada TSO (Time Sharing Operation).

Hora	Lunes		Martes		Miercoles		Jueves		Viernes	
	Carga-Utilización		Carga-Utilización		Carga-Utilizacion		Carga-Utilización		Carga-Utilización	
9	4.3	.36	5.4	.39	5.2	.39	5.3	.45	4.2	.29
10	4.6	.38	5.5	.43	5.0	.40	4.5	.47	4.8	.38
11	4.1	.36	4.7	.34	4.6	.38	5.0	.43	4.3	.34
12	2.5	.16	3.0	.21	3.3	.21	2.7	.19	2.6	.17
13	5.0	.42	4.7	.36	4.0	.32	4.0	.28	4.5	.31
14	4.7	.43	4.7	.40	5.3	.39	5.1	.42	4.6	.40

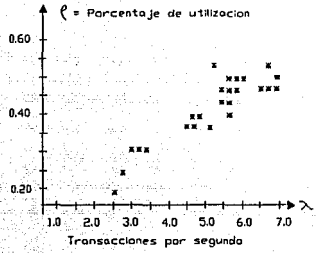
Estadística de las mediciones de TSO



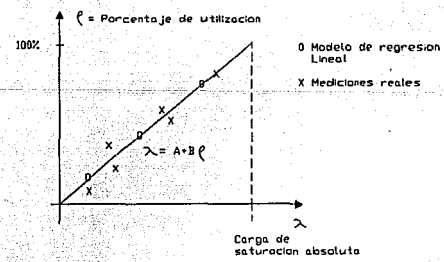
Carga del TSO vs. Hora del día



Utilización de TSO vs. Hora del día



Utilización vs. Carga (TSO)



Utilización vs. Carga de Trabajo

Se observa que la regresión lineal no es un procedimiento complejo, pero si nos proporciona una gran ayuda para el proceso de modelado dentro de la metodología expuesta.

Aplicando regresión lineal en el capítulo 4, se expone el caso práctico de análisis de capacidad en unidades de disco.

2.4 TEORIA DE COLAS

La teoría de colas, que fue originalmente desarrollada por el matemático Danés A.K. Erlang, para el diseño de sistemas telefónicos, ha sido adaptada y extendida para convertirse en una útil herramienta en el diseño y análisis de sistemas de computación. La teoría de colas puede ser utilizada para predecir medidas del rendimiento de la computadora tales como: El tiempo de espera para usar una terminal en línea, los requerimientos de almacenamiento en centros de conmutación de mensajes, y los estimados de los efectos de la asignación de prioridades en un sistema interactivo.

Una cola es una línea de espera y, en comunicaciones y computación, la teoría de colas es el estudio del fenómeno *línea de espera-línea de servicio*. Por lo regular se habla de una cola de requerimientos esperando a ser procesada por un sistema en línea, dicha cola puede crear también otras colas. Esa cola puede más adelante causar una línea de espera de solicitudes de Entrada/Salida (E/S), lo cual puede, a su tiempo, generar una cola de peticiones de canal, y así sucesivamente.

Dentro de los elementos de la teoría de colas existe una población o fuente de clientes potenciales, donde el término de "cliente" significa alguien quien compra o usa un producto o servicio. En comunicaciones y computación un cliente puede ser un mensaje a ser transmitido, un requerimiento de proceso o una solicitud de E/S.

El cliente desea algún tipo de servicio- la transmisión de un mensaje, el procesamiento de una requisición o el servicio para una solicitud de E/S- de una fuente de servicios. En la fuente de servicios existen uno o más servidores, los cuales son unidades que proveen el servicio solicitado por el cliente. Si todos los servidores están ocupados cuando un cliente entra al sistema, se integra a una cola hasta que un servidor esté disponible.

Algunas variables aleatorias usadas en el estudio de sistemas de colas se encuentran ilustradas en la figura 9. El anexo 1 recopila las definiciones de teoría de colas utilizadas en este capítulo. Ciertas relaciones claves entre las variables de sistemas de colas son mostradas en el anexo 2.

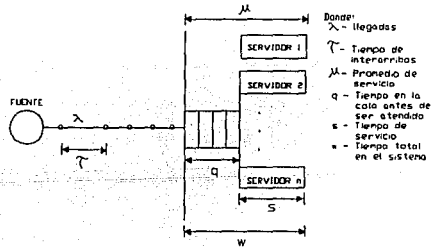


Fig. 9 Variables usadas en teoría de Colas

Especificaciones de una Cola

El estudio matemático o el modelado de un sistema de colas requiere de la explicación de las siguientes especificaciones:

a) Fuente

La población fuente puede ser finita o infinita. Un sistema de fuente finita no puede tener un largo de cola arbitrario para el servicio, y el número de clientes en el sistema afecta la tasa de llegadas. En el caso extremo, si cada cliente está esperando o recibiendo servicio la tasa de llegadas cae a cero. Si la fuente es finita pero grande, asumimos una población de clientes infinita para simplificar la parte matemática.

b) Proceso de Arribo

Asumimos que los clientes entran a la cola del sistema en tiempos $t_1 < t_2 < t_n$. Las variables aleatorias $\tau_k = t_k - t_{k-1}$ (donde $k \geq 1$) son llamadas tiempos de interarribo. Asumimos el τ_k de una secuencia de variables aleatorias independientes e idénticamente distribuidas y usamos el símbolo τ para un tiempo de interarribos arbitrario. Especificamos el proceso de arribos, dando una función A de distribución del tiempo de interarribo, $A(t) = P[\tau \leq t]$. Los patrones de arribo más comunes en la terminología de teoría de colas son: entrada aleatoria, patrón de arribo aleatorio, o un proceso de arribo de Poisson. Si la distribución del tiempo de interarribo es exponencial, esto es si $P[\tau \leq t] = 1 - e^{-\lambda t}$ para cada tiempo de interarribo, entonces la probabilidad de n arribos en cualquier intervalo de tiempo de longitud t es $e^{-\lambda t}(\lambda t)/n!$ donde $n = 0,1,2,.$ Aquí λ es el promedio de la tasa de arribo, y los arribos tienen una distribución de Poisson. Otras distribuciones de tiempos de interarribos son Erlang-k y distribuciones constantes.

c) Tiempo de Distribución del Servicio

Sea S_k el tiempo de servicio requerido por el K-ésimo cliente. Asumiremos que el S_k es independiente y con variables de distribución idénticamente distribuidas. Asimismo, podemos hacer referencia a un tiempo de servicio como s. De la misma manera asumimos para el tiempo de servicio la función de distribución común:

$W_S(t) = P[s \leq t]$.

La distribución del tiempo-servicio más común en teoría de colas es exponencial, la cual define un servicio llamado *servicio aleatorio*: El símbolo μ está reservado para el promedio de la tasa de servicio y la función de distribución para el servicio aleatorio está dada por $W_S(t) = 1-e^{-\mu t}$, donde $t \geq 0$. Otras distribuciones comunes para el tiempo de servicio son el Erlang-k y la distribución constante.

Un parámetro estadístico muy usado como una medida del carácter de las distribuciones de probabilidad para el tiempo de interarribo y para el tiempo de servicio es el coeficiente de variación cuadrado C^2_x, el cual está definido por la siguiente ecuación:

$$C^2_x = \frac{VAR[X]}{E[X]}$$

Si X es una variable aleatoria constante, entonces $C^2_x = 0$; si X tiene una distribución exponencial, entonces $C^2_x = 1$; y si X tiene una distribución Erlang-k, entonces $C^2_x = 1/k$. Concluimos que, para C^2_T cerca o igual a cero , el proceso de arribos tiene un patrón regular; si C^2_T esta cerca ó es igual a 1 , el proceso de arribos esta cerca de un carácter aleatorio; y si C^2_T es mayor que 1, los arribos tienden a agruparse. Declaraciones similares pueden ser hechas para la distribución del tiempo de servicio, donde valores pequeños de C^2_S, corresponden a tiempos de servicio más ó menos constantes y valores grandes de C^2_S corresponden a tiempos de servicio con gran variabilidad.

d) Capacidad máxima de la Cola del Sistema

En algunos sistemas de colas, la capacidad de la cola se asume como infinita. Esto es, que a todos los clientes que llegan se les permite esperar hasta que el servicio pueda ser provisto.

Otros sistemas llamados *loss systems* (*sistemas desbalanceados*) no tienen capacidad para mantener una linea de espera. Esto es, si un cliente arriba cuando el servicio esta totalmente utilizado, el cliente es rechazado. Otros sistemas de colas tienen una capacidad positiva (pero no infinita).

e) Número de Servidores

El sistema de colas más simple es el sistema de un solo servidor , el cual puede servir a un solo cliente a la vez. Un sistema de múltiples servidores tiene *c* servidores idénticos y puede servir hasta c clientes simultáneamente. En un sistema con un número infinito de servidores, todos los arribos de clientes son inmediatamente provistos con un servidor.

f) Disciplina de Cola

La disciplina de cola, algunas veces llamada disciplina de servicio, es la regla para seleccionar al siguiente cliente que recibirá el servicio. La disciplina de cola más común es "el primero en entrar primero en servir", abreviado como PEPS (o más comunmente conocido por sus siglas en ingles FIFO). Otra disciplina de colas usada constantemente es "ultimo en entrar primero en servir" abreviado como UEPS (o por sus siglas en ingles LIFO).

Una notación corta llamada notación Kendall ha sido desarrollada para especificar sistemas de colas y tiene la forma A/B/c/K/m/Z. Aquí A especifica la distribución del tiempo de interarribo, B la distribución del tiempo de servicio, c el número de servidores, K la capacidad del sistema, m el número en la fuente, y Z la disciplina de cola. La notación corta A/B/c es constantemente usada cuando no existe límite en la línea de espera, la fuente es infinita, y la disciplina de cola es FIFO. Los símbolos usados para A y B son los siguientes:

GI: Tiempo de interarribo independiente general.

G : Tiempo de servicio General, asumido usualmente con independencia.

E : Erlang-k de interarribo o distribución del tiempo de servicio.

M : Interarribo exponencial ó distribución del tiempo de servicio.

D : Interarribo determinísticos (constante) ó distribución del tiempo de servicio.

Así por ejemplo, un M/E/3/20/x/FIFO el sistema tiene un tiempo de interarribo exponencial, tres servidores con idéntica distribución de tiempo de servicio Erlang-4, capacidad del sistema de 20 (3 en servicio y 17 en la cola), infinita fuente de clientes y servicio con política FIFO

g) Intensidad del Tráfico

La intensidad del tráfico es la razón del promedio del tiempo de servicio E[s] entre el promedio del tiempo de interarribos E[τ]. Esta razón es uno de los parámetros más importantes de los sistemas de colas y está definida por la siguiente formula:

$$u = \frac{E[s]}{E[\tau]} = \lambda E[s] = \frac{\lambda}{\mu}$$

La intensidad de tráfico u determina el número mínimo de servidores que son requeridos para mantener el flujo de entrada de clientes. Así, por ejemplo, si E[τ] es 10 segundos y E[s] es 15 segundos , al menos son requeridos dos servidores. La unidad de la intensidad de tráfico es el Erlang, nombrado en honor de A.K. Erlang.

h) Utilización del Servidor

Otro parámetro importante es la intensidad de tráfico por servidor o u/c, llamada *utilización del servidor* ρ cuando el tráfico se encuentra dividido igualmente entre los servidores. La utilización del servidor es la probabilidad que algún servidor dado se encuentre ocupado y así, por la ley de los Grandes Números, ρ es la fracción aproximada del tiempo que cada servidor está ocupado.

Probabilidad que n clientes estén en el sistema a un tiempo t

Esta probabilidad $P_n(t)$ depende no únicamente de t sino también de las condiciones iniciales del sistema de colas, esto es, el número de clientes presentes cuando el servicio se inicia. Para la mayoría de los sistemas de colas utilizados, cuando t se incrementa, $P_n(t)$ se acerca al valor de P_n, el cual es independiente tanto de t como de las condiciones iniciales. Se dice entonces que el sistema está en una condición de estado estable. Dentro de este capítulo consideramos únicamente soluciones de problemas de colas en estado estable, ya que soluciones dependientes del tiempo o transitorias son generalmente demasiado complejas para usos prácticos.

La teoría de colas provee mediciones estadísticas para el rendimiento de sistemas de colas, así como también ayudas para que el ingeniero de sistemas pueda diseñar un sistema, con un mínimo costo que provea el nivel de servicio requerido. Estas mediciones estadísticas y sus varianzas incluyen lo siguiente:

Tiempo promedio de espera en la cola W_q

Tiempo promedio de espera en el sistema **W**

Número promedio de espera para el servicio L_q

Número promedio en el sistema **L**

Estas medidas no son independientes, y se asume que las distribuciones del tiempo de interarribo y del tiempo de servicio son conocidas, el conocimiento de cualquiera de ellas hace posible calcular las otras tres fácilmente (ver anexo 2). Así, si el valor de W_q es calculado primero, entonces los siguientes valores son obtenidos:

$$L_q = \lambda W_q$$

$$W = W_q + W_s$$

$$L = \lambda W$$

Otra medida del rendimiento muy usada es el 90avo del valor porcentual del tiempo de respuesta del sistema $\pi u(90)$, el cual está definido como la suma del tiempo dentro del sistema tal que el 90avo porcentual de todos los arribos de clientes gastan menos que esta suma de tiempo en el sistema. Expresado simbólicamente $\pi u(90)$ está definido por la ecuación $P[\ w \leq \pi u(90)] = 0.9$. El 90avo del valor porcentual del tiempo en la cola $\pi q(90)$ está definido en forma similar.

Modelos M/M/1 de un solo Servidor

Consideremos ahora modelos de sistemas de un solo servidor que específicamente son muy usados. El modelo de colas M/M/1 es usado extensivamente debido a que las distribuciones exactas de las variables aleatorias de interés pueden ser determinadas, además de tener una forma simple. Este modelo tiene distribuciones de interarribos y tiempos de servicio exponenciales (M/M) y un solo servidor (1). A este respecto, el sistema M/M/1 es marcadamente diferente a muchos modelos de colas para los cuales únicamente los valores promedio y, posiblemente, la desviación estandard de las variables aleatorias de interés pueden ser calculadas. Otras razones del uso extendido del sistema M/M/1 es que constantemente se asume un patrón de arribos aleatorio. Sin embargo, para las distribuciones del tiempo de servicio de la UCP, la desviación estandard puede ser mucho más grande que la media, y el modelo M/M/1 da predicciones optimistas aproximadas. Las fórmulas de estado estable para modelos de sistemas de colas M/M/1 están dadas en el anexo 3.

Algunas de las variables aleatorias para el modelo M/M/1 tienen una forma familiar. El número de clientes en el sistema N tiene una distribución geométrica. El tiempo de espera o de respuesta del sistema w tiene una distribución exponencial. El tiempo en que un cliente dado espera en la cola q tiene una distribución mezclada que es discreta en el origen ($P[q=0] = 1-\rho$) y es continua en otra parte. El tiempo de espera del estado estable en la cola tiene una función de distribución que está dada por la siguiente formula:

$$W_q(t) = 1 - \rho e^{-t/E[w]}$$

válida para todo $t \geq 0$

Uno de los puntos que resaltan en las fórmulas para el sistemas M/M/1, es la alta dependencia no lineal de las variables aleatorias para el número de clientes en estado estable, en la cola N_q y en el sistema n (mas q y W anteriormente definido) en la utilización ρ del servidor. Así el promedio de la razón del tiempo de espera en la cola E[q]/E[s] se incrementa de 0.111 cuando ρ = 0 0.1, a 4 cuando ρ = 0.8 y a 9 cuando ρ =0 0.9. La no linealidad es ilustrada en la figura 10. Esta figura también muestra como una alta utilización del servidor conduce a largos tiempos de espera, en la cola antes de recibir servicio, q, y largos tiempos en el sistema, w, incluyendo la espera en la cola. Por supuesto que estos valores se incrementan sin límite cuando se aproxima a 1.

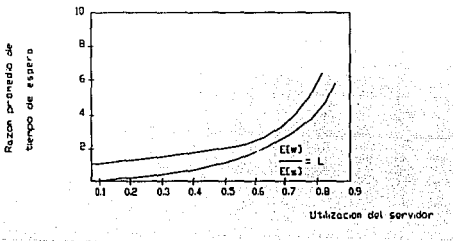


Fig. 10 No linealidad

Esta figura también muestra que para valores de arriba de 0.8, pequeños incrementos en la tasa de arribo degradan dramáticamente el rendimiento del sistema. Por esta razón, los sistemas con alta utilización del servidor son indeseables para sistemas sin políticas de prioridades para el cliente. Los sistemas de prioridades de colas bien diseñados pueden funcionar bien con alta utilización del servidor.

El sistema M/M/1 puede ser usado para ilustrar el efecto de escala. Este efecto es que, dado un sistema M/M/1 con una tasa promedio de arribo λ y una tasa promedio de servicio μ, si ambas λ y μ son dobladas (con ρ sin cambiar) el efecto resultante es que se divide tanto el promedio del tiempo de espera en la cola E[q] como el promedio del tiempo gastado en el sistema E[w]. El número promedio de espera en la cola y el número promedio en el sistema permanecen inalterables. De hecho si para el nuevo sistema reemplazamos λ por $n\lambda$ y μ por $n\mu$ entonces tenemos las siguientes relaciones de escala:

$$\frac{E[q]_{nuevo}}{E[q]_{viejo}} = \left[\frac{\rho/n\mu}{1-\rho}\right] \Big/ \left[\frac{\rho/\mu}{1-\rho}\right] = \frac{1}{n}$$

y

$$\frac{E[w]_{nuevo}}{E[w]_{viejo}} = \left[\frac{1/n\mu}{1-\rho}\right] \Big/ \left[\frac{1/\mu}{1-\rho}\right] = \frac{1}{n}$$

El siguiente argumento sigue una línea de razonamiento intuitivo que es respaldado por el efecto de escala. Si la carga de trabajo de una computadora grande es dividida igualmente entre n computadoras pequeñas, cada una con 1/n veces la velocidad del sistema grande, entonces el tiempo de respuesta no cambia, y los clientes tienen computadoras más convenientemente localizadas. El efecto de escala muestra sin embargo, que el tiempo de respuesta se incrementa por un factor de n , en promedio.

Ejemplo de un sistema M/M/1

En este ejemplo se modelará un sistema con un solo servidor y con un número limitado de consumidores.

En un conmutador de líneas de comunicación llegan los mensajes con un patrón aleatorio, con una taza promedio de llegadas de 240 mensajes por hora. La longitud de los mensajes está distribuida aproximadamente en forma exponencial, con una media de 150 carácteres. El tiempo de transmisión de un mensaje es directamente proporcional a su longitud; la rapidez es de 15 carácteres por segundo. Asumiendo que se cuenta con un *buffer* muy grande, encontrar las siguientes características del sistema:

- El número promedio de mensajes en espera L_q

- El tiempo promedio de espera en la cola W_q

- El tiempo promedio de espera para mensajes retrasados $E[q \mid q>0]$

- El número promedio de mensajes en el sistema L

- El tiempo promedio de estadía dentro del sistema

Solución.

De la descripción del ejemplo y de las fórmulas para un sistema M/M/1 del anexo 3 tenemos:

1) La tasa de llegadas de mensajes es:

λ = 240 mensajes/hora

= 1/15 mensajes/seg

2) Un tiempo promedio de servicios de:

$$E[s] = 150/15 = 10 \text{ seg}$$

3) Una utilización del servidor de:

$$\rho = \lambda E[s] = 10/15 = 2/3$$

con los datos anteriores, podemos calcular:

$L_q = E[N_q] = \rho^2/1-\rho = (2/3)^2 \div (1-2/3) = 1.33$ mensajes

(Número promedio en la cola)

$W_q = E[q] = \rho E[s]/1-\rho = (2/3)(10) \div (1-2/3) = 20$ seg.

(Tiempo promedio de espera en la cola)

$E[q \mid q>0] = E[s]/1-\rho = 10 \div (1-2/3) = 30$ seg.

(Tiempo promedio de espera en la cola para mensajes retrasados)

$L = E[M] = \rho/1-\rho = 2/3 \div (1-2/3) = 2$ mensajes

(Número de promedio de mensajes dentro del sistema)

$W = E[s] = E[s]/(1-\rho) = 30$ seg.

(Tiempo promedio de estadía dentro del sistema)

Prioridad de Colas

En muchos sistemas de colas los clientes son divididos en clases de prioridad, digamos de 1 a n, en donde el número de clase más bajo designa la prioridad más alta. Así, los clientes de prioridad clase i tendrán preferencia sobre los clientes de prioridad clase j si i<j, y los clientes de prioridad clase 1 tendrán preferencia sobre todos los demás clientes. Los clientes con la misma clase de prioridad serán servido con base a su orden de arribo (FIFO).

Existen dos políticas de control básicas para la situación en donde un cliente de clase i arriba para encontrar a un cliente de clase j que esta siendo atendido (i<j), llamadas prioridades *de derecho* y *de no derecho*. En una prioridad *de derecho* el servicio es interrumpido y el nuevo cliente con prioridad más alta es atendido. Como un refinamiento, si el sistema *de derecho* es un sistema con prioridad *de derecho-continúa*, el cliente con prioridad más baja, y cuyo servicio fue interrumpido , reasumirá el servicio en el punto donde fue interrumpido. En otra variación el cliente de baja prioridad repite su entrada al servicio desde el principio.

En un sistema de colas con prioridad *de no derecho*, el nuevo cliente que está arribando espera hasta que el cliente que se encuentra haciendo uso del servicio termine. Entonces se le permite hacer uso del servicio. Tal sistema es también llamado cabeza de línea.

No es difícil mostrar que si las prioridades de clases son puestas para favorecer a los clientes con promedios pequeños de tiempos de servicio, entonces el tiempo medio en el sistema decrecerá.

Muchos centros de procesamiento utilizan esta técnica para dar un mejor servicio a todos los clientes . Por otro lado, si por alguna razón una porción importante de clientes con largos tiempos de servicio deben ser favorecidos entonces el rendimiento total del sistema se decrementa.

CAPÍTULO 3
METODOLOGÍA

Hasta aquí se han descrito en forma independiente todos los elementos que conforman la metodología por delinear. En este capítulo definiremos paso a paso las etapas que la componen y daremos una descripción detallada de cada una de ellas, dando énfasis a los elementos utilizados dentro de cada etapa.

El planteamiento de la metodología es de carácter general, lo que significa que no es exclusiva de una marca ni de una configuración en particular. Por otro lado, no existe una forma específica en la cual la planeación de la capacidad deba ser implantada, esto es debido a los diferentes niveles de sofisticación en que se encuentran las instalaciones de procesamiento de datos. Como un esfuerzo para entender la compleja área de planeación de la capacidad y su relación con una instalación específica, definiremos a continuación los tipos de instalación existentes, desde el punto de vista del análisis de capacidad.

Instalación tipo 1.- Es aquella en la que no existen esfuerzos reales para realizar planeación de la capacidad, son instalaciones pequeñas que no pueden dedicar recursos para realizar este tipo de actividades. La mayoría de estas instalaciones probablemente realizan, por medio de alguna herramienta, mediciones de su sistema pero no son salvadas ni usadas para ningún tipo de análisis. El personal no está capacitado en técnicas de medición y análisis.

Instalación tipo 2.- Dentro de este ambiente es bien conocida la necesidad de la planeación de la capacidad y se encuentran tratando de iniciar esfuerzos al respecto.

Los planes son emplear recursos específicos para realizar las tareas de la planeación de la capacidad; sin embargo, únicamente el 10 ó 20 por ciento del tiempo disponible del administrador del sistema es dedicado a estas labores. En estas instalaciones son utilizadas un buen número de herramientas de medición y generación de reportes pero el personal técnico usualmente no es muy experimentado en la medición y análisis de sistemas computacionales.

Instalación tipo 3.- En este tipo de instalaciones los esfuerzos de planeación son serios y en la mayoría de los casos han comprometido a un grupo de personas de tiempo completo para esta tarea. Normalmente esta instalación se encuentra probando diferentes enfoques de análisis (por ejemplo: guías, coleo, simulación etc.) y emplea herramientas de medición y generación de reportes de muchos tipos. Constantemente se encuentran buscando las tendencias de las herramientas de selección, procedimientos de análisis, y cuestionando ciertos productos de *hardware* y *software*. Bajo una dirección adecuada este tipo de instalaciones pueden implantar un programa de planeación de la capacidad muy exitoso.

Instalación tipo 4.- En este tipo de instalaciones su comportamiento ha sido rastreado durante un buen número de años y tienen funcionando un programa de planeación muy confiable. Esta instalación usualmente obtiene buenos resultados con el uso de varias herramientas de medición y técnicas de análisis, es por lo general un departamento dedicado a las labores de planeación. No necesariamente buscan asesoría o tendencias en la planeación de la capacidad, más bien se encuentran en un estado de intercambio de información. En muchos casos, estas instalaciones se encuentran utilizando técnicas simples de análisis (por ejemplo: guías, análisis lineal, etc.) así como investigando otros métodos para afinar sus procedimientos de análisis (por ejemplo: coleo, simulación, etc.).

La metodología propuesta está dirigida a los cuatro tipos de instalación, todas ellas tienen la necesidad de hacer planeación. Por ende, la fuente de requerimientos apropiados, el nivel correcto de detalle en las mediciones y la técnica de análisis, deben ser seleccionadas de acuerdo al tipo de instalación.

Esto significa, en la mayoría de los casos, que un plan de capacidad para una instalación tipo 1 será significativamente diferente al de una instalación tipo 4.

En el caso práctico descrito en el capítulo siguiente, la metodología es aplicada a una instalación tipo 4.

3.1 Etapas de la metodología

La siguiente figura nos muestra el proceso que se debe seguir en la planeación de la capacidad.

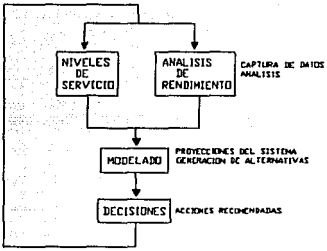


Fig. 11 Etapas para el desarrollo de planeación de la capacidad

La metodología se encuentra dividida en cuatro etapas:

a) Niveles de servicio

b) Análisis de rendimiento

c) Modelado

d) Toma de decisiones

Dentro de cada una de éstas etapas deberán ser realizados una serie de pasos, los cuales serán descritos a continuación.

3.1.1 Niveles de servicio

El objetivo de esta etapa es cuantificar tanto la demanda futura de servicio para el sistema, como los niveles de servicio requeridos actualmente.

Para lograr el objetivo se deberán realizar las siguientes actividades:

Proyección de las cargas de trabajo

Esta actividad consiste en determinar las cargas de trabajo futuras por cada una de las aplicaciones que son ejecutadas en el sistema, las cuales son definidas de la siguiente manera:

i) Que el usuario especifique en base a sus expectativas, los porcentajes de crecimiento de sus aplicaciones.

ii) Con base en el comportamiento histórico de las cargas de trabajo dentro del sistema determinar mediante un análisis estadístico las demandas de trabajo futuras.

Es importante hacer notar que estas proyecciones deberán ajustarse lo más que sea posible a la realidad, tratando de abarcar períodos como mínimo de seis meses y un máximo a dos años; debido a la frecuencia con que deben desarrollarse los estudios de planeación de la capacidad o bién por los cambios en las tendencias tecnológicas.

Niveles de servicio requeridos

Dentro de esta actividad deberán ser establecidos por medio de acuerdos y negociaciones con los usuarios, los requerimientos de servicio que se van a instituir.

Estos niveles de servicio están principalmente enfocados a las transacciones en línea, en donde los tiempos de respuesta requeridos juegan un papel muy importante. No obstante, deberán tomarse en cuenta también aspectos como: espacio requerido en discos y cinta para los nuevos proyectos, tiempo de proceso, número de líneas de impresión, tiempo de entrega cuando el proceso sea por lotes y cualquier otro aspecto que impacte a los niveles de servicio.

Un punto importante al momento de establecer los requerimientos es que, una vez mejorado el servicio, éste debe mantenerse. Algunos ejemplos de niveles de servicio son: el número de transacciones por segundo, el número de procesos por hora, el tiempo de respuesta del sistema en segundos, la capacidad en discos, entre otros.

3.1.2 Análisis de Rendimiento

En esta etapa el objetivo es comprender cual es la capacidad de carga y utilización que esta ofreciendo actualmente el sistema.

Para lograr el objetivo se deberán realizar las siguientes actividades:

Recopilación de parámetros clave

Los parámetros clave del rendimiento y dependiendo del tipo de componente del sistema son:

i) Porcentaje de espacio utilizado (discos, cintas y cartuchos)

ii) Número de trabajos terminados (UCP e impresoras)

iii) Tiempo promedio en que un requerimiento es servido (UCP, controladores, canales y equipo de comunicaciones)

iv) Tiempo promedio en espera de servicio (UCP, controladores, canales, equipo de comunicaciones)

v) Tiempo promedio de respuesta (iii + iv)

Todos estos parámetros se obtienen por medio de las herramientas mencionadas en la actividad de mediciones descrita posteriormente.

Análisis de situación actual

Dentro de esta actividad se deben identificar todos los elementos que permitan obtener un panorama detallado de los componentes del sistema. Los elementos principales que deben tomarse en cuenta son:

i) Configuración del sistema

Identificar el número de unidades de cinta, discos, impresoras y cartuchos

Identificar el número de UCPs

Identificar la cantidad de equipo de comunicaciones

ii) Criterios de saturación. Aplicar los valores de los niveles de saturación y subutilización proporcionados por el fabricante para cada componente del sistema que permitan operar con un nivel de servicio óptimo. Cabe hacer mención que estos criterios de saturación son, la mayor de las veces, los resultados de pruebas realizadas en sus laboratorios. Sin embargo, estos criterios pueden modificarse en base a las experiencias manifestadas por los usuarios.

Mediciones del Sistema

Esta actividad es una de las más importantes basadas en las mediciones y su exactitud dependerá el grado de complejidad y certeza del estudio de planeación.

Los datos que se deberán obtener durante esta actividad son:

a) Para dispositivos de almacenamiento secundario: capacidad en Mb, velocidad de transferencia en Mb/seg, utilización en Mb.

b) Para UCPs, controladores y equipo de comunicaciones: transacciones por unidad de tiempo, total de transacciones, tiempo de respuesta.

c) Para dispositivos de salida: capacidad de impresión en líneas por minuto u hojas por minuto.

Para aquellos dispositivos no incluídos en los puntos anteriores, se deben identificar los parámetros clave para realizar las mediciones necesarias.

Los valores obtenidos deberán ser divididos por:

a) Departamento (Finanzas, Banca menudeo, Banca empresarial, etc.)

b) Aplicación (Cheques, Bursátil, Tarjeta de crédito, Ahorro, etc.)

c) Ambiente (en línea, lotes, desarrollo, etc.)

Las mediciones realizadas pueden ser tan generales o específicas como sea necesario dependiendo del nivel de profundidad con el que se desee realizar el estudio. Considerando que en el caso práctico del capítulo siguiente se trabaja en un ambiente I.B.M., se utiliza como herramientas de medición el SMF y el SLR.

3.1.3 Modelado (proyecciones)

El objetivo de esta actividad es predecir, en base a un modelo estadístico, el comportamiento del sistema con los diferentes datos de entrada y de esta manera proponer alternativas de solución para los requerimientos planteados.

Los modelos pueden ser sencillos o complejos, la elección de la técnica de modelado que se va a emplear depende de las características que muestra el sistema a través de las mediciones efectuadas durante un intervalo de tiempo. Evidentemente los costos se incrementan en proporción directa al grado de su complejidad.

Existen sistemas para el modelado en donde se define una estructura de *hardware* y *software*, realizando la simulación de los procesos de producción, obteniendo así la información que necesaria para el planteamiento de alternativas.

El éxito de la aplicación de los modelos depende de que tan confiable es la información manejada .

En la figura 12 se muestran las diversas técnicas de modelado para el desarrollo del estudio de la capacidad. Se deben considerar en la selección: la complejidad, los costos, la confiabilidad y la adaptación del modelo del sistema.

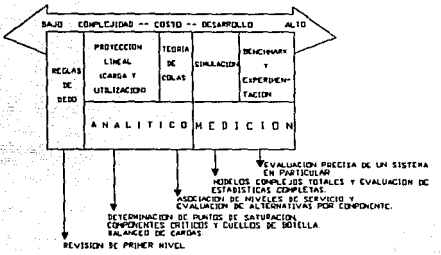


Fig. 12 Técnicas de Modelado.

3.1.4 Toma de decisiones

El objetivo de esta etapa es elegir, con base en las alternativas propuestas en la etapa anterior, aquella que más se adecúe a las necesidades de la empresa. Considerando los siguientes puntos:

1. Análisis de *costo-beneficio* de cada una de las alternativas propuestas.

2. Impacto en el servicio.

3. Adquisición de tecnología de punta, considerando la oportunidad en el servicio e imágen de la empresa.

4. Análisis de las necesidades de capacitación y/o contratación de personal.

5. Tiempo requerido para implantar adecuaciones.

El nivel ejecutivo (directores, subdirectores, y gerentes) juega un papel muy importante y es el de apoyar en todo lo que sea posible la aplicación del estudio de la capacidad. Es necesario contemplar que los cambios efectuados al sistema afectan directamente a los niveles de servicio comprometidos y a los criterios de saturación establecidos, por lo que se deben determinar los nuevos valores.

De lo expuesto con anterioridad, la secuencia de las etapas de la planeación de la capacidad se resumen en:

a) Recopilación de los niveles de servicio y análisis de rendimiento.

b) Modelado con la etapa anterior como entrada.

c) Decisión con los resultados del modelo.

El estudio madura conforme se cumplen las diferentes etapas, llegando hasta la toma de decisiones donde el ciclo se reanuda. La aplicación de la metodología debe ser flexible a los cambios requeridos por los usuarios, las nuevas estrategias gerenciales y los cambios que dicten las nuevas tecnologías.

CAPÍTULO 4
CASO PRÁCTICO

4.1 Antecedentes

Para la presentación de un caso práctico se eligió una institución bancaria, dado que el manejo óptimo de la información en una organización de esta naturaleza es directamente proporcional a los niveles de servicio que ofrece al cliente.

Un estudio de planeación de la capacidad de los recursos informáticos es una labor que conjunta información de diversas áreas dentro de la organización. Las áreas involucradas para el desarrollo del estudio son las siguientes:

La Dirección General

Integrada por los niveles directivos, es la responsable de fijar las guías de acción que rigen a la organización, incluyendo las políticas informáticas, con la finalidad de hacerla más competitiva en su ramo.

La Dirección de Finanzas

Es la responsable de la toma de decisiones para manejar los aspectos financieros y proveer los recursos económicos de acuerdo a las nececidades que van surgiendo durante el desarrollo de las actividades cotidianas de la organización.

La Dirección de Informática

Esta área es la responsable de mantener siempre a punto tanto a los recursos de *hardware* como al conjunto de aplicaciones (*software*) que se están explotando. A su vez se encuentra organizada por los siguientes departamentos:

Soporte técnico en *SOFTWARE*

Esta área es responsable de mantener a punto las aplicaciones instaladas y al sistema operativo. Es en primera instancia, la que atiende los problemas de rendimiento del equipo. Aporta al estudio, las herramientas de monitoreo de los diferentes recursos; obteniendo a través de éstas las estadísticas de utilización del sistema. Es importante señalar que también evalúa el impacto que estas herramientas tienen en el rendimineto del equipo.

Soporte técnico en *HARDWARE*

Esta área es la encargada de coordinar, conjuntamente con el proveedor el mantenimiento a los equipos de cómputo, conservándolos en condiciones óptimas de operación y verificando la correcta configuración física de los mismos.

Instalaciones físicas

Esta área debe proveer todas las necesidades para brindar el ambiente adecuado de operación en los sistemas de cómputo como son: instalación eléctrica, acondicionamiento del centro de cómputo, temperatura de operación, humedad relativa, cableado, etc.

Usuarios finales

Es la diversidad de usuarios que utilizan los recursos del sistema y que son responsables de que las operaciones brinden los beneficios esperados, así como de establecer los niveles de servicio requeridos (tiempo de respuesta, calidad de la información, etc.). En el estudio de la planeación de la capacidad, éstas son las áreas beneficiadas y deben reportar los resultados obtenidos en la implantación de las adecuaciones efectuadas. Ellos evalúan la efectividad de estos cambios al realizar sus funciones cotidianas, retroalimentando al proceso de planeación de la capacidad.

4.2 Creando la necesidad

La institución bancaria a la que se esta aplicando el caso práctico de la tesis, es una instalación de tipo 4, en la cual se cuenta con personal dedicado a estas labores.

El problema a resolver es el siguiente: Soporte técnico en *hardware* recibió una carta del proveedor donde se le indica que las unidades de disco tipo 3350 han quedado fuera de la póliza de mantenimiento debido a que solo se tendrán refacciones durante el año en curso ya que dichas unidades tienen siete años en el mercado.

En primera instancia esta área hace una evaluación apriori del problema y detecta que en las intalaciones se tienen 27 unidades de disco tipo 3350 (de un total de 41), lo cual equivale al 66% de las unidades de disco con las que cuenta el centro de cómputo. Soporte técnico hace constar esta situación a la dirección de informática.

La dirección de informática envía un memorandum a la dirección general, en el que se le pone al tanto de dicha situación y en el cual se manifiesta que será necesario tomar acciones que afectarán las operaciones normales del sistema para efectuar una evaluación y, que de acuerdo con los resultados obtenidos, se requerirá hacer una inversión.

La dirección general solicita a la dirección de informática que desarrolle un estudio completo sobre el impacto que esta situación tendrá dentro de la organización. Para esta labor, la dirección de informática canaliza la petición a su equipo de *planeación de la capacidad.*

4.3 El equipo de planeación de la capacidad.

Está integrado por personal debidamente capacitado en el manejo de las técnicas usadas para el desarrollo de un estudio de planeación de la capacidad en sistemas de cómputo, constituido por seis elementos y organizado de la siguiente manera:

Un líder responsable de coordinar las actividades del grupo y es quien mantiene informada de los resultados obtenidos a la dirección.

Un auxiliar administrativo que realiza tareas de organización y presentación de la información.

Cuatro ingenieros los cuales se encargan de depurar, analizar y modelar la información obtenida, así como de utilizar y evaluar las nuevas herramientas de monitoreo que existan en el mercado.

4.4 Desarrollo del estudio

De acuerdo con la figura 11 del capítulo 3, las etapas de análisis de los niveles de servicio y de análisis de rendimiento se arrancan de manera simultánea con la finalidad de poder contar con la información necesaria para desarrollar el modelo sobre el cuál se va a trabajar. Una vez recopilada toda la información, se elabora el modelo en el que se basa el planteamiento de las diferentes alternativas de solución y la toma de decisiones.

A continuación se desarrollan cada una de éstas actividades.

Análisis de rendimiento

El centro de cómputo cuenta en agosto de 1990 con la siguiente configuración de

Cantidad	Tipo de Equipo
27	Unidades de disco modelo 3350 marca IBM
12	Unidades de disco modelo 3380-D marca IBM
2	Unidades de disco modelo 3380-E marca IBM

unidades de disco:

Total: 41

Recopilación de los parámetros clave y análisis de la situación actual

Modelo	Capacidad (Mb)	Cantidad de Unidades	Total por modelo (Mb)
3350	635	27	17145
3380-D	2521	12	30252
3380-E	5042	2	10084

Capacidad de almacenamiento en Mb por modelo

Capacidad total instalada: 57481 Mb

Niveles de servicio con respecto a la capacidad:

Promedio de capacidad por unidad: 1401.98 Mb/unidad

Tomando una muestra de la actividad de los dispositivos, observamos una tasa de transacciones de I/O por hora λ = 107366 [Trans/hora]

$$\lambda = 29.8156 \text{ [Trans/seg]}$$

El tiempo promedio de servicio W_S = 0.0394 [seg], haciendo el análisis de colas tenemos que:

La tasa media de servicio es μ = 25.3807 [seg]

La intensidad de tráfico para peticiones de I/O es $u = \lambda/\mu$

$$\rho = u = 1.1747$$

Rendimiento relativo por unidad promedio con respecto a:

Unidades 3350 2.21 Veces

Unidades 3380-E 0.28 Veces

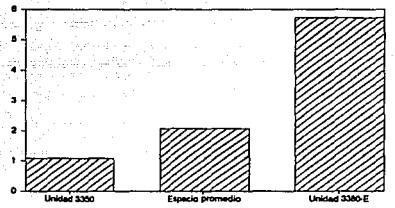


Fig. 13 Comparativo de niveles de servicio en capacidades por unidad (situación actual)

Transferencia de datos

Modelo	3350	3380-D	3380-E
Vel. de Transferencia (Mb/seg)	1.2	3	3
Cantidad de unidades	27	12	2
Vel. por modelo (Mb/seg)	32.4	36	6

Velocidad de transferencia de datos por modelo

Niveles de servicio con respecto a la velocidad de transferencia:

Velocidad promedio por unidad: 1.8146 Mb/seg

Rendimiento relativo por unidad promedio con respecto a:

Unidades 3350 1.51 Veces

Unidades 3380-E 0.60 Veces

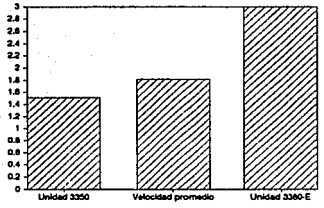


Fig. 14 Comparativo de nivel de servicio en velocidad de transferencia por unidad (situación actual)

Espacio físico

Modelo	3350	3380-D	3380-E
Espacio (m^2)	0.5844	0.8525	0.88
Cantidad de unidades	27	12	2
Espacio por modelo (m^2)	15.778	10.23	1.76
Mb/m^2	1086.6	2956	5727

Niveles de servicio con respecto al espacio físico requerido:

Mb por metro cuadrado promedio: 069.985019 Mb/m^2

Rendimiento relativo por unidad promedio con respecto a:

Unidades 3350 1.91 Veces

Unidades 3380-E 0.36 Veces

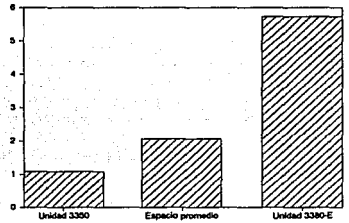


Fig. 15 Comparativo de nivel de servicio en espacio físico en Mb/m^2 (situación actual)

Análisis de los niveles de servicio y requerimientos

Nivel de servicio promedio general en discos

El nivel de servicio óptimo se toma con respecto a la unidad 3380-E, de esta manera podemos caracterizar la situación actual de la siguiente forma:

Capacidad por unidad	0.28 Veces
Velocidad de transferencia	0.60 Veces
Espacio físico	0.36 Veces
Nivel de servicio promedio general:	0.41 Veces

Mediciones del sistema.

Para el desarrollo de esta actividad, las herramientas de medición utilizadas son: SMF, JARS y SLR. La base de datos de SLR recaba información de los parámetros emitidos por SMF y JARS de tal manera que se pudieron generar los siguientes datos históricos:

Ambiente\Año	1987	1988	1989	1990
IDBX	1220	1600	2100	6450
TOTF	3310	3800	5100	15450
CMNX	1750	2400	2800	4650
PSIF	1050	2000	2060	5581
VARX	4750	5700	13600	15350
CNTF	800	1000	1000	2650
TSOF	2200	2650	7900	7350
TOTAL	15080	19150	34560	57481

Histórico de capacidades (Mb)

Definición de cada uno de los ambientes:

IDBX Ambiente de uso para la base de datos de clientes (consultas a estados de cuenta, etc.)

TOTF Ambiente de base de datos para tarjetas de crédito

CMNX Ambiente de contabilidad y nómina de la institución

PSIF Ambiente del sistema operativo

VARX Ambiente de producción (Aplicaciones, programas, etc.)

CNTF Ambiente complementario al ambiente de contabilidad y nómina

TSOF Ambiente de bibliotecas para operaciones de TSO, de desarrollo y pruebas

Ambiente\Año	1987	1988	1989	1990
IDBX	1086	1536	1470	1742
TOTF	2946	3230	4794	6798
CMNX	1330	1872	1260	1581
PSIF	1040	1700	1607	3739
VARX	3800	3990	7072	7061
CNTF	520	730	780	1855
TSOF	1430	1166	3555	3234
TOTAL	12151	14224	20538	26010

Histórico de utilización (Mb)

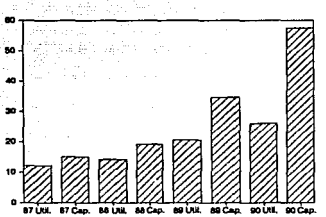


Fig. 16 Relación de utilización vs. capacidad (Histórico 1987-1990)

Ambiente\Año	1987	1988	1989	1990
IDBX	89.00%	96.00%	70.00%	27.00%
TOTF	89.00%	85.00%	94.00%	44.00%
CMNX	76.00%	78.00%	45.00%	34.00%
PSIF	99.00%	85.00%	78.00%	67.00%
VARX	80.00%	70.00%	52.00%	46.00%
CNTF	65.00%	73.00%	78.00%	70.00%
TSOF	65.00%	44.00%	45.00%	44.00%
TOTAL	80.43%	75.86%	66.00%	47.43%

Histórico de utilización en porcentajes

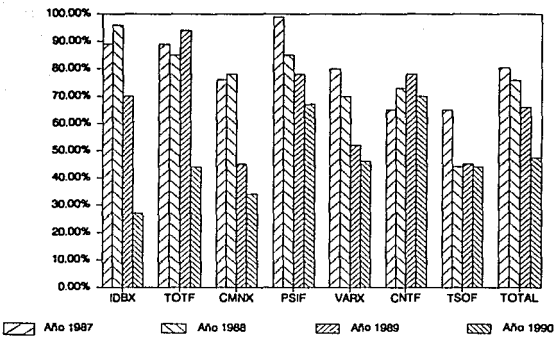


Fig. 17 Histórico de porcentajes de utilización: Años 1987-1990

El fabricante recomienda que para mantener un nivel adecuado de servicio los equipos nunca deben de exceder el 60% de ocupación, de aquí que el umbral máximo que se maneja es del 70% de ocupación y, por consiguiente, el criterio de saturación es el 60% de ocupación.

Del histórico de utilización podemos elaborar un gráfico en el cuál comparamos los niveles de utilización contra el criterio de saturación para evaluar el comportamiento de estos.

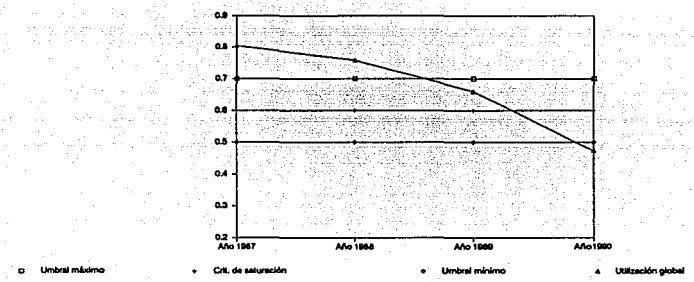


Fig. 18 Histórico de Utilización vs. criterio de utilización

Requerimientos actuales, modelado y alternativas de solución

Del análisis de la situación actual, y tomando en cuenta el hecho de que el porcentaje que representan las unidades 3350 del total que se tienen actualmente es muy alto, se propone la siguiente alternativa de modificación en la configuración de las unidades de disco:

- Eliminar todas las unidades 3350 debido a la obsolencia de las mismas

- Subsistuir éstas por 4 unidades 3380-D y 2 unidades 3380-E

- Evaluar estas modificaciones bajo los parámetros que fueron utilizados en la estimación de la situación actual

Cantidad	Tipo de Equipo
16	Unidades de disco modelo 3380-D marca IBM
4	Unidades de disco modelo 3380-E marca IBM

Configuración propuesta

Análisis general de la configuración propuesta.

Con la configuración propuesta, la capacidad instalada en Megabytes es:

Modelo	Capacidad (Mb)	Cantidad de Unidades	Total por modelo (Mb)
3380-D	2521	16	40336
3380-E	5042	4	20168

Total: 60505 Mb

Niveles de servicio con respecto a la capacidad:

Promedio de capacidad por unidad: 3025.2 Mb/unidad

Rendimiento relativo por unidad promedio con respecto a 3380-E:

Actual	Propuesto
0.28	0.6

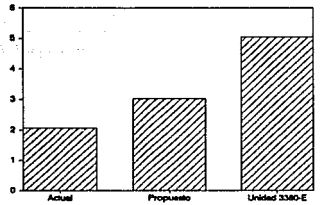


Fig. 19 Comparativo de nivel de servicio en capacidad por unidad (situación propuesta)

Análisis de rendimientos de la situación propuesta.

Modelo	Vel. de Transf. (Mb/seg)	Cantidad de unidades	Vel. por modelo (Mb/seg)
3380-D	3	16	48
3380-E	3	4	12

Total: 60 Mb/seg

Niveles de servicio con respecto a la velocidad de transferencia:

Velocidad promedio por unidad: 3 Mb/seg

Rendimiento relativo por unidad promedio con respecto a 3380-E:

Actual	Propuesto
0.60	1

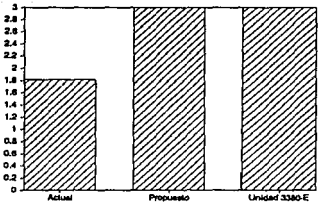


Fig. 20 Comparativo de niveles de servicio en vel. de transferencia por unidad (situación propuesta)

Espacio físico

Modelo	Espacio (m^2)	Cantidad de unidades	Espacio por modelo (m^2)	Mb/m^2
3380-D	0.8525	16	13.64	2956
3380-E	0.88	4	3.52	5727

Total: 17.16 m^2

Niveles de servicio con respecto al espacio físico requerido:

Mb por metro cuadrado promedio: 3525.87 Mb/m^2

Rendimiento relativo por unidad promedio con respecto a 3380-E

Actual	Propuesto
0.36	0.61

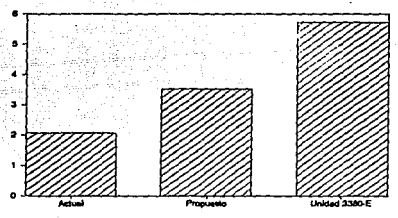


Fig. 21 Comparativo de nivel de servicio en espacio físico en Mb/m^2 (situación propuesta)

Nuestro nivel de servicio óptimo se toma con respecto a la unidad 3380-E, la situación propuesta entonces se puede resumir de la siguiente manera:

	Actual	Propuesto
Capacidad por unidad	0.28	0.60 Veces
Velocidad de transferencia	0.60	1.00 Veces
Espacio físico	0.36	0.53 Veces
Nivel de servicio promedio general	0.41	0.71 Veces

Niveles de servicio promedio generales en discos

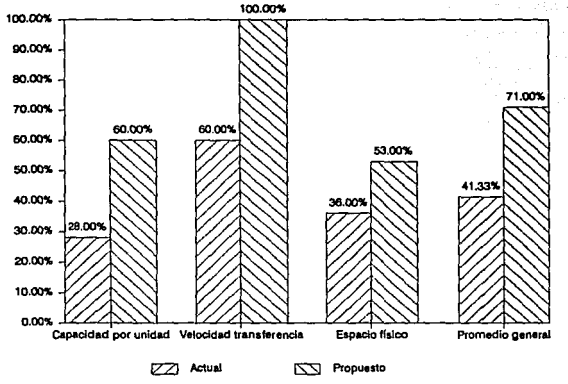


Fig. 22 Niveles de servicio propuestos comparados con la situación actual

Se puede hacer una rápida evaluación de la capacidad propuesta con respecto a la utilización. Observese la siguiente tabla:

Ambiente	Utilización	Cap. Actual	%Util. Actual	Cap. Prop.	%Util. Prop.
IDBX	1742	6450	27.00%	6789	25.65%
TOTF	6798	15450	44.00%	16263	41.80%
CMNX	1581	4650	34.00%	4895	32.30%
PSIF	3739	5581	67.00%	5875	63.65%
VARX	7061	15350	46.00%	16157	43.70%
CNTF	1855	2650	70.00%	2789	66.50%
TSOF	3234	7350	44.00%	7737	41.80%
Total	26010	57481	45.25%	60504	42.99%

Evaluación de Capacidad propuesta vs. Utilización

Evidentemente, en el *string* de discos se deben balancear las cargas por ambiente de tal manera que se tiene la siguiente distribución en función de lo que el ambiente requiere así como su porcentaje de participación en la capacidad del disco, la capacidad balanceada y el porcentaje de utilización:

Ambiente	Utilización	% Particp.	Cap. Balanceada	%Util.
IDBX	1742	6.70%	4051	42.99%
TOTF	6798	26.14%	15814	42.99%
CMNX	1581	6.08%	3678	42.99%
PSIF	3739	14.38%	8698	42.99%
VARX	7061	27.15%	16425	42.99%
CNTF	1855	7.13%	4315	42.99%
TSOF	3234	12.43%	7523	42.99%
Total	26010	100.00%	60504	42.99%

Balanceo de cargas por ambiente

Gráficamente la evaluación sobre la capacidad y el balanceo de las cargas se puede representar de la siguiente manera:

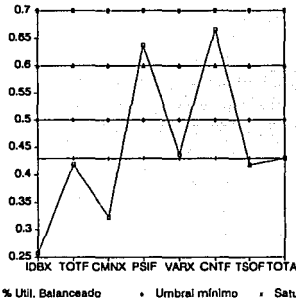


Fig. 23 Evaluación de capacidad y balanceo de cargas

Como puede observarse, el nivel de rendimiento que se obtiene al eliminar las unidades 3350 es bastante bueno (40% de utilización del string de discos). Ahora se deben considerar las cargas de trabajo proyectadas para los años 1991, 1992 y 1993 basándose en los históricos de utilización y en la siguiente fórmula para el cálculo de la tasa de crecimiento anual (TCA):

K_0 Dato inicial

$$K_1 = K_0 + K_0(i) = K_0(1+i)$$

$$K_2 = K_1 + K_1(i) = K_1(1+i) = K_0(1+i) = K_0(1+i)^2$$

$$K_3 = K_2 + K_2(i) = k_2(1+i) = K_0(1+i)(1+i)(1+i) = K_0(1+i)^3$$

Para encontrar el valor de i, conociendo K_0 y K_3

$$K_3 = K_0(1+i)^3$$

$$K_3/K_0 = (1+i)^3$$

despejando para i obtenemos la siguiente relación:

$$i = (K_3/K_0)^{1/3} - 1$$

de aquí que el porcentaje de crecimiento anual se puede calcular de la siguiente manera:

%Crecimiento anual = (%Utilización 1900 / %Utilización 1987)$^{1/3}$ - 1

de esta forma, observemos el comportamiento sobre el histórico de utilización y proyectando los porcentajes de crecimiento anual obtenemos los siguientes resultados:

Ambiente	1987	1988	1989	1990	TCA
IDBX	0.00%	41.46%	-4.30%	18.47%	17.06%
TOTF	0.00%	9.64%	48.42%	41.80%	32.15%
CMNX	0.00%	40.75%	-32.69%	25.48%	5.93%
PSIF	0.00%	63.54%	-5.48%	132.72%	53.22%
VARX	0.00%	5.00%	77.24%	-0.16%	22.94%
CNTF	0.00%	40.38%	6.85%	137.82%	52.80%
TSOF	0.00%	-18.46%	204.69%	-9.03%	31.26%
Total	0.00%	17.06%	44.39%	26.64%	28.88%

Porcentajes de crecimiento anual

De los porcentaje de crecimiento se proyectan los incrementos aplicando la TCA, de aquí se obtiene la siguiente tabla:

Ambiente\Año	1990	1991	1992	1993
IDBX	1742	2039	2386	2793
TOTF	6798	8983	11871	15687
CMNX	1581	1675	1774	1879
PSIF	3739	5729	8779	13451
VARX	7061	8681	10672	13120
CNTF	1855	2834	4331	6617
TSOF	3234	4245	5572	7314
Total	26010	33520	43200	55674

Proyecciones aplicando la tasa de crecimiento anual (Mb)

Una vez que se obtienen las proyecciones, se hace una evaluación de los crecimientos contra la capacidad del sistema:

Año	1990	1991	1992	1993
Proyección	26010	33520	43200	55674
Capacidad	60504	60504	60504	60504
% Utilización	42.99%	55.40%	71.40%	92.02%

Proyecciones vs. capacidad del sistema (Mb)

Graficamente, el comportamiento es de la siguiente manera:

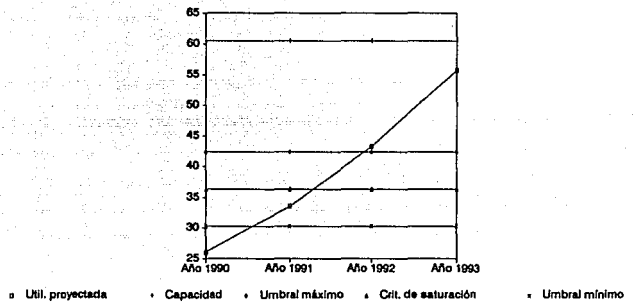


Fig. 24 Proyección de utilización de acuerdo a la tasa de crecimiento anual

De esta manera, se pueden proyectar las necesidades que se tendrán para 1991, 1992 y 1993. Así, para 1991 se tienen los siguientes requerimientos:

Niveles de Saturación\Año	1991	1992	1993
Umbral mínimo (Mb)	6537	25896	50844
Criterio de saturación (Mb)	-4637	11496	32286
Umbral máximo (Mb)	-12618	1210	19031

Requerimientos para 1991

Para satisfacer las necesidades en 1991 se necesita:

Cantidad	Modelo	Capacidad Unitaria	Capacidad Total
4	3380-D	2521	10084

Debido a que con esta propuesta se cubren las necesidades del '91, se proyectan los requerimientos para 1992 de la siguiente manera:

Año	1990	1991	1992	1993
Proyección	26010	33520	43200	55674
Capacidad	60504	70588	70588	70588
% Utilización	42.99%	47.49%	61.20%	78.87%

y se obtienen los requerimientos para 1992:

Niveles de Saturación\Año	1992	1993
Umbral mínimo (Mb)	15812	40760
Criterio de saturación (Mb)	1412	22202
Umbral máximo (Mb)	-8874	8947

Requerimientos para 1992

Ahora, para cubrir las necesidades de 1992 se necesita:

Cantidad	Modelo	Capacidad Unitaria	Capacidad Total
4	3380-E	5042	20168

De aquí que el comportamiento para 1993 sea:

Año	1990	1991	1992	1993
Proyección	26010	33520	43200	55674
Capacidad	60504	70588	90756	90756
% Utilización	42.99%	47.49%	47.60%	61.34%

por lo cual se obtiene la siguiente tabla:

Niveles de Saturación\Año	1993
Umbral mínimo (Mb)	20592
Criterio de saturación (Mb)	2034
Umbral máximo (Mb)	-1121

Requerimientos para 1993

Y para satisfacer las necesidades en 1993 se necesita:

Cantidad	Modelo	Capacidad Unitaria	Capacidad Total
4	3380-E	5042	20168

Cubriendo todas las necesidades se propone la siguiente tabla de referencia:

Año	1990	1991	1992	1993
Proyección	26010	33520	43200	55674
Capacidad	60504	70588	90756	110924
% Utilización	42.99%	47.49%	47.60%	50.19%

y graficamente, el comportamiento es el siguiente:

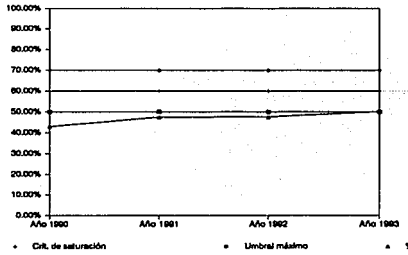


Fig. 25 Proyección de utilización en base a las recomendaciones

De esta manera, la configuración propuesta para soportar las proyecciones hasta 1993 es la siguiente:

Modelo	3380-D	3380-E
Capacidad (Mb)	2521	5042
Cantidad de unidades	20	12
Total por modelo (Mb)	50420	60504
Total	110924	Mb

y , en resumen, los crecimientos proyectados son los siguientes:

Año	1990	1991	1992	1993
Capacidad (Mb)	60504	70588	90756	110924
Vel. Transferencia (Mb/seg)	60	66	78	90
Espacio físico (m^2)	17.16	20.57	24.09	27.61
Número de Unidades	20	24	28	32

y su interpretación gráfica para cada uno de estos conceptos se da a continuación:

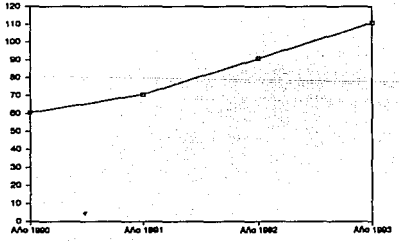


Fig. 26 Capacidad proyectada en Megabytes

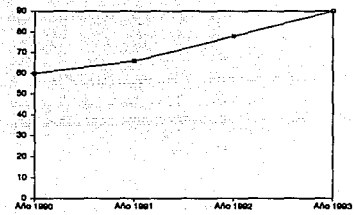


Fig. 27 Velocidad de transferencia en Megabytes por segundo (Mb/seg)

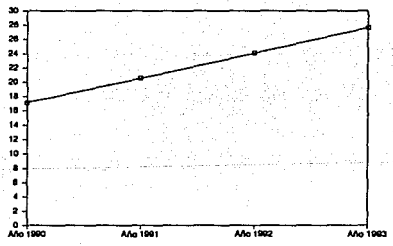


Fig. 28 Espacio físico proyectado en metros cuadrados

Finalmente, los niveles de servicio esperados se pueden resumir de la siguiente manera:

Año	1990	1991	1992	1993
Capacidad (Mb)	60.00%	58.33%	64.29%	68.75%
Vel. Transferencia (Mb/seg)	100.00%	100.00%	100.00%	100.00%
Espacio físico (m^2)	61.57%	59.92%	65.78%	70.15%

cuyo comportamiento está representado por las figuras 29, 30 y 31

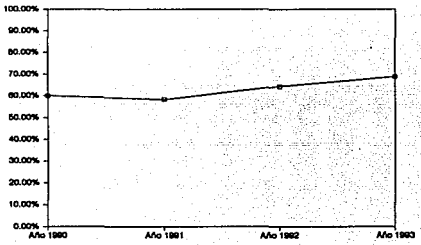


Fig. 29 Capacidad (Mb)

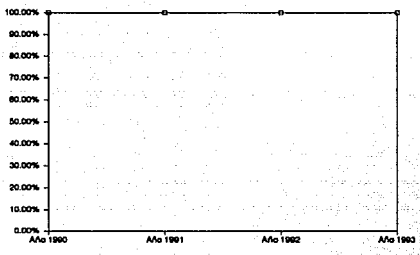


Fig. 30 Velocidad de transferencia (Mb/seg)

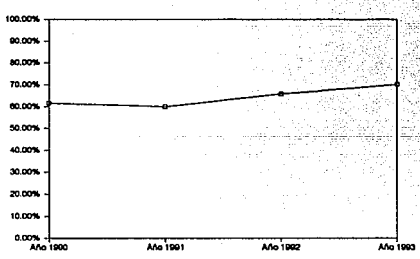


Fig. 31 Espacio físico en metros cuadrados

CONCLUSIONES

El estudio de planeación de la capacidad, o *capacity planning* como se le conoce en el medio, es una metodología destinada a la *evaluación* y *proyección* de los crecimientos y niveles de servicio que en todo centro de cómputo se deben considerar para brindar una óptima atención a los usuarios.

Es evidente que un estudio de esta naturaleza requiere de personal adecuadamente capacitado, lo cual implica la realización de inversiones tanto en el rubro de entrenamiento como en el de actualización.

La interpretación de la información generada por las herramientas de monitoreo que haga el analista de la capacidad debe ser natural, es decir, que debe comprender el significado de los resultados generados de tal manera que se puedan detectar los puntos de alertamiento antes de que se presenten.

Por esta razón, es necesario que a los niveles directivos de la organización, independientemente del ramo al que se dedique, estén concientes de la importancia que tiene el poder contar con un equipo que pueda resolver las contingencias que surjan de la manera más adecuada.

Esta labor de concientización es hasta cierto punto compleja si en la organización no se le dá la debida importancia al área informática, y depende en gran medida de la habilidad y experiencia del gerente (o director) de informática el crear la necesidad.

Es indiscutible que un adecuado manejo de la información permite a cualquier organización ser más competitiva, por ello la tendencia general se orienta hacia la automatización de funciones que tradicionalmente se desarrollaban manualmente.

Cuando una organización se decide a adquirir equipo de cómputo es porque necesita mantener un control adecuado y tal vez hasta estricto de *su* información.

Para la elección del equipo de cómputo adecuado a las necesidades de la organización se deben realizar pruebas de *benchmark* para evaluar aspectos como tiempos de respuesta, alternativas de crecimiento, portabilidad, etc., de tal manera que se pueda elegir la configuración de *hardware* apropiada.También es necesario evaluar y definir los perfiles del personal que va a trabajar con el equipo, incluyendo a los candidatos para el departamento de *capacity planning*.

La metodología propuesta en este trabajo es tan solo uno de los múltiples enfoques que se pueden dar a un estudio de *capacity planning*: los modelos matemáticos se pueden complicar tanto como sea necesario.

Las herramientas de monitoreo usadas en en el desarrollo del caso práctico dependen en gran medida del sistema operativo con el cuál se trabaja y no son únicas, por ejemplo, en los sistemas *UNIX* existe el subsistema de auditoría (*audit subsystem*) que permite monitorear al *Kernel*, al *device driver*, y a las facilidades de *reducción/análisis* de datos. Siempre debe considerarse el efecto que pueden tener sobre el rendimiento del sistema, de tal manera que la asignación de recursos sea balanceado con respecto al consumo general que los diferentes ambientes que se esten procesando en el computador.

El objetivo de este trabajo fue el presentar de manera general una metodología que puesta en práctica permite una racional explotación y adquisición de recursos informáticos, así como el de motivar a los profesionales de la informática al desarrollo de estudios de esta naturaleza.

Una recomendación en el desarrollo del proceso de planeación de la capacidad es hacerlo tan simple como sea posible, guardando un mínimo de datos a ser analizados.

En palabras de L.K. Bronner la planeación de la capacidad se define como:"Planeación de la capacidad es un proceso desarrollado para proveer un método sistemático para entender y predecir la capacidad de producción de proceso de datos... pronosticando las cargas futuras de los usuarios, determinando los requerimientos de capacidad de cómputo en forma efectiva y eficiente, administrando los recursos (gente, *hardware* y *software*) para lograr los objetivos de servicio a los usuarios."

BIBLIOGRAFIA

R.M. AMSTRONG, "Capacity Planning Overview", IBM Washington System Center, TechnicalL Bulletin, July 1986.

DR. L. BRONNER, "Basic Hand Analysis", IBM, December 1983.

R.M. AMSTRONG, "Management Methodology", IBM, August 1982.

DR. L. BRONNER, "Capacity Planning an Introducction", IBM GG229001, Technical Bulletin, January 1977.

DR. L. BRONNER, "Capacity Planning Implementation", IBM GG229015, Technical Bulleting, January 1979.

OS/VS2 MVS RESOURCE, "Reference and User Guide - Measurement Facility (RMF)", IBM SC28-0922-3, 1980

OS/VS2 MVS, "System Programming Library (SMF)", IBM GC 28-0706-1, July 1977.

JOHNSON SYSTEMS INC. "JARS, OS, Job Accounting Report System User Reference Manual", JS0002-004, May 1981.

"IBM 3380, Direct Access Storage Description and User Guide", IBM GA26-1664-1, December 1981.

"IBM 3380, Storage Control Description", IBM GA26-1661-3, May 1981.

JOHNSON SYSTEM GUIDE, "JARS, OS, Job Accounting Report System, General Information Manual", July 1981.

JHONSON SYSTEM GUIDE, "JARS, OS, Job Accounting Report System, Working Set Guide" JS009-004, June 1991.

J.C. COOPER, "A Capacity Planning Methodology", IBM Systems Journal, Vol 90 No. 1, 1980

A.O. ALLEN, "Elements of Queuing Theory for System Design", IBM Systems Journal, Vol. 40 No. 2, 1975.

J.M. JENKINGS and P.C. HOWARD, "Measuring System Capacity", EDP Performance Review 5, No. 4, April 1977.

ANEXOS

ANEXO 1

NOTACION Y DEFINICION DE TEORIA DE COLAS

$A(t)$ Función distribución de tiempos de interarribo $A(t) = P(r \leq t)$

c Número de servidores idénticos.

D Distribución determinística (constante) de tiempos de interrarribo o servicio.

E_k Distribución Erlang-k de tiempo de interarribo o de servicio.

$E[N_q \mid N_q > 0]$ Media (esperada o promedio) de longitud de la cola de colas no vacías.

$E[q \mid q > 0]$ Tiempo promedio en la cola para colas no vacías.

FCFS Disciplina de teoría de colas: primero en entrar primero en ser atendido.

FIFO Disciplina de teoría de colas: primero en entrar primero en salir.

G Probabilidad general de distribución del tiempo de servicio, usualmente independiente.

GI Tiempo independiente de interarribos general, algunas veces usado para describir la distribución del tiempo de servicio.

K Máximo número permitido en el sistema de colas, incluyendo a aquellos que están en espera como a los que están siendo atendidos.

L $E[N]$, valor promedio en el sistema de colas (estado estable).

L_q $E[N_q]$ valor promedio en la cola, sin incluir a quienes están siendo atendidos (estado estable).

LCFS Disciplina de teoría de colas: último en llegar primero en ser atendido

LIFO Disciplina de teoría de colas: último en llegar primero en salir

λ Razón promedio de llegadas al sistema de colas.

M Distribución exponencial de interarribos o tiempo de servicio.

μ Tasa promedio de servicio por servidor.

N(t) Variable aleatoria para describir la longitud en la cola para cualquier valor de *t* (tiempo).

N Variable aleatoria para describir el número en la cola (condiciones de estado estable).

$N_q(t)$ Variable aleatoria para describir la longitud en la cola (excluyendo a quienes están siendo atendidos) para cualquier valor de t.

N_q Variable aleatoria para describir la longitud en la cola para condiciones de estado estable.

$N_s(t)$ Variable aleatoria para describir el número de usuarios que están siendo atendidos para cualquier valor de t.

N_s Variable aleatoria para describir el número de usuarios que están siendo atendidos para condiciones de estado estable.

$p_n(t)$ Probabilidad de que haya n usuarios en el sistema para cualquier valor de t.

p_n Probabilidad de que haya n usuarios en el sistema para condiciones de estado estable.

PRI Disciplina de prioridad en la cola.

q Variable aleatoria para describir el tiempo que un usuario pasa en la cola antes de ser antendido.

RSS Selección aleatoria para disciplina de servicio en la cola.

ρ Utilización del servidor $\rho = \lambda/c\mu$.

s Variable aleatoria para describir el tiempo de servicio.

SIRO Servicio en orden aleatorio (idéntico a RSS)

τ Variable aleatoria para describir los tiempos de interarribos.

u Intensidad de tráfico.

w Variable aleatoria que describe el tiempo total que un usuario pasa en el sistema.

$W(t)$ Función distribución para w, $W(t) = P[w \leq t]$.

W $E[w]$, tiempo promedio en el sistema.

W_q E[q], tiempo promedio de espera en la cola.

$W_s(t)$ Función distribución del tiempo de servicio $W_s(t) = P[q \leq t]$

W_s $E[s]$, tiempo promedio de servicio, $W_s = 1/\mu$.

ANEXO 2

Relaciones usadas en los modelos de teoría de colas

$u = E[s]/E[\tau] = \lambda E[s] = \lambda/\mu$

$\rho = u/c$

$w = q + s$

$W = E[w] = E[q] + E[s] = W_q + W_s$

$N(t) = N_q(t) + N_s(t)$

$N = N_q + N_s$

$L = E[N] = \lambda W = E[N_q] + E[N_s]$

$L_q = E[N_q] = \lambda W_q$

ANEXO 3

Fórmulas de estado estable para los modelos de colas tipo M/M/1.

$P_n = P[N = n] = (1 - \rho)\rho^n$

$P[N=n] = \Sigma P_k = \rho^n$ donde n = 1,2,3,... y k va desde n hasta infinito en la sumatoria

$L = E[N] = \rho/(1-\rho)$

$s^2_N = \rho/(1-\rho)^2$

$W(t) = 1 - e^{-t/E[w]}$

$W = E[w] = 1/\mu(1-\rho)$

$L_q = \rho^2/1-\rho$

$W_q = \rho E[s]/(1-\rho)$

$E[q \mid q > 0] = E[s]/(1-\rho)$

$\pi_W(90) = 2.3E[w]$

$\pi_q(90) = E[w]\log(20\rho)$