Capítulo 1: Introducción

En este capítulo se da el contexto bajo el cual se desarrolla todo el resto de esta tesis. Inicialmente, se hace una breve descripción de las aplicaciones numéricas y de la relación existente entre la multiplicación de matrices y las demás operaciones y métodos numéricos para las operaciones provenientes del álgebra lineal en general.

Desde el punto de vista del hardware de procesamiento, se describen las redes locales de computadoras como una alternativa más de cómputo paralelo y se detalla brevemente su relación con las demás plataformas de procesamiento paralelo que se utilizan actualmente. Avanzando sobre este punto, se identifican con mayor nivel de detalle los costos asociados con el cómputo paralelo en las redes de computadoras instaladas, comparándolos con los de las demás arquitecturas de cómputo paralelo.

Finalmente, se hace un resumen de los objetivos, aportes y método utilizado en el desarrollo de esta tesis así como se explica la organización del contenido.

1.1 Aplicaciones y Arquitecturas Paralelas

Las arquitecturas de procesamiento paralelo han sido extensiva e intensivamente utilizadas en numerosas aplicaciones, y el área de los problemas numéricos ha sido normalmente el punto de partida donde se ha estudiado y aprovechado la posibilidad de cómputo paralelo. Se han desarrollado numerosos métodos de solución para los problemas numéricos, la mayoría de ellos con una clara orientación hacia al menos dos sentidos [1] [59] [101] [102] [84]:

- Estabilidad numérica cuando se resuelven con aritmética que involucra algún tipo de error. La aritmética que se suele asumir es la que normalmente se denomina “de punto flotante”.
- Implementación directa o con costo mínimo en algún lenguaje de programación para ser resuelto en una computadora.

Desde el punto de vista de quien tiene un problema numérico a resolver, la utilización de una computadora no es nada más (y nada menos) que una forma de obtener lo que necesita. Más aún, si la computadora es paralela o no, o la forma en que una computadora obtiene los resultados correctos tampoco es de importancia excepto por el tiempo que debe esperar para utilizar el resultado. De hecho, el término *supercomputadora* (*supercomputer*) o la denominación *computadora de alto rendimiento* (*high-performance computer*) son utilizados desde hace mucho tiempo y refleja esta realidad: lo importante es la velocidad, si eso se logra con procesamiento paralelo entonces se lo utiliza [113].

El área de problemas provenientes del álgebra lineal tradicionalmente ha aprovechado el rendimiento que proporcionan las arquitecturas de cómputo (paralelo) disponibles. Quizás unos de los esfuerzos más significativos de los investigadores en esta área ha sido enfocado hacia el desarrollo de una biblioteca de rutinas que se consideran de vital importancia. Este esfuerzo ha dado como resultado la biblioteca denominada LAPACK (Linear Algebra PACKage) [7] [8] [LAPACK]. Más específicamente relacionado con el cómputo paralelo se ha desarrollado ScaLAPACK (Scalable Linear Algebra PACKage o simplemente Scalable LAPACK) [21] [27] [29] [ScaLAPACK].

Dentro de las aplicaciones del álgebra lineal se han identificado un conjunto de operaciones o directamente rutinas de cómputo que se consideran y a partir de las cuales, por ejemplo, se puede definir todo LAPACK. Tales rutinas se han denominado BLAS (Basic Linear Algebra Subroutines) y tanto para su clasificación como para la identificación de requerimientos de cómputo y de memoria de cada una de ellas, se las divide en tres niveles: nivel 1, nivel 2 y nivel 3 (Level 1 o L1 BLAS, Level 2 o L2 BLAS y Level 3 o L3 BLAS). Desde el punto de vista del rendimiento, las rutinas de nivel 3 (L3 BLAS) son las que se deben optimizar para obtener rendimiento cercano al óptimo de cada máquina y de hecho, muchas empresas de microprocesadores estándares proveen bibliotecas BLAS con marcado énfasis en la optimización y el consiguiente rendimiento de las rutinas incluidas en BLAS de nivel 3.

A partir de la definición misma de las rutinas del nivel 3 de BLAS y más específicamente a partir de [77], la multiplicación de matrices es considerada como el pilar o la rutina a partir de la cual todas las demás incluidas en este nivel de BLAS se pueden definir. Quizás por

esta razón y/o por su simplicidad, la mayoría de los reportes de investigación en esta área de procesamiento paralelo comienza por el "problema" de la multiplicación de matrices en paralelo y existen numerosas propuestas y publicaciones al respecto [20] [144] [57] [30] [142] [26]. Expresado de otra manera, al optimizar la multiplicación de matrices de alguna manera se optimiza todo el nivel 3 de BLAS y por lo tanto se tendrían optimizadas la mayoría de las aplicaciones basadas en álgebra lineal y que dependen de la optimización de las rutinas que llevan acabo las operaciones provenientes del álgebra lineal. Aunque esta optimización no sea necesariamente directa, sí se puede afirmar que el tipo procesamiento que se debe aplicar para resolver la multiplicación de matrices es muy similar al del resto de las rutinas definidas como BLAS de nivel 3 e incluso muy similar también a los problemas más específicos que se resuelven recurriendo a operaciones del álgebra lineal. En este sentido, es muy probable que lo que se haga para optimizar la multiplicación de matrices (en paralelo o no) sea utilizable y/o aprovechable en otras operaciones. En términos de problema a resolver o de procesamiento a llevar a cabo en paralelo, esta tesis se orienta específicamente hacia la multiplicación de matrices con algunos comentarios hacia la generalización dada su importancia y representatividad en el área de álgebra lineal.

La Figura 1.1 resume esquemáticamente la relación de la multiplicación de matrices con los problemas numéricos en general. Dentro de los problemas numéricos que se resuelven computacionalmente existe un área relativamente bien definida que es la que corresponde al álgebra lineal. Como se ha explicado anteriormente, en esta área se ha definido una biblioteca estándar *de facto* que se ha denominado LAPACK. Un conjunto básico de rutinas de cómputo bien definidas puede ser utilizado para construir todo LAPACK: BLAS. Este conjunto básico de rutinas se ha dividido en tres niveles con el objetivo de identificar más claramente cuáles de ellas son las más apropiadas para la obtención del mejor rendimiento posible en una arquitectura de cómputo dada. Las rutinas del nivel 3 de BLAS o L3 BLAS son las que se deben analizar e implementar con mayor énfasis en la optimización y todo este nivel de BLAS puede ser definido en función de la multiplicación de matrices.

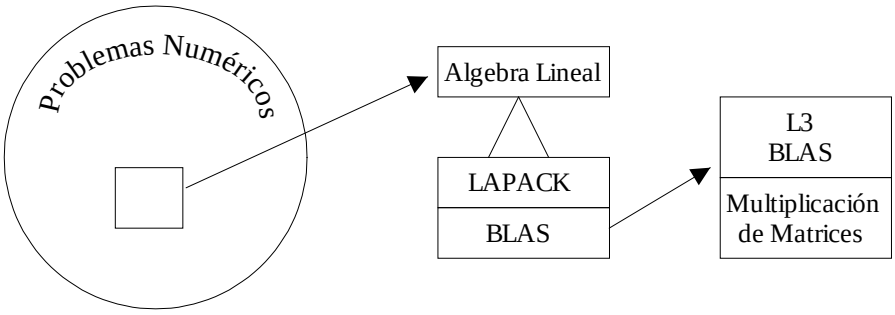

Figura 1.1: La Multiplicación de Matrices dentro de los Problemas Numéricos.

Las computadoras secuenciales tienen varios límites respecto de la capacidad máxima de procesamiento por lo que se recurre al procesamiento paralelo en general [2]. En realidad, las dos formas clásicas de obtener mayor rendimiento de cómputo se han explotado simultáneamente y de forma complementaria:

- Aumento de la velocidad de cómputo secuencial de un procesador
- Aumento de las unidades funcionales y/o procesadores que se pueden utilizar de forma

simultánea.
De hecho, se puede afirmar que la principal razón de ser del procesamiento paralelo es el rendimiento.

El área de cómputo paralelo ha evolucionado en muchos sentidos, desde el hardware de cómputo hasta los algoritmos y bibliotecas especializados en determinadas tareas. Desde el punto de vista del hardware de procesamiento ha habido una gran variación en cuanto a los objetivos de diseño y costos asociados. De hecho, la gran mayoría de las empresas creadoras de diseños y máquinas paralelas consideradas como muy destacables en su época han dejado de existir o han sido absorbidas por otras corporaciones. Solamente a modo de ejemplo se pueden mencionar empresas como Cray o Thinking Machines (creadora de la Connection Machine). La que puede considerarse como una "gran lección" al respecto es: los costos asociados a una computadora (paralela) específica no siempre se pueden afrontar y existen alternativas menos costosas con resultados similares o por lo menos apropiados para los problemas que se necesitan resolver. Los costos de computadoras (paralelas) específicas están relacionados con [113]:

- Diseño. La computadora paralela se desarrolla en términos de hardware casi desde cero. Por lo tanto se debe involucrar a especialistas con una alta capacidad y experiencia.
- Tecnología de construcción. Tanto en términos de materiales como de mecanismos de fabricación, los costos son mucho más altos que los involucrados en cualquier computadora de fabricación y venta masiva. La diferencia se cuantifica en varios órdenes de magnitud.
- Instalación. Las instalaciones tanto del lugar físico como del ensamblado de la computadora misma involucra muchas características específicas que tienen un muy alto costo. Esto involucra desde el personal que realiza la instalación hasta las condiciones de temperatura, por ejemplo.
- Mantenimiento del hardware. Siempre ha sido y es un porcentaje del costo total de una máquina y por lo tanto está en el mismo o similar orden de magnitud.
- Software. Tanto el software de base como el de desarrollo y ejecución de programas que es tanto o más importante, son específicos. En este caso, la especificidad del hardware se combina con la propia complejidad de un sistema operativo o de un ambiente de desarrollo y depuración de aplicaciones paralelas. En este caso el costo no solamente involucra tiempo y dinero sino que además hace más probables los errores en el software desarrollado para estas computadoras.
- Operación. Tanto el personal de producción de software como el de monitoreo y sintonización del sistema debe ser específicamente capacitado.

Por otro lado, el hardware básico de procesamiento que se utiliza masivamente en las computadoras de escritorio ha incrementado su capacidad también en órdenes de magnitud a la vez que ha reducido sus costos a los usuarios finales. Tanto el alto costo de las computadoras paralelas como la reducción de costo y el incremento de capacidad del hardware de procesamiento que se puede denominar *estándar* y de uso masivo ha llevado a que las computadoras paralelas actuales tengan una fuerte tendencia a incorporar este hardware estándar como básico. Los beneficios tienen una fuerte relación con la reducción de todos los costos mencionados y además se ha probado que es factible.

A medida que la cantidad de computadoras instaladas en una misma oficina y también a nivel de las instituciones, se ha incrementado, también se incrementaron las propuestas y el

estudio de los problemas y soluciones en lo que se ha denominado redes locales. De manera simultánea y desde distintos puntos de vista, la existencia de una cantidad relativamente grande de computadoras interconectadas en red ha dado lugar a los sistemas operativos distribuidos y a la utilización masiva de las redes locales como formas útiles de resolver problemas en ambientes acotados respecto de usuarios y aplicaciones. Sin embargo, a medida que estas redes locales se han aumentado pronto se identificaron varias alternativas de procesamiento que son posibles y de muy bajo costo:

- Utilización de los períodos de inactividad de las máquinas interconectadas en redes locales [90].
- Utilización de más de una computadora interconectada en red para resolver un problema, haciendo uso de los conceptos de cómputo paralelo [9].

A partir de la interconexión masiva de las redes locales a Internet también se han introducido las ideas de “Internet Computing” y actualmente se está impulsando con fuerza la idea de “Grid Computing” como una forma más amplia de compartir recursos en general [54] [75] [53] [55] o “Metacomputing” [24] [13].

La idea de cómputo paralelo en las redes locales de alguna manera es posible a partir de la interconexión misma de las computadoras. Sin embargo, es evidente que se deben resolver varios problemas para que esta forma de cómputo paralelo se utilice de manera confiable y con rendimiento aceptable. Si bien es bastante difícil definir el término *aceptable*, se tienen dos ideas subyacentes al respecto que son muy importantes:

- Aprovechamiento máximo de los recursos disponibles, especialmente de los recursos de cómputo (procesadores).
- Tiempo de procesamiento necesario para resolver un problema.

Puede ser posible que el tiempo de procesamiento para resolver un problema sea aceptable sin que se utilicen al máximo los recursos disponibles, pero dado que usualmente las computadoras interconectadas en una red local son las de menor rendimiento del mercado, la probabilidad de que esto suceda es bastante baja en general.

Una alternativa de cómputo paralelo con computadoras estándares de escritorio que ha probado ser muy satisfactoria en algunas áreas es la que se asocia con las instalaciones o con el proyecto Beowulf [19] [103] [99] [143] [111] [BEOWULF]. Aunque casi cualquier red local usada para cómputo paralelo se puede considerar como una instalación Beowulf [37], existe cierto consenso con respecto a que:

- Las computadoras de una instalación Beowulf son PCs homogéneas, con a lo sumo una computadora dedicada a la administración del sistema completo. De hecho, estas instalaciones son creadas para cómputo paralelo y en principio no tiene sentido instalar múltiples tipos de computadoras (heterogéneas).
- La red de interconexión básica es Ethernet [73] de 100 Mb/s y el cableado debería incluir la utilización de *switches* de interconexión que son capaces de aislar las comunicaciones entre pares de computadoras. De todas maneras, mejores redes de interconexión nunca son descartadas aunque siempre se tiende al menor costo.

Con el objetivo de estudiar y distinguir diferentes características se han propuesto múltiples clasificaciones de las arquitecturas de procesamiento [50] [51] [33]. Desde el punto de vista de capacidad y costo, las plataformas de cómputo paralelo actuales se pueden organizar de mayor a menor en una pirámide tal como lo muestra la Figura 1.2.

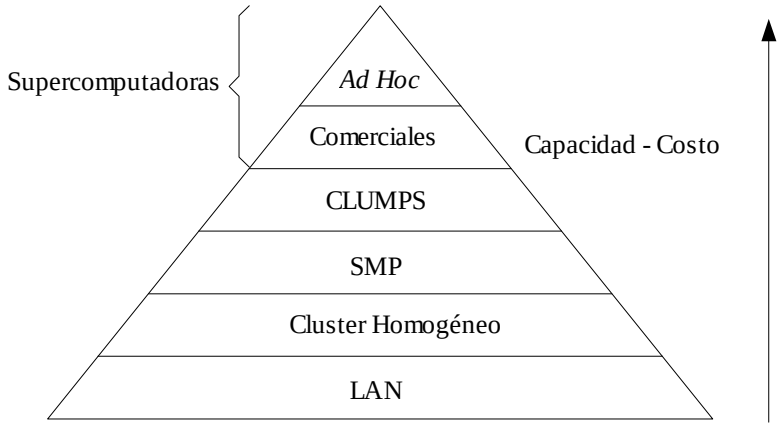

Figura 1.2 Una Clasificación de Plataformas de Cómputo Paralelo.

Ad Hoc. Computadoras paralelas planificadas/construidas específicamente bajo programas o aplicaciones definidas con anticipación. Dos de los más conocidos sobre el final de la década 1990-2000 son los denominados Accelerated Strategic Computing Initiative (ASCI) del Departamento de Energía de Estados Unidos de América [61] y EA (Earth Simulator) de varios departamentos y/o centros de investigación de Japón [EA] [139] [140]. Están entre las de mayor potencia de cálculo absoluta, tal como se reporta en [86] [TOP500]. Son las clásicas computadoras paralelas construidas *ad hoc* y se les asignan nombres como [TOP500]: Earth Simulator, ASCI Q, ASCI White, ASCI Red, ASCI Blue-Pacific, ASCI Blue Mountain. Llegan a varios Tflop/s (*Tera*flop/s, o 10^{12} operaciones entre números de punto flotante por segundo) y son proporcionalmente costosas (en realidad, tan costosas como construidas a medida). De hecho, pueden considerarse como las *únicas* computadoras paralelas construidas a medida de una aplicación o conjunto de aplicaciones en la actualidad.

Comerciales. Son casi específicamente dedicadas a cómputo científico en general y la mayoría de los centros con grandes requerimientos de cálculo tienen alguna/s. Suelen construirse con hardware básico que es promocionado como "escalable" dado que se pueden agregar procesadores, memoria, discos, etc. según las necesidades y siempre dentro de un rango definido. Pueden considerarse como ejemplos de esta clase: IBM SP, Compaq AlphaServer, Hitachi SR, SGI Origin, Cray T3E y las computadoras con procesadores vectoriales (Fujitsu).

CLUMPs: Clusters de SMP (Symmetric MultiProcessing) [12] o SMPs conectadas en red, no parece estar muy explorado en general y en/para cómputo paralelo en particular porque es bastante nuevo, al menos no hay muchas publicaciones al respecto. Desde el punto de vista de la paralelización es muy importante la combinación de modelos:

- De memoria compartida en un SMP que puede ser considerado como MIMD fuertemente acoplado.
- De pasaje de mensajes o al menos MIMD de memoria distribuida dada por las utilización de más de una máquina SMP interconectada por una red.

SMP: Symmetric MultiProcessing, suelen estar directamente orientadas a almacenamiento y procesamiento de grandes volúmenes de datos en disco y/o "servidores web". No hay

muchos reportes de su utilización en aplicaciones científicas y sí hay una gran cantidad de publicaciones de su utilización/utilidad en el campo de recuperación de información almacenada en disco. Muchas empresas directamente las promocionan como “servidores”. Su utilidad en las aplicaciones paralelas en general y particularmente en aplicaciones científicas es inmediata. Aunque tienen límites en cuanto a escalabilidad y más precisamente en cuanto a la cantidad máxima de procesadores que pueden tener, ponen a disposición del usuario una cantidad relativamente grande de procesamiento sobre una única memoria compartida, lo cual hace inmediata la utilización de los algoritmos paralelos orientados a los multiprocesadores.

Clusters Homogéneos: Conjunto de computadoras dedicadas a cómputo paralelo. Estaciones de trabajo “clásicas” (Sun, SGI), o PCs homogéneas. Desde el punto de vista técnico pueden considerarse como las instalaciones Beowulf que nacieron como un conjunto de PCs conectadas en una red local (normalmente Fast Ethenet hasta ahora). Desde el inicio mismo se hace una gran inversión en la red de comunicaciones (básicamente a través de switches en el cableado de la red Ethernet. Se han desarrollado múltiples herramientas tanto para administración de todo el sistema paralelo como para la producción de software paralelo. Los demás clusters homogéneos (con estaciones de trabajo “clásicas” no tienen muchas diferencias con las instalaciones Beowulf, pero en [99] [111] se las excluye casi sin discusión por no tener Linux como sistema operativo en cada nodo básico de cómputo. El costo de hardware es bastante mayor que el de de los sistemas Beowulf por la relación de costos entre una PC y una estación de trabajo Sun o SGI por ejemplo. En términos de rendimiento y confiabilidad la diferencia no es tan clara.

Por la relación de costo de hardware existente entre una PC y una estación de trabajo “clásica” es muy difícil establecer cuál configuración es más potente. En general se cumple que a igual costo, la capacidad de cálculo (al menos la suma de capacidad de los procesadores) es mayor en los sistemas Beowulf. Sin embargo, la diferencia es aún más difícil de cuantificar y establecer *a priori* cuando se tiene en cuenta el costo de las herramientas, confiabilidad, etc. de cada uno de los sistemas.

LAN: redes locales de computadoras instaladas y que se pueden aprovechar para cómputo paralelo. Cada computadora tiene uno o un conjunto de usuarios que la deja/n disponible para cómputo paralelo por determinados períodos de tiempo. Otra de las alternativas en este contexto es la de ceder un porcentaje o fracción de cómputo de cada máquina. Es muy difícil diferenciar los términos usados en este contexto tales como “clusters”, “NOW” (Networks of Workstations), “COW” (Cluster of Workstations) y “Workstation Clusters”. En [12], por ejemplo, se afirma directamente que son sinónimos y se diferencia las distintas posibilidades según las características de procesamiento, objetivos y hardware de base entre otros índices. De hecho, a lo largo de toda esta tesis se hará referencia a esta clase de plataforma de cómputo paralelo como “redes locales” y también como “clusters”, que aparentemente es una de las denominaciones más utilizadas en la bibliografía. Sin embargo, es importante notar que las redes locales que no se han construido con el objetivo específico de cómputo paralelo como en los casos anteriores de los clusters homogéneos y tienen características propias de costo y de procesamiento que no tienen las demás. Por lo tanto, es muy importante identificar las formas de paralelizar aplicaciones o las características propias de las aplicaciones paralelas para utilizar al máximo y/o de manera optimizada la capacidad de las redes locales instaladas y que se pueden aprovechar para

cómputo paralelo.

Las definiciones y el alcance de cada término o expresión no está aún muy bien especificado en general. Solamente como ejemplo, se puede notar que en [48] [49] toda red de computadoras interconectadas en red se denomina *cluster* y se dividen en dos clases generales:

- NOW (Network of Workstations): cada computadora tiene uno o más usuarios que permiten la utilización de la computadora en los períodos de inactividad.
- PMMPP ("Poor Man's" MPP): cluster dedicado a ejecución de aplicaciones paralelas con requerimientos de alto rendimiento.

Sin embargo, en [12] por ejemplo, se mantiene que todas las computadoras en red que se usan para cómputo paralelo son clusters y se los clasifica según

- Objetivo (alto rendimiento o disponibilidad).
- Relación de pertenencia de los usuarios (cada computadora se dedica exclusivamente a cómputo paralelo o no).
- Hardware de cada nodo o computadora (PC, SMP, etc.).
- Sistema operativo de cada nodo o computadora.
- Configuración de cada computadora (Homogéneo o Heterogéneo).
- Niveles de "clustering".

Por lo tanto, de alguna manera se combinan términos de hardware, software y configuración de la red completa para cada caracterización. Más allá de que se utilizarán los términos "redes locales" y "clusters" de aquí en más, las redes locales instaladas y que se aprovechan para cómputo paralelo en esta tesis se consideran como:

- Clusters, dado que son un conjunto de computadoras interconectadas en red que se utilizan para cómputo paralelo.
- NOW, en el sentido de que cada computadora tiene un usuario o conjunto de usuarios que son quienes la ceden para cómputo paralelo.
- PMMPP, dado que por el tiempo que se dispone una computadora se tiene de manera completa. Aún en el caso de tener una fracción de cada computadora, esa fracción se tiene siempre disponible.
- Heterogénea, dado que es muy difícil que en las redes locales instaladas todas las computadoras tengan el mismo hardware y el mismo software de base. Esto se debe principalmente a
    - Objetivo de cada computadora o aplicaciones usuales que resuelve. La existencia de cada máquina se debe a un usuario o a un conjunto de aplicaciones que deben ser resueltas de manera más o menos periódica. Ni los usuarios ni las aplicaciones necesariamente tienen los mismos requerimientos de cómputo, almacenamiento, etc. Por lo tanto, en general la existencia de cada computadora no tiene relación o tiene muy poca relación con las demás (al menos en cuanto a características físicas).
    - Tiempo de instalación de la red. El ingreso de nuevas máquinas a la red local así como la reparación y actualización de las máquinas existentes, que se puede denominar evolución de la red hace muy difícil mantener la homogeneidad. La evolución de las redes locales es innegable, y la relación entre la evolución y la heterogeneidad tampoco. En general, a mayor tiempo de instalación de una red local mayor será la heterogeneidad. Solamente como ejemplo, se puede mencionar que la disponibilidad de un tipo de computadora es bastante limitada.

> En el caso de las PCs, este tiempo de disponibilidad suele contabilizarse en meses y a lo sumo quizás en algo más de dos años.

Es muy interesante estimar la evolución de los clusters homogéneos que actualmente cuentan con un período relativamente corto de instalados y en producción. Aunque con aplicaciones y usuarios controlados y bien administrados, la probabilidad de fallas en el hardware sigue presente. Y a mayor cantidad de computadoras y hardware en general, mayor será la probabilidad de que algo falle. Por lo tanto, la reparación y/o reposición no es algo *fuera de lo común* ni lo será en los clusters homogéneos. Cuando no sea posible mantener homogéneo un cluster por la disponibilidad de hardware a reparar/reponer se tienen, en general, dos alternativas (descartando la fabricación *ad hoc*, cuyo costo es sumamente alto):

- Mantener la homogeneidad y no reparar ni reponer. Es válida siempre y cuando las aplicaciones se sigan resolviendo. Teniendo en cuenta que en general las aplicaciones tienden a aumentar sus requerimientos esta alternativa parece poco útil en general.
- Reparar y/o reponer con hardware disponible. Automáticamente el cluster se transformará en heterogéneo. La heterogeneidad se traduce directamente en diferentes valores para las métricas de rendimiento cuando se trata de procesadores y/o memoria en el caso de las tareas de cómputo intensivo.

Otra de las razones por las cuales un cluster no necesariamente se puede mantener homogéneo es la necesidad de resolver aplicaciones con mayores requerimientos de cómputo y/o almacenamiento. En este sentido, a medida que transcurre el tiempo, la disponibilidad del hardware es menor y por lo tanto, si se decide mantener la homogeneidad o resolver las aplicaciones con hardware homogéneo se debería instalar un *nuevo* cluster. Esta decisión tiene el costo inmediato de la adquisición e instalación de *todo* un cluster, además de la necesidad de tomar alguna decisión respecto del cluster homogéneo que ya no tiene la capacidad suficiente para resolver la/s aplicación/es. La alternativa que parece ser más lógica, es decir la instalación del hardware necesario para *aumentar* la capacidad total del cluster, en general implica heterogeneidad.

Desde la perspectiva de que es muy difícil el mantenimiento homogéneo aún de los clusters dedicados a cómputo paralelo, todo aporte que se haga en el contexto de las redes de computadoras heterogéneas tiene un amplio espectro de utilización. De hecho, cada vez que, por la razón que sea, un cluster se “transforma” en heterogéneo las aplicaciones y los usuarios se tendrán que adaptar de una manera u otra al hardware para obtener rendimiento óptimo (que, como se afirma antes, es la principal razón para la existencia misma del cómputo paralelo).

1.2 Costos de Cómputo Paralelo en las Redes Locales Instaladas

Tal como se ha adelantado, las redes locales instaladas son la plataforma de cómputo paralelo de “costo cero” en cuanto a hardware. De hecho, los costos de

- Instalación

- Administración y monitoreo
- Mantenimiento

de cada computadora no tienen relación con el cómputo paralelo y por lo tanto no debería ser asumido por quienes las aprovechan para cómputo paralelo, dado que

- Las redes locales ya están instaladas y funcionando, y entre los objetivos no ha estado el de cómputo paralelo (al menos en la *gran* mayoría de las redes locales instaladas). El hecho de aprovecharlas para cómputo paralelo no cambia esta realidad, aunque de alguna manera se agrega como una de las utilizaciones de las redes locales.
- Cada computadora y la red local misma tiene al menos un administrador que se encarga de la configuración, interconexión y una mínima verificación de funcionamiento.
- Cada computadora tiene al menos un usuario o un conjunto de usuarios que se encarga de una manera u otra del mantenimiento dado que la utilizan. También es común que estos usuarios se encarguen de gestionar y/o llevar a cabo la actualización del hardware, con lo que se tendría costo cero también de actualización del hardware.

Pero el costo de hardware no es todo el costo relacionado con el procesamiento de las redes locales instaladas que se pueden aprovechar para cómputo paralelo. Los costos que se deben asumir en este contexto son los relacionados con:

- Las herramientas de administración de la red local para cómputo paralelo y para desarrollo y ejecución de programas paralelos.
- La disponibilidad de las computadoras de la red local y de la misma red local.
- La paralelización de aplicaciones.

Las herramientas de desarrollo y ejecución de programas paralelos son imprescindibles para aprovechar las redes locales para cómputo paralelo. Lo primero que ha sido desarrollado en este sentido son bibliotecas de pasaje de mensajes tales como PVM (Parallel Virtual Machine) [44] que se estableció como un estándar *de facto* en esta área. Posteriormente se propuso el estándar MPI (Message Passing Interface) [88] [107] [92] y actualmente se utilizan ambos, con la tendencia de desarrollo de las nuevas aplicaciones paralelas en las implementaciones de MPI disponibles para la arquitectura paralela que se utiliza. Específicamente para las redes de computadoras ambas bibliotecas están implementadas y su utilización es libre [PVM] [LAM/MPI] [MPICH] y se obtienen vía Internet. En este sentido, el único costo de estas herramientas es el de instalación. Aunque de manera no tan centralizada o estandarizada como PVM y MPI, también se han desarrollado múltiples herramientas de administración de redes de computadoras para cómputo paralelo y también están disponibles en Internet. Las instalaciones Beowulf son una de las principales fuentes de herramientas para administración de redes que se utilizan para cómputo paralelo y que se pueden aprovechar. Sin embargo, existe cierto consenso en que el costo de administración de una red de computadoras que se utiliza para cómputo paralelo es bastante mayor que en el caso de las computadoras paralelas clásicas [17].

La disponibilidad de las computadoras de una red local instalada para cómputo paralelo no es completa. Si bien este costo es muy difícil de cuantificar, la utilización de las redes locales tiene varias alternativas, dos de las cuales son:

- Períodos de casi total inactividad tales como las noches o los días no laborables. Si bien es dependiente de características específicas en cada red local, incluso en horarios considerados como de mucha utilización, las computadoras no necesariamente se aprovechan al máximo [90].

- Determinar *a priori* un porcentaje de cada computadora para ser utilizado por procesos de aplicaciones paralelas. En este caso es como tener la misma cantidad de computadoras pero cada una de ellas con menor capacidad.

Desde el punto de vista de las aplicaciones paralelas, ambas alternativas son similares: se tiene un conjunto de recursos disponibles. Este es el contexto de "disponibilidad" de las computadoras que se utilizarán. Más específicamente, la experimentación se llevará a cabo durante los períodos en los cuales las redes locales no se utilizan para ninguna otra cosa y por lo tanto la disponibilidad será total (durante la ejecución de los programas paralelos).

Quizás el costo más considerable o al menos más desconocido es el de la paralelización y/o desarrollo de programas paralelos para las redes locales instaladas y cuyos recursos se pueden aprovechar. El problema mayor en este contexto es el de la paralelización misma. Siempre ha sido considerado uno de los grandes problemas (y con su costo asociado), dado que no hay métodos generales. Una de las grandes razones para que esto suceda es justamente la razón de ser del cómputo paralelo: el rendimiento. Si bien se pueden diseñar algoritmos paralelos sintáctica, semántica y estilísticamente muy buenos, no son útiles cuando no obtienen rendimiento aceptable. En general, es muy difícil cuantificar el término *aceptable*, pero se pueden mencionar dos posibles acepciones:

- Utilizan al máximo los recursos disponibles. En general se acentúa la importancia en términos de utilización de los procesadores disponibles.
- Tiempo de respuesta mínimo. En este caso es dependiente de la aplicación. En el caso de paralelizar la tarea de predicción de las condiciones meteorológicas (estado del tiempo) para un día determinado, es claramente inadecuado obtener la respuesta de predicción en uno o varios días posteriores.

Varias métricas de rendimiento de sistemas paralelos se concentran en la utilización de los recursos disponibles dado que:

- Es independiente de las aplicaciones y en este sentido las métricas son generales.
- Si se utilizan al máximo los recursos disponibles se asume que no se puede obtener nada mejor, como mínimo en la máquina paralela con el algoritmo paralelo que se utiliza.

Sin embargo, como se ha afirmado antes, muchos de los problemas de álgebra lineal han sido resueltos satisfactoriamente con máquinas paralelas. Una de las primeras tareas que se deben llevar a cabo entonces es la revisión de los algoritmos paralelos ya desarrollados para aprovecharlos en el contexto de las redes de computadoras. En este punto no hay que perder de vista que las redes de computadoras y más específicamente cada computadora que se puede conectar en una red (con una interfase de red) no ha sido ni en principio es diseñada para cómputo paralelo distribuido en múltiples máquinas. Por lo tanto, se tiene que

- Como mínimo se deben analizar los algoritmos paralelos propuestos y su efectividad en cuanto a rendimiento en las redes de computadoras.
- Si no hay ningún algoritmo paralelo que se pueda considerar apropiado para cómputo paralelo asociado a un problema en una red se debería proponer al menos otro para considerar la efectividad de este tipo de arquitecturas paralelas.

Si bien el análisis de los algoritmos paralelos propuestos hasta el momento se debe hacer de manera exhaustiva y quizás "caso por caso" (o al menos por área de aplicaciones o características de procesamiento) se tiene un inconveniente desde el principio: las máquinas paralelas han sido y son diseñadas para cómputo paralelo y las redes de computadoras no. Es esperable, por lo tanto, que los algoritmos paralelos no sean

directamente utilizables en las redes locales que han sido instaladas y son utilizadas con múltiples propósitos y el de cómputo paralelo no es uno de ellos, o no es uno de los más importantes.

La paralelización de aplicaciones para este tipo de arquitectura paralela tiene varios inconvenientes o al menos varias características que no se conocen en las máquinas paralelas tradicionales. Los dos inconvenientes que pueden considerarse como más importantes son:

- Heterogeneidad de los nodos de cómputo.
- Características de la red de interconexión.

Tanto las máquinas paralelas tradicionales como los clusters homogéneos tal como se los ha caracterizado antes (construidos o instalados para cómputo paralelo) tienen elementos de procesamiento (procesadores) homogéneos. Esto simplifica de manera notable la distribución de la carga de cómputo, dado que para lograr balance de carga de procesamiento *simplemente* es necesario distribuir la misma cantidad de operaciones que cada elemento de cómputo debe ejecutar. En general, la definición del término *simplemente* que se ha utilizado no es trivial, pero en el contexto de las aplicaciones de álgebra lineal normalmente implica la distribución de la misma cantidad de datos (o porciones iguales de datos de matrices) entre los procesadores. Por lo tanto, la paralelización de las aplicaciones provenientes del área de álgebra lineal no ha tenido muchos inconvenientes asociados al balance de carga en las computadoras paralelas homogéneas. En las redes locales que se aprovechan para cómputo paralelo evidentemente también se tendrá que resolver el problema causado por las diferencias entre las capacidades de cómputo de las distintas máquinas utilizadas.

La red de interconexión más extendida en cuanto a redes locales de computadoras instaladas es sin lugar a dudas la definida por el estándar Ethernet [73] de 10 Mb/s y de 100 Mb/s. De hecho, se reconoce que la red Ethernet de 10 Mb/s no es adecuada en general para cómputo paralelo [91]. Más allá de las características de rendimiento máximo, las redes Ethernet tienen varios inconvenientes por su propia definición y dependencia del protocolo CSMA/CD (Carrier Sense-Multiple Access/Collision Detect), que implica rendimiento en general muy dependiente del tráfico individual de cada una de las máquinas interconectadas. Sin embargo, la enorme potencia de cálculo instalada en las redes locales hace valioso cualquier aporte al respecto. Las aplicaciones provenientes del álgebra lineal, por otro lado, tiene una cantidad relativamente grande de usuarios potenciales y de redes locales *disponibles* para cómputo paralelo. Desde el punto de vista del costo, las redes Ethernet no solamente son las más difundidas y por lo tanto *aprovechables* en cuanto a instalaciones actuales, sino que además tiene una gran *inercia* en cuanto a instalaciones nuevas dado que

- Existe una gran cantidad de personal técnico capacitado y con experiencia que puede seguir instalando estas redes. Cambiar de tecnología en general implica mayor costo por lo menos de capacitación de personal técnico.
- Existe una gran cantidad de software desarrollado y que en general tiende a ser estable y utilizado.
- La norma Ethernet se ha definido con mayores capacidades de transferencia de datos, tales como 1 Gb/s (10^6 bits por segundo) [76] y 10 Gb/s (10^7 bits por segundo) [110].

Desde otro punto de vista, hay otros costos asociados que no son tan relacionados con los estrictamente técnicos como los que se han mencionado. Tal como se puntualiza en [18] (en el contexto de proporcionar *high throughput*), la utilización de redes locales es un problema tecnológico y sociológico. En este sentido, existe de un "evangelista" (tal como se lo denomina en [18]) que desarrolla y crea con sus propios recursos y con los de "aliados" un *high throughput computing cluster* y con él "genera demanda" hacia un conjunto más amplio de usuarios potenciales que a su vez aportarán sus propios recursos individuales. Pero además, toda red local o conjunto de redes locales debe tener apoyo explícito de una organización para posibilitar su aprovechamiento para cómputo paralelo. Sin embargo, en esta tesis este/estos costos no se abordarán.

1.3 Resumen de Objetivos y Aportes de la Tesis y Organización del Contenido

El principal objetivo de esta tesis es la evaluación de los problemas y soluciones para el cómputo paralelo en las redes de computadoras instaladas, con sus características de cómputo y comunicaciones. El problema a través del cual se realiza la evaluación es el de la multiplicación de matrices por varias razones:

- Representatividad del tipo de procesamiento con respecto al resto de las operaciones (y aplicaciones) provenientes del álgebra lineal.
- Representatividad en cuanto a requerimientos de cómputo y de almacenamiento, específicamente con respecto a las rutinas incluidas en BLAS de nivel 3.
- Cantidad de publicaciones identificando tanto algoritmos paralelos propuestos como rendimiento obtenido.

Toda la tesis está orientada a la paralelización de aplicaciones para la obtención del máximo rendimiento posible, por lo tanto aunque inicialmente se hará uso de herramientas estándares y ampliamente utilizadas en este contexto como PVM la tendencia será mejorar y/o reemplazar todo lo que imponga una excesiva penalización en el rendimiento. Esto abarca desde la forma de llevar a cabo cómputo local hasta la monitorización y evaluación de las rutinas de comunicaciones que se utilizan.

El método con el que se hará la evaluación tiene, en principio dos partes:

- Análisis de algoritmos, métodos y herramientas ya propuestos para llevar a cabo procesamiento paralelo. Si este análisis implica *a priori* una clara penalización en cuanto a rendimiento, se propondrá al menos una alternativa considerada apropiada.
- Experimentación exhaustiva. Se definirán un conjunto de experimentos a llevar a cabo y se analizará la validez o no del rendimiento obtenido.
  En el caso de que el rendimiento obtenido no sea satisfactorio, se identificará la fuente de la penalización de rendimiento y se propondrá al menos una alternativa de solución (en términos de algoritmos o implementaciones de rutinas específicas) con la cual se realizarán nuevamente los experimentos propuestos.

Los aportes se pueden resumir en:

- Identificación del nivel de aprovechamiento de los algoritmos paralelos propuestos

específicamente para la multiplicación de matrices en el contexto de cómputo paralelo en redes de computadoras instaladas.
- Identificación de las características de procesamiento paralelo con las cuales se debe resolver específicamente la multiplicación de matrices.
- Identificación de las características de procesamiento paralelo con las cuales se deben resolver las aplicaciones provenientes del álgebra lineal en general.
- Identificación y propuestas de solución para los posibles problemas de rendimiento ocasionados tanto por el rendimiento local de cada computadora como de las rutinas de comunicaciones que se utilizan.

La utilización de estos aportes tiende a:
- Aprovechar la capacidad de cómputo disponible en las redes locales instaladas, aunque no tengan como objetivo el procesamiento paralelo.
- Aprovechar las actuales instalaciones de clusters homogéneos que se utilizan para procesamiento paralelo cuando, por la evolución en en hardware y/o en las aplicaciones, sean heterogéneos.

Más específicamente, a lo largo de esta tesis se tendrá:
1. Del análisis de los algoritmos paralelos de multiplicación de matrices surgirá claramente que no son directamente utilizables en la plataforma de cómputo paralelo que representan las redes locales de computadoras.
2. Aún cuando se proponga un algoritmo específico de multiplicación de matrices para aprovechar al máximo las características de la arquitectura de cómputo paralelo que proveen las redes locales, el rendimiento optimizado no está asegurado. Se mostrará por experimentación y monitorización (vía instrumentación) del rendimiento del algoritmo que la biblioteca de pasaje de mensajes PVM impone penalización de rendimiento inaceptable.
3. Se analiza y explica brevemente las razones por las cuales las bibliotecas de pasaje de mensajes de propósito general (entre las cuales se incluyen la propia biblioteca PVM y las implementaciones de MPI, por ejemplo) no pueden asegurar rendimiento optimizado en las redes locales de computadoras utilizadas para cómputo paralelo.
4. Se propone e implementa una única operación de pasaje de mensajes directamente orientada al aprovechamiento de las características de las redes Ethernet utilizadas para la de interconexión de computadoras en las redes locales.
5. Se evalúa, vía experimentación, el rendimiento de los algoritmos propuestos incluyendo la operación de pasaje de mensajes también propuesta, mostrando que el rendimiento puede ser considerado como aceptable u optimizado para las redes locales de computadoras utilizadas para cómputo paralelo.
6. Se muestra también cómo uno de los algoritmos propuestos para multiplicación de matrices puede ser utilizado para evaluar la capacidad de las máquinas de una red local de computadoras de llevar a cabo comunicaciones en *background* (*solapadamente*) mientras realiza cómputo local.
7. Finalmente, también se compara la propuesta algorítmica de esta tesis con los algoritmos específicos de la biblioteca ScaLAPACK, que es considerada como la que implementa los mejores algoritmos (en cuanto a rendimiento y a escalabilidad) de cómputo paralelo en arquitecturas paralelas de memoria distribuida. En este caso específico de comparación solamente se tienen en cuenta redes de computadoras homogéneas, dado que ScaLAPACK no tiene ninguna previsión para el caso de las

arquitecturas con elementos de cómputo heterogéneos. Esta comparación también da resultados favorables para la propuesta de esta tesis, dado que se tienen mejores resultados de rendimiento y escalabilidad, al menos hasta donde fue posible evaluar en cuanto a cantidad de máquinas utilizadas para cómputo paralelo.

El siguiente capítulo, **Capítulo 2: Multiplicación de Matrices**, se dedicará al análisis de la multiplicación de matrices en general, en el contexto de las operaciones de álgebra lineal y también se analizarán los algoritmos paralelos que se han propuesto.

En el **Capítulo 3: Clusters Heterogéneos** se analiza en detalle las características de las redes locales instaladas desde el punto de vista de cómputo paralelo. De acuerdo con este análisis se identifican las características principales de cómputo paralelo en esta plataforma de procesamiento y también se proponen dos algoritmos específicos para multiplicar matrices en paralelo.

El **Capítulo 4: Experimentación**, detalla los experimentos realizados y analiza los resultados obtenidos con la biblioteca de pasaje de mensajes PVM. Muestra con detalle los problemas de rendimiento que genera esta biblioteca en particular y también se comentan en general las expectativas con respecto a las demás bibliotecas de pasaje de mensajes para las redes de computadoras. Se propone una rutina de comunicaciones propia que, al rehacer los experimentos muestra cómo se logra utilizar al máximo o de manera optimizada tanto los recursos disponibles de cómputo como los de comunicaciones y esta combinación proporciona rendimiento satisfactorio.

En el **Capítulo 5: Comparación con ScaLAPACK** se presentan dos aspectos importantes en cuanto a la validez y utilización de los aportes de esta tesis: 1) Aplicación de los principios de paralelización en ambientes homogéneos dedicados a cómputo paralelo para dos casos específicos: multiplicación de matrices y factorización LU de matrices, y 2) Comparación de resultados obtenidos por experimentación en cuanto a rendimiento con respecto a la biblioteca ScaLAPACK. Esta biblioteca está dedicada específicamente a las plataformas de cómputo paralelo homogéneas con memoria distribuida, y es aceptada en general como la que implementa los mejores algoritmos paralelos existentes en cuanto a escalabilidad y optimización de rendimiento.

En el **Capítulo 6: Conclusiones y Trabajo Futuro**, se resumen las principales conclusiones a partir de la tarea de análisis y experimentación realizada. También se adelantan de manera estimada las principales líneas de investigación que siguen a partir del trabajo realizado en esta tesis.

Posteriormente se dan los detalles de la **Bibliografía** a la cual se hace referencia a lo largo de esta tesis y se incluyen los **Apéndices**. En general, la idea de los apéndices es que sean autocontenidos y es por eso que tienen su propia lista de referencias bibliográficas.

El **Apéndice A: Características de las Redes Locales** muestra todo el detalle del hardware las computadoras y del cableado de las redes locales que se utilizaron para la experimentación.

El **Apéndice B: Rendimiento de Procesamiento Secuencial de las Computadoras**

muestra el impacto de los niveles de optimización de código de procesamiento en cada una de las computadoras, su rendimiento máximo (que se aprovecha en el procesamiento paralelo) así como algunos comentarios respecto del código que se utiliza y que se debería utilizar en general en la experimentación con y en la producción de programas paralelos.

El **Apéndice C: Comunicaciones en la Red Local del CeTAD** muestra el rendimiento punto a punto de las comunicaciones en una de las redes locales utilizadas y también de las comunicaciones con la rutina *broadcast* provista por la biblioteca PVM entre procesos de una aplicación paralela. También se hacen comentarios respecto del rendimiento de las comunicaciones en las demás redes locales utilizadas y con algunas referencias a otras bibliotecas de pasaje de mensajes disponibles para cómputo paralelo en redes de computadoras.