

EDITORIAL

Estimados lectores de *Buran*. A través de estas líneas queremos agradecerles y darles las bienvenida de nuevo a este mundo llamado *Buran*. Un mundo que nació hace ocho años y del que han formado parte tantas y tantas personas con su esfuerzo y su desinterés. A los que se incorporen nuevos esperamos que estas páginas sean de su agrado, y a los que llevan más tiempo con nosotros, que sigan disfrutando con *Buran*.

Personalmente nos sentimos orgullosos de ver como poco a poco *Buran* va creciendo, madurando y consolidándose como un medio de comunicación independiente y abierto a cualquier persona que sienta identificado con nuestra política de hacer de la ingeniería el medio a través del cual mejorar nuestras vidas y de las personas que participan de ella, y alcanzar una satisfacción personal y profesional que se nos haga seguir trabajando. *Buran* es en si, el resultado de un interés común.

Nuestra única pretensión es hacer que *Buran* siga creciendo, sin afán de lucro, sin intereses que vayan más allá de los propiamente intelectuales. Por eso desde aquí, sólo nos queda agradecer a todas las personas que han sabido ver en estas páginas nuestra idea, nuestro concepto de *Buran*.

Por eso, el sentir la responsabilidad de no defraudar a nuestros colaboradores y lectores hace que intentemos poco a poco mejorar, sin encerrarnos, sin establecer fronteras, y poder dejar este legado a futuras generaciones que sientan *Buran* como algo que ha formado, forma y formará parte de nuestras vidas, no sólo por el hecho de hacer una publicación, sino por haber encontrado un hueco donde compartir momentos, buenos y malos, con nuestra gente, con nuestros compañeros, con nuestros amigos...

Quizás sea difícil concebir este sentimiento a los que leen *Buran*, pero créannos si les decimos que la sensación de hacer algo que a la gente le gusta y de la cual participan, no es comparable a nada. Por eso estamos abiertos a cualquier sugerencia que nos permita mejorar *Buran*. Aunque suene tópico, *Buran* es una revista hecha por unos pocos, para muchos. E intentaremos cambiar en continente, que no en contenido, para que los que hemos trabajado durante esta etapa podamos ser recordados como los que lucharon porque *Buran* dure cuanto menos otros tantos años más. Estoy seguro que será más, y esperamos no caer en la monotonía.

En este número no podemos olvidarnos de toda la gente que ha compartido con nosotros su trabajo, desde otras comunidades, desde otros países, al otro lado del charco... A todos ellos, gracias por vuestra ayuda, porque sin ustedes quizás estas líneas no tendrían sentido.

COORDINACIÓN BARCELONA

Jorge Sáiz Fernández

EDICIÓN BARCELONA

José A. López Salcedo

Xavier Palau Marqués

Carles Ruiz Floriach

Jorge Sáiz Fernández

Jose Castor Vallés Martínez

Luis Almajano

REVISIÓN

José A. López Salcedo

Laura Mascaró Rotger

Xavier Palau Marqués

Carles Ruiz Floriach

Jorge Sáiz Fernández

Miguel Ángel Sastre Serra

Jose Castor Vallés Martínez

EDICIÓN GRÁFICA

José A. López Salcedo

Jorge Sáiz Fernández

AGRADECIMIENTOS

II. Dir. Antoni Elias Fusté, Ángel Cardama,
IEEE Internacional, Jorge Luis Sánchez Ponz,

Pere Camps y a los puntos de
distribución en la UPC:

Abacus, CPET, CPDA, Kiosk Campus Nord
y Reprografía Sant Just.

IMPRESIÓN

RET, s.a.l.

FOTOMECLÁNICA

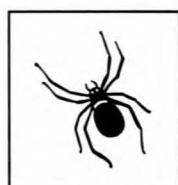
Sistemes d'Edició

DEPÓSITO LEGAL

B-19.950-96

La organización se reserva el derecho de publicar los artículos. La opinión expresada en los artículos no tiene por qué coincidir con la de la organización.

Agradecemos las colaboraciones hechas desinteresadamente, y a causa de la falta de espacio, pedimos disculpas a todas aquellas personas a las cuales no se les ha publicado su colaboración. Esperamos que en un próximo número tengan cabida.



IMPLEMENTACIÓN DE UNA RED NEURONAL EN UNA CALCULADORA HP 48GX PARA EL CONTROL DE UN-NITI, UN MINIROBOT CAMINADOR CON ALAMBRE NITINOL

Camilo Andrés Cortés G.¹, Juan Pablo Sáenz E.², Ing. Alberto Delgado³

¹camilokmi@ieee.org ²jpsaenz@ieee.org

³Profesor Facultad de Ingeniería adelgado@ieee.org

UNIVERSIDAD NACIONAL DE COLOMBIA

RESUMEN

El diseño y construcción de un robot móvil es un arte y una ciencia. Un diseñador de robots debe poseer un compendio de habilidades básicas de varios campos, como de ingeniería mecánica, ingeniería eléctrica, ciencias de la computación e inteligencia artificial [9].

La calculadora Hewlett Packard 48G(X) es una herramienta que tiene gran acogida en el ámbito universitario, posee un procesador Saturn de 4 MHz, una memoria expandible hasta 4MB, un puerto infrarrojo y un puerto serial RS-232 [11], exhibiendo características deseables para utilizarse como controlador.

El desarrollo de programas de control secuencial ha sido ampliamente estudiado por los investigadores en robótica en el mundo, demostrando ser complejos, extensos, poco flexibles y de escasa robustez. Las nuevas técnicas de control basadas en los nuevos paradigmas en Inteligencia Artificial (IA) como las redes neuronales son simulables fácilmente en un PC, aportando mayor flexibilidad y robustez que los métodos tradicionales.

Una alternativa para producir movimiento mecánico utilizando una señal eléctrica sin utilizar motores es el nitinol (aleación de níquel y titanio que se contrae al aplicar en sus extremos una diferencia de potencial). Las investigaciones mundiales sobre este material han producido múltiples aplicaciones como robots, máquinas giratorias hasta de 1000 RPM, acoplos de tuberías, conectores de tarjetas de computadores y displays de caracteres braile [6].

UN-NITI es un minirobot cuyo movimiento se produce mediante actuadores de nitinol acoplados a sus 8 patas. Su control electrónico se realiza gracias a cuatro módulos. La calculadora HP 48GX sirve como controlador secuencial o inteligente del movimiento del robot, todo mediante comunicación serial RS-232.

INTRODUCCIÓN

El proyecto «Diseño y construcción de un robot Caminador utilizando alambre NITINOL y una calculadora HP 48GX como plataforma de Inteligencia Artificial», UN-NITI (Universidad Nacional - NITInol), fue un trabajo de grado de ingeniería Eléctrica en el área de robótica y control inteligente, el cual tuvo por objetivo construir un minirobot caminador de ocho patas (tipo araña). El robot funciona con alimentación externa, su movimiento (sin motores) se produce con alambre nitinol y es controlado por una red neuronal implementada en una calculadora HP48GX utilizando el puerto serial para la comunicación (Figura 1).

ESTRUCTURA MECÁNICA DEL ROBOT

Físicamente el robot se elaboró con diferentes materiales, su cuerpo consiste en una lámina de balsó en la cual se encuentran acopladas las tarjetas correspondientes a los módulos de control, además se utilizaron dos piezas rectangulares de madera a las cuales se fijaron las patas del robot elaboradas con cuerda de piano de 0.95 mm de diámetro. Las patas tienen unos «zapatos» hechos con tubos capilares de cobre que facilita

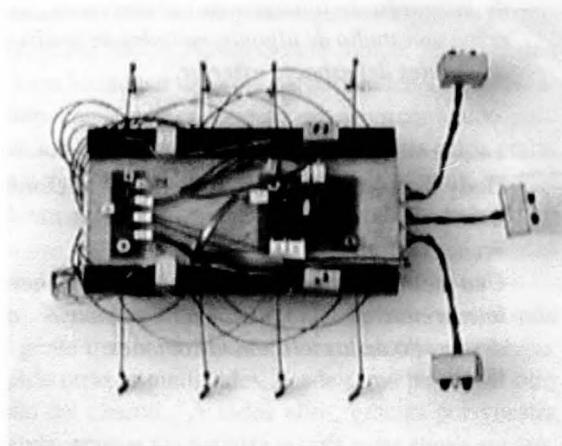


Figura 1a). UN-NITI. Vista superior

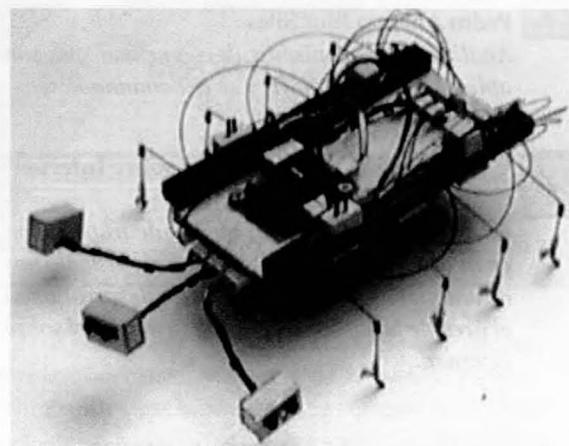


Figura 1b). UN-NITI. Perspectiva

tan el desplazamiento. En la figura 2 se observa la estructura mecánica del robot.

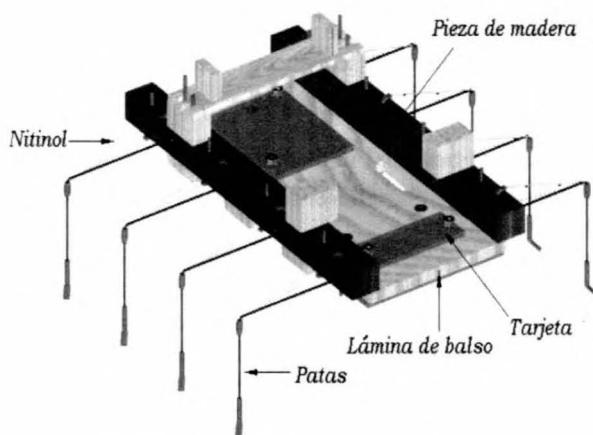


Figura 2. Diseño mecánico de UN-NITI.

El movimiento de las patas se realiza con alambre nitinol fijado a las patas y a las piezas de madera mediante presión mecánica ejercida sobre tubos capilares de cobre (esta técnica se implementó para no utilizar métodos calientes, como la soldadura, que pueden dañar la estructura cristalina del nitinol). En la figura 3 se ilustra la forma en que se encuentra acoplado el nitinol al cuerpo del robot.

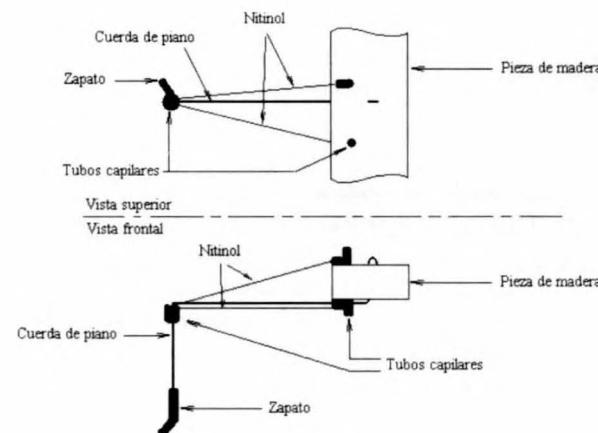


Figura 3. Detalles del acople del nitinol en UN-NITI.

Actuadores de Nitinol

El nitinol es un alambre hecho de una aleación entre níquel y titanio que presenta las características y propiedades de las aleaciones con memoria de forma (Shape Memory Alloys)[2]. El término SMA es aplicado a un grupo de materiales metálicos que han demostrado la habilidad de retornar a una forma o tamaño preestablecidos mediante un adecuado procedimiento térmico y mecánico.

Las ventajas de utilizar nitinol en cambio de motores son: tamaño y peso reducidos, bajo consumo de potencia, control preciso, operación con AC o DC, larga vida y capacidad de soportar gran peso [6].

En UN-NITI los actuadores de nitinol mueven las patas al contraerse debido a la circulación de corriente por ellos. Esta corriente origina un aumento en la temperatura del nitinol produciendo un cambio en la estructura cristalina que disminuye la longitud del alambre (figura 4).

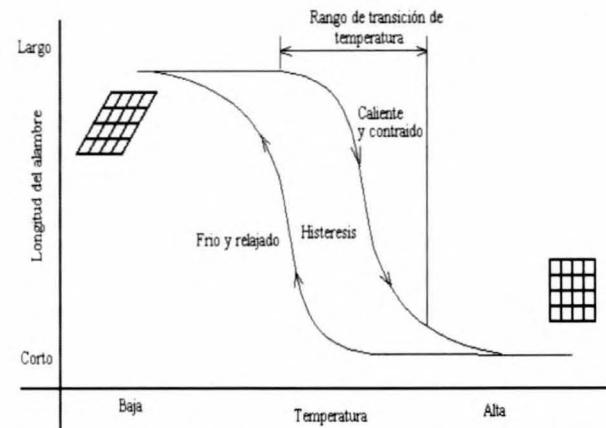


Figura 4. Longitud del alambre de nitinol sometido a una fuerza constante como función de su temperatura

Cuando la corriente cesa y la temperatura disminuye, el alambre recupera su longitud al ser aplicada una fuerza contraria ejercida por la pata. El nitinol utilizado en UN-NITI fue de 100 mm de diámetro (marca flexinol 100).

EL CONTROL ELECTRÓNICO

El sistema de control del robot está dividido en 4 módulos dedicados a tareas específicas (figura 5) con el fin de dar mayor flexibilidad al diseño del robot .

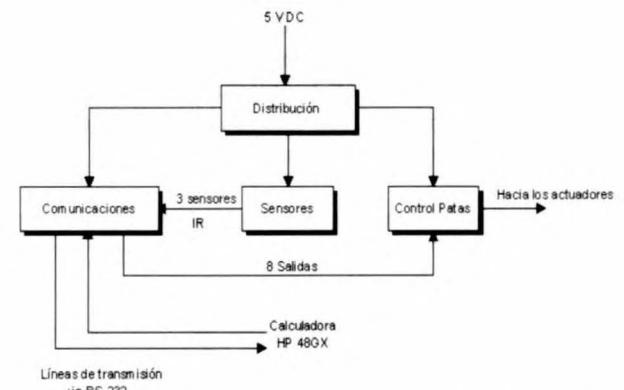


Figura 5. Diagrama de bloques ilustrativo del sistema de control de UN-NITI

Módulo de Distribución

Debido a la magnitud elevada de corriente para el movimiento del robot (aproximadamente 2.5 A para el funcionamiento de todos los módulos) fue necesario utilizar una fuente de alimentación externa de 5V DC. Este módulo tiene por función distribuir la energía con el

propósito de utilizar sólo un cable externo para permitir el movimiento libre del robot.

Módulo de Sensores

La habilidad de un robot para identificar su mundo por medio de sensores y cambiar su comportamiento es lo que hace a un robot una cosa interesante de construir y un artefacto útil cuando es terminado [4].

En UN-NITI se encuentran tres sensores infrarrojos (IR) que detectan la proximidad de obstáculos con un rango variable de detección de 3 a 30 cm dependiendo de la calibración de la tarjeta y del tipo de superficie del obstáculo. Cada sensor consta de dos LEDs emisores IR y un fotodiodo receptor. La señal de recepción es tratada por un Amplificador operacional y cada salida es comparada con el nivel de luz IR en el ambiente obtenido por un fotodiodo detector adicional, el diagrama esquemático de un sensor se presenta en la figura 6, este diseño es una modificación del propuesto para el detector de proximidad IR de SCORPIO [2].

Módulo de Control de Patas

Es el encargado de suministrar la corriente necesaria para mover los actuadores de nitinol dependiendo de la orden enviada por la tarjeta de comunicaciones, esta tarjeta es capaz de manejar 8 salidas cada una de las cuales mueve 2 actuadores conectados en paralelo.

Módulo de Comunicaciones

Es el más importante de todos por servir de interfaz entre la calculadora y el robot. Recibe las señales de los sensores enviando su estado a la HP 48GX, y traduce la orden enviada por la calculadora hacia el módulo de control de patas. Para estos fines se utiliza un convertidor de señales TTL - RS 232 (ICL 232) y un microcontrolador PIC 16F84 [8] programado con rutinas especiales para permitir la comunicación serial de UN-NITI.

ALGORITMOS DE CONTROL

La calculadora permite la implementación de software de control de diversos tipos (secuencial, inteligente, etc.). En la actualidad se encuentra implementado un algoritmo de control secuencial y una red Neuronal multicapa.

La red neuronal implementada en el software de control inteligente es la red multicapa 3-N-3 mostrada en la figura numero 7.

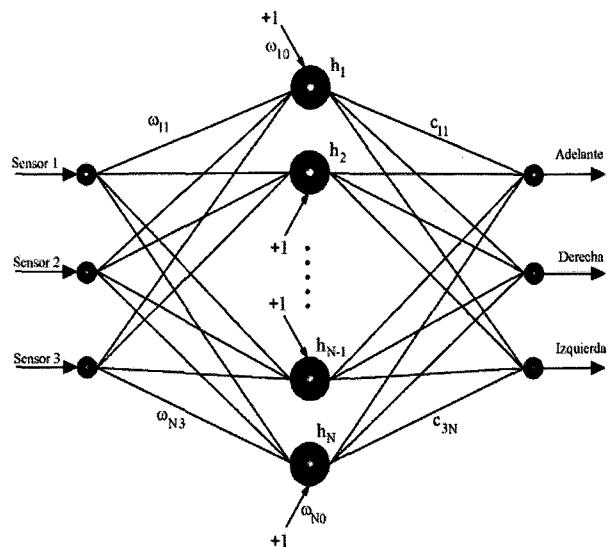


Figura 7. Red neuronal 3-N-3 usada en el software de control de UN-NITI

Es una RNA estática, con entrenamiento fuera de línea (Offline), que utiliza como algoritmo de entrenamiento el de propagación inversa (backpropagation) [3]. La capa de entrada y la de salida tienen función de activación lineal, la capa oculta cuenta con función de activación Tanh. La red fue entrenada con los patrones mostrados en la tabla 1.

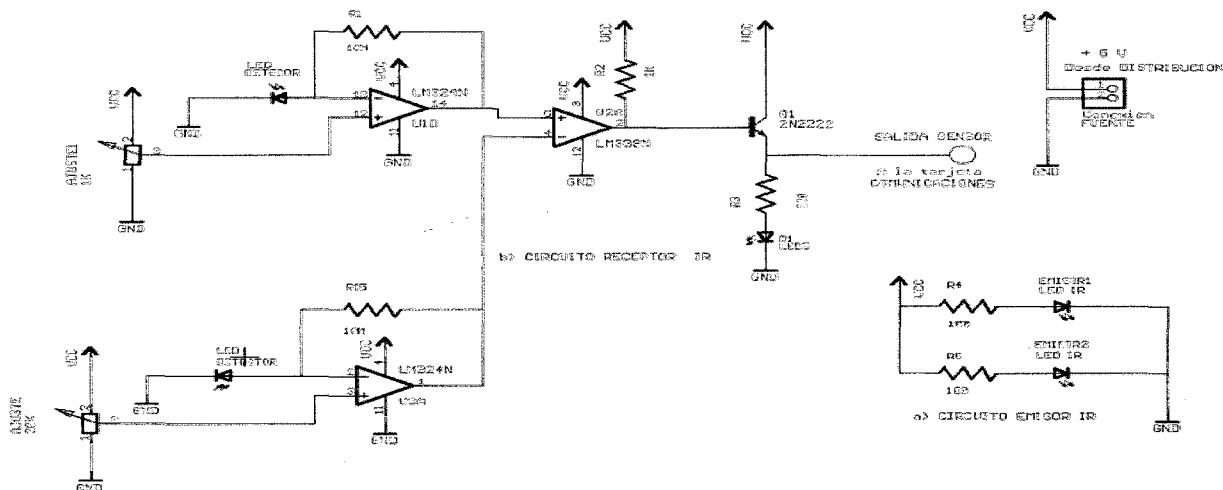


Figura 6. Diagrama esquemático de un sensor. a) Circuito Emisor b) Circuito receptor y comparador

Estado de los sensores			Acción a realizar		
Sensor 3	Sensor 2	Sensor 1	Adelante	Giro Der.	Giro Izq.
-0.9	-0.9	-0.9	0.9	-0.9	-0.9
-0.9	-0.9	0.9	-0.9	0.9	-0.9
-0.9	0.9	-0.9	-0.9	0.9	-0.9
-0.9	0.9	0.9	-0.9	0.9	-0.9
0.9	-0.9	-0.9	-0.9	-0.9	0.9
0.9	-0.9	0.9	0.9	-0.9	-0.9
0.9	0.9	-0.9	-0.9	-0.9	0.9
0.9	0.9	0.9	-0.9	-0.9	0.9

Tabla 1. Patrones de entrenamiento de la red neuronal.

En la aplicación con la HP 48 (figura 8) se utilizaron 4 neuronas ocultas y una rata de aprendizaje de 0.015 con el fin de lograr mayores velocidades a la hora de propagar la red y poder utilizar la calculadora como controlador en tiempo real.



Figura 8. Programa REDNEU en la HP48, para una red multicapa 3-N-3

La implementación de programas de control en la HP 48GX usando User RPL [7] (lenguaje de programación de más alto nivel de la HP 48GX) busca establecer la utilidad de la calculadora como controlador en tiempo real, además de analizar las facilidades que brinda al programador. Este lenguaje cuenta con estructuras de información, como pilas, listas y matrices, que permiten desarrollar programas complejos de manera rápida. Estas características del User RPL permitieron desarrollar los programas de control y en especial el de la RNA de forma



Figura 9. Control de UN-NITI desde un PC utilizando un emulador de la HP48.

más fácil que en los lenguajes tradicionales de programación para PC.

El software de control puede implementarse también desde un PC, utilizando para ello una interfaz adecuada (ej. Visual Basic, MATLAB, entre otros), o utilizando una versión emulada de la calculadora [1] como se ilustra en la figura 9. El uso del emulador tiene la ventaja de trabajar con velocidades de procesamiento mayores a las de la calculadora, por ejemplo, según pruebas con el entrenamiento de la red neuronal, se pudo observar que operaciones que toman 1 minuto en la calculadora (4MHz), en un computador Pentium III de 450MHz se reducen a 4 o 5 segundos.

Una de las aplicaciones futuras de UN-NITI es la programación inspirada en modelos biológicos para el estudio del comportamiento adaptativo utilizándolo como un ANIMAT [10], esta nueva tendencia de programación es una modificación de la Inteligencia Artificial (IA) tradicional en donde se simula el comportamiento animal (Inteligencia Artificial basada en el comportamiento).

EXPERIMENTOS Y RESULTADOS

El movimiento de las patas de UN-NITI se basó en el estudio de la forma y secuencia de caminar de las arañas [5] para obtener un movimiento más eficiente. Hasta la fecha, se han realizado diferentes experimentos, algunos de los cuales se exponen a continuación:

Al medir la velocidad de UN-NITI, se encontró que en línea recta avanza 12 cm/min sin cargar la calculadora y 7.5 cm/min cargándola. Esta velocidad es buena debido a que la velocidad del robot más conocido con nitinol, Stiquito [2], es de 3 a 10 cm/min con cargas hasta de 50 gr.

Se descubrió que la superficie óptima para UN-NITI depende del peso cargado, por ejemplo, cuando no carga la calculadora ni otro peso se desplaza muy bien sobre cartón paja. Las pruebas de giro mostraron que para

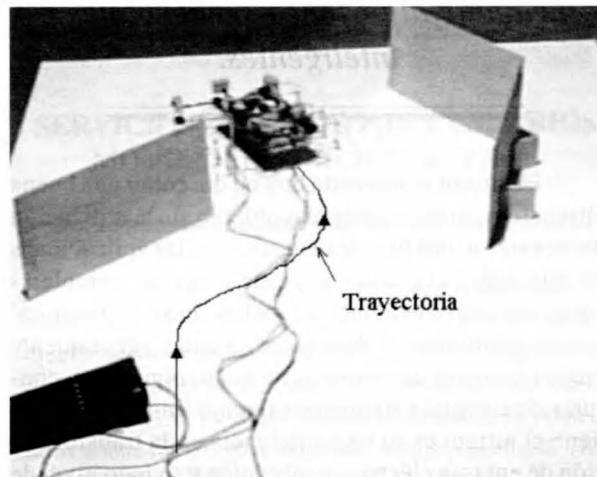


Figura 10. Configuración de obstáculos para la prueba de control inteligente.



realizar un giro de 90° en cualquier dirección el robot demora 6 min utilizando un radio de giro de 25 cms.

Para probar el comportamiento del control por medio de la RNA se dispuso de 3 obstáculos cuya disposición se puede apreciar en la figura 10. El minirobot sorteó los obstáculos en 12 min de manera exitosa siguiendo la trayectoria mostrada en la figura. Esta prueba demostró la viabilidad de usar la HP 48GX como controlador de sistemas físicos usando algoritmos inteligentes.

Actualmente UN-NITI ha funcionado con cargas hasta de 400 gr aproximadamente (264 de la calculadora y 140 de los módulos de control). Es un peso aceptable pues iguala las máximas capacidades obtenidas con las variaciones de Stiquito.

CONCLUSIONES

La calculadora HP 48GX demostró ser un adecuado controlador en tiempo real de un sistema físico (UN-NITI) mediante algoritmos secuenciales e inteligentes. Al ver la popularidad de esta calculadora entre los estudiantes de ingeniería en el mundo surge como una buena herramienta en la enseñanza del control y robótica, de fácil manejo y gran accesibilidad.

El lenguaje de programación utilizado, User RPL, es de fácil aprendizaje y cuenta con características propias de programación orientada a objetos. Además cuenta con estructuras de datos predefinidas (ej. listas) que ayudan en la programación de redes neuronales y cuenta con acceso a todas las funciones matemáticas de una calculadora científica (manejo de matrices, funciones hiperbólicas, etc).

La calculadora HP 48GX demostró ser un adecuado controlador en tiempo real de un sistema físico (UN-NITI) mediante algoritmos secuenciales e inteligentes.

El nitinol se presenta hoy en día como una buena alternativa para producir movimiento sin la utilización de motores u otro tipo de dispositivos, las aplicaciones de este material pueden ir desde áreas tan complejas como la electromedicina y el control, hasta el diseño de avisos publicitarios dinámicos, siendo seguramente menos costosos en términos de mantenimiento y consumo de energía, sin embargo algunas limitaciones que tiene el nitinol es su baja eficiencia en la transformación de energía eléctrica a mecánica y su bajo nivel de deformación (8% máximo), lo cual posiblemente sea superado gracias a investigaciones en esta área.

La flexibilidad del sistema de control del minirobot permite conectarlo a la HP 48GX, a un PC o a la reciente HP 49G dependiendo de la disponibilidad de hardware de la persona que desee utilizar a UN-NITI. Estas características son deseables para nutrir con plantas los laboratorios de control e inteligencia artificial de universidades con recursos físicos limitados.

REFERENCIAS

- [1] Carlier, Sébastien; GieBelink, Christoph (1999): «EMU 48 Ver. 1.09». URL: <http://www.gulftel.com/~pattersc/win48/>
- [2] Conrrad, James M.; Mills, Jonathan W. (1998): «STIQUITO, advanced experiments whit a simple and inexpensive robot». Los Alamitos, CA., IEEE computer Society.
- [3] Delgado, Alberto. (1988): «Inteligencia Artificial y Minirobots». Santafé de Bogotá, Colombia, Ecoe Ediciones.
- [4] Everett, H. R. (1995): «Sensors for Mobile Robots, theory and application». Wellesley, Massachusetts, A K Peters.
- [5] Foelix, Rainer F. (1982): «Biology of spiders». London, England, Harvard University Press.
- [6] Gilbertson, Roger G. (1994): «Muscle Wires, Project Book». San Anselmo, CA., Mondo-Tronics, Inc.
- [7] Hewlett Packard. (1994): «HPG series, Advanced User's reference manual». Corvallis, OR, U.S.A., Hewlwt Packard.
- [8] Hurtado, Jaime. (1996): «Aplicaciones con Microcontroladores». Santafé de Bogotá, Colombia, Proyecto de grado en Ingeniería de Sistemas, Universidad Nacional.
- [9] Jones, Joseph L.; Flynn, Anita M. (1993): «Mobile Robots, inspiration to implementation». Wellesley, Massachusetts, A K Peters.
- [10] Rojas, Sergio A. (1998): «Disertación teórica sobre simulaciones inspiradas biológicamente para el estudio del comportamiento adaptativo». Santafé de Bogotá, Colombia, Proyecto de grado en Ingeniería de Sistemas, Universidad Nacional.
- [11] Teuwen, Philippe (1997): «Guide to the HP48G/GX Hardware». Version 0.90.
URL: <http://freezone.exmachina.net/doegox/Default.html>,
e-mail:Philippe.Teuwen@student.ulg.ac.be

REDES DE DATOS PARA USUARIOS MÓVILES. ESTUDIO DE LA ARQUITECTURA MOBILEIP/CELLULARIP

Marta Bordes¹, Anna Calveras

marta.bordes@cselt.it, acalveras@mat.upc.es

Grup de Comunicacions Mòbils

Departament de Matemàtica Aplicada i Telemàtica (DMAT)

Universitat Politècnica de Catalunya (UPC)

BARCELONA, ESPANYA

Paolo Pellegrino

Infrastructures, Switching Platform, Switching Systems

CSELT (Centro Studi e Laboratori Telecomunicazioni)

TORINO, ITALY



La creciente difusión de terminales portátiles está llevando a un crecimiento vertiginoso del número de usuarios que piden poder acceder a Internet o a su propia red empresarial independientemente de su posición geográfica y de la tecnología de red disponible en el acceso (LAN, PTSN, GSM...). Tanto los operadores de telefonía móvil como los proveedores de servicios de Internet (ISP), por tanto, están cada vez más interesados en satisfacer la demanda de servicios de datos para usuarios móviles. Pero para conseguir esto son necesarios nuevos protocolos que permitan el acceso remoto del terminal móvil y la continuidad de las comunicaciones durante el movimiento. Las soluciones propuestas son varias y provienen fundamentalmente de dos ámbitos tan diversos como Internet o el entorno radiomóvil GSM y su ulterior evolución UMTS. Todo esto hace pensar en un escenario futuro en el que las compañías ofrecerán servicios de datos utilizando diversas redes de acceso con diferentes protocolos. Será necesario, por tanto, un protocolo que gestione la movilidad de una red a otra. Una posible solución que está teniendo una gran aceptación en el ámbito IETF es el uso de MobileIP / CellularIP.

NUEVAS EXIGENCIAS DE MOVILIDAD

La telefonía móvil e Internet son las dos áreas que están teniendo un mayor desarrollo en estos últimos años en el mercado de las telecomunicaciones. La evolución de los terminales móviles nos hace pensar que probablemente en un futuro se creen nuevas necesidades intentando aprovechar la ventaja de no tener que conectar el terminal a la línea telefónica. Situaciones que hoy en día no son posibles, como el poder descargar un fichero de Internet mientras se camina por la calle puede llegar a convertirse en una práctica bastante común en el futuro.

La demanda de servicios de datos para usuarios móviles también se extiende al mundo empresarial. Las grandes empresas tienen empleados que trabajan normalmente a distancia, o que, debido al tipo de trabajo que realizan, están obligados a viajar frecuentemente. Sería muy interesante en estos casos el poderse conectar de manera remota a Internet y a su red empresarial mediante un PC con un interfaz wireless de modo que pudieran disponer de los recursos de la empresa y efectuar comunicaciones o transferencias de ficheros aun encontrándose en una localización diversa de las oficinas de la empresa, o incluso viajando en un tren.

La conexión remota a las redes empresariales ya es posible mediante tecnologías como las Redes Privadas Virtuales (VPN) [<http://www.vpnc.org>]. Las VPN permiten un acceso transparente a los recursos de la empresa de una manera económica, utilizando la red Internet o el

backbone de la red del ISP (Internet Service Provider), y mecanismos de tunnelling seguro, utilizando protocolos como L2TP [RFC2661] o IPSec [RFC2401]. No obstante, ésta es una solución válida para terminales portátiles con un interfaz de red tradicional (por ejemplo, Ethernet), y no para terminales móviles que, a parte de conseguir el acceso seguro, necesitan mecanismos que garanticen la continuidad de las comunicaciones incluso durante el movimiento.

Actualmente ya existen algunas soluciones que vienen del ámbito de las comunicaciones radiomóviles. A la oferta de servicios de voz se pretende añadir la de servicios de datos. No obstante, hay otras soluciones basadas en el protocolo IP [RFC791] que están teniendo gran aceptación debido a que poseen grandes ventajas respecto a aquellas basadas en radioenlaces.

SERVICIOS DE DATOS PARA USUARIOS MÓVILES. SITUACIÓN ACTUAL

Soluciones propuestas por los operadores radio

En los últimos años los operadores radiomóviles han empezado a añadir la oferta de servicios de datos a aquella de servicio de voz. Sobre la red GSM actualmente están disponibles los servicios de mensajería SMS (Short Message Service) y también el servicio de transmisión de datos sobre un circuito dedicado a baja velocidad CSD

¹ Actualmente realizando el Proyecto Final de Carrera de la titulación de Ingeniería Superior de Telecomunicaciones en CSELT (CSELT - Centro Studi e Laboratori Telecomunicazioni)



(Circuit Switched Data), que permite el acceso a Internet a los usuarios provistos de un teléfono móvil. A corto plazo, se propone GPRS (General Packet Radio Service), que es una solución basada en tecnología de paquetes diseñada para soportar IP y X.25, y con el objetivo de proveer de más velocidad al usuario. Con el GPRS se podrá ofrecer al usuario una velocidad máxima de 171 kbps.

Como solución a medio plazo tenemos los sistemas móviles de tercera generación UMTS (Universal Mobile Telecommunications System) que adoptarán un nuevo acceso radio W-CDMA que permitirá la transferencia de datos a velocidad muy superior respecto a los sistemas actuales, hasta un máximo de 2Mbps en ambientes indoor.

La ventaja de estas soluciones es que los operadores radiomóviles ya poseen la infraestructura de redes celulares. Pero por otra parte tienen las desventajas asociadas a los sistemas basados en radioenlaces, que son la limitada capacidad de banda y el bajo aprovechamiento de los recursos radio.

Soluciones propuestas por los proveedores de servicios de Internet

Actualmente, los ISP responden a las estas nuevas exigencias de los usuarios ofreciendo un conjunto de servicios dial-up que incluyen el acceso remoto a Internet pero también la posibilidad de acceder de manera segura a una red empresarial a través de protocolos de tunnelling como L2TP o IPsec. Este es el caso de las VPN, de las que se ha hablado anteriormente. Por otra parte, muchos ISP se reúnen en confederaciones para ofrecer a los usuarios servicios de roaming entre las redes de las organizaciones miembros.

Una ulterior evolución que estudian los ISP es la posibilidad de ofrecer acceso wireless a los usuarios que tengan exigencias de movilidad, de este modo se les puede permitir seguir conectados incluso durante el movimiento. Las opciones disponibles incluyen la utilización de soluciones simples y económicas pertenecientes a la categoría de las Wireless LAN (IEEE 802.11 [<http://grouper.ieee.org/groups/802/11/index.html>], Bluetooth [<http://www.bluetooth.com>], HomeRF [<http://www.homerf.org>], etc.) en ambientes indoor y la utilización de la cobertura wireless proporcionada por los operadores radiomóviles en ambientes outdoor urbanos y extraurbanos.

Pero el inconveniente principal de estas soluciones es el hecho de que la movilidad del terminal se realiza a nivel IP, pero se gestiona casi completamente por la infraestructura de red wireless, utilizando mecanismos propietarios. Esto puede comportar un direccionamiento no óptimo del tráfico de datos, pero sobretodo no permite

el roaming transparente entre las redes de acceso que utilizan tecnologías heterogéneas.

Con el objetivo de superar los problemas mencionados en el apartado anterior, la IETF (Internet Engineering Task Force) está estudiando actualmente nuevos protocolos para gestionar la movilidad a nivel de red. [<http://www.ietf.org/html.charters/mobileip-charter.html>]

EL PROTOCOLO MOBILEIP

MobileIP (MIP) [RFC2002, RFC2003, RFC2004] es la solución estándar propuesta por el IETF para gestionar la movilidad de terminales entre subredes IP. El objetivo de MobileIP es permitir a un host cambiar de manera transparente el punto de conexión a Internet. MIP es un protocolo que trabaja a nivel IP actuando sobre el direccionamiento de los paquetes dirigidos hacia el nodo móvil y por este motivo está en grado de gestionar fácilmente la movilidad entre redes independientemente del tipo de acceso. La principal característica de MobileIP es que es independiente a la aplicación usada, ya que no cambia las direcciones IP saliente y destino, con lo cual las sesiones a nivel de transporte TCP [RFC793] activas no se interrumpen durante el movimiento. A continuación se especificará el funcionamiento de MobileIPv4, aunque también está en desarrollo MobileIPv6, es decir el correspondiente a la versión 6 del protocolo IP [RFC2460], que en realidad es análogo al anterior y difieren en solo unas pocas características.

FUNCIONAMIENTO DE MOBILEIP

El nodo móvil, también llamado Mobile Host (MH) tiene asignadas dos direcciones IP. La primera es propia de su red originaria, se llama Home Address (HAddr), y no cambia en todo el proceso. La segunda es la llamada Care-of Address (COAddr) y generalmente es la dirección IP de un router o estación base radio de la red visitada, que adquiere el papel de Foreign Agent (FA). Para que el terminal móvil pueda continuar comunicándose utilizando su propia dirección IP, es necesario que un router de su subred original adquiera el papel de Home Agent (HA). La Figura 1 muestra el esquema de funcionamiento.

Los llamados mobility agents (es decir, los routers que hacen el papel de Home Agent y Foreign Agent) envían periódicamente una serie de mensajes llamados Agent Advertisement, que son señales de beacon que permiten al terminal móvil conocer su localización actual y por tanto darse cuenta de si se ha movido o no respecto a su localización anterior. Si el MH se ha desplazado comienza un proceso de intercambio de paquetes entre el mismo y su Home Agent, y en el cual el Foreign Agent también está involucrado de manera que, al final del proceso, el HA tiene registrada la nueva posición del MH.

Cuando un nuevo terminal envía paquetes destinados al MH éstos, mediante enrutamiento IP standard,

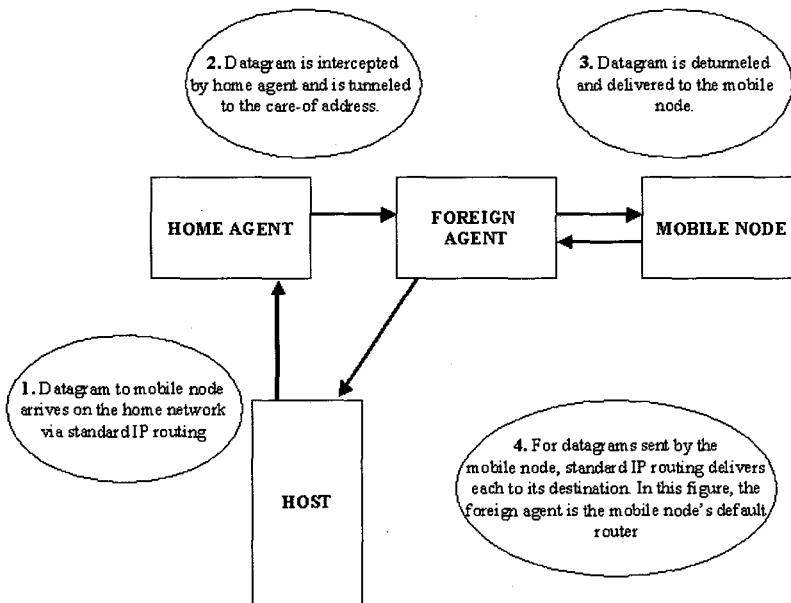


Figura 1. Esquema de funcionamiento del protocolo MobileIP

llegaran a la subred originaria del móvil y serán captados por el HA. Este último ya sabe que el MH no está conectado a su subred, sino que está conectado a la subred perteneciente al FA, así que encapsula el paquete añadiendo una nueva cabecera IP (encapsulado IP en IP) en la cual la dirección destino será la COAddr. Mediante ese tunnelling los paquetes llegarán a la red del FA, éste los desencapsulará y se los entregará al MH. Por otra parte, el nodo móvil puede responder a la estación transmisora de modo directo utilizando como dirección saliente la Home Address.

GESTIÓN DE LA MICROMOVILIDAD EN REDES WIRELESS

Con MobileIP es posible acceder a cualquier red con un terminal wireless manteniendo la dirección IP original, y por tanto, siendo alcanzable por cualquier otro terminal que quiera establecer una comunicación con él mediante routing IP. Esto significa que el usuario, una vez que se conecte a la red y empiece a mantener una comunicación, podrá tener la necesidad de moverse y en ese caso se debe mantener la comunicación mientras se está produciendo el desplazamiento. Este requisito se traduce en la necesidad de ser capaz de llevar a término los procedimientos de gestión de la movilidad en el mínimo tiempo posible y con la mínima pérdida de datos.

En una red wireless se pueden llevar a cabo dos tipos de movimientos (handoff) que deben ser gestionados del modo oportuno para poder garantizar la continuidad de las comunicaciones. Uno de ellos es el movimiento entre estaciones base que pertenecen a la misma subred (handoff de nivel 2) y otro es el movimiento entre estaciones base pertenecientes a subredes distintas (handoff de nivel 3). El

handoff de nivel 2 se puede gestionar de modo rápido utilizando protocolos de señalización ad-hoc para la coordinación de estaciones base adyacentes.

En cambio, un handoff de nivel 3 implica un cambio de subred IP y por tanto requiere modificar el direccionamiento de los paquetes que van hacia el nodo móvil. La Figura 2 muestra los niveles 2 y 3 del handoff.

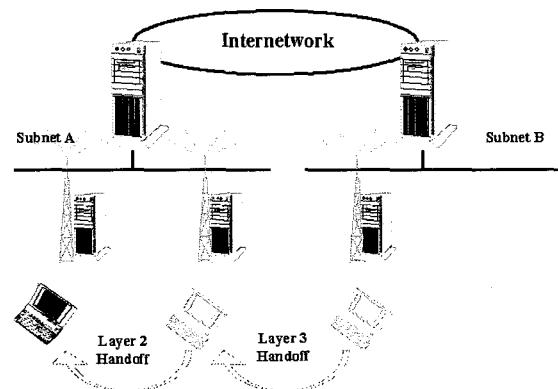


Figura 2. Handoffs de nivel 2 y 3

El protocolo MobileIP, como se ha visto anteriormente, realiza esto mismo a través de un mecanismo de tunnelling. Esta modalidad de funcionamiento puede llegar a provocar una pérdida bastante grande de paquetes durante el handoff, ya que el procedimiento de registro con el Home Agent puede requerir un tiempo de latencia mayor que un segundo, durante el cual los paquetes que estaban ya viajando se suelen perder. A parte de esto está



el problema de la gran carga de tráfico de señalización que provoca un nuevo registro con el Home Agent.

Existen ya varias propuestas para solucionar las limitaciones del protocolo MobileIP en cuanto a la gestión de micromovilidad y movilidad frecuente. Estos protocolos pueden dividirse en dos categorías:

- * En la primera categoría aparecen los protocolos que gestionan el direccionamiento del tráfico hacia los terminales móviles en roaming mediante protocolos de routing ad-hoc que no requieren la presencia de agentes de movilidad en el interior del dominio. La idea es tener solamente un Foreign Agent localizado en el router de a bordo del sitio visitado y hacerlo actuar de manera que parezca que este único Foreign Agent y el terminal móvil estén siempre localizados en la misma LAN, independientemente de la posición real del móvil en el interno del sitio.
- * En la segunda categoría aparecen los protocolos que, manteniendo la necesidad de tener un Foreign Agent en cada una de las subredes IP, gestionan localmente (lo cual significa, sin informar al Home Agent) todos los mensajes de registro MobileIP que el terminal móvil transmite después de cada movimiento.

CellularIP es un ejemplo de protocolo de gestión de la micromovilidad perteneciente a la primera categoría, y es una solución que está adquiriendo un gran consenso en el ámbito IETF. Otros protocolos que se están desarrollando son HAWAII [RLT99] (perteneciente al primer grupo) y Regional Tunnel Management [GJP98] y THEMA [MHW99], pertenecientes al segundo grupo.

A continuación vamos a especificar el funcionamiento de CellularIP.

CELLULAR IP

CellularIP [CGW00] es un protocolo optimizado para redes de acceso wireless y para usuarios que se mueven muy velozmente. Una red CellularIP es en general una red de routers IP, algunos de los cuales poseen además una o varias interfaces wireless a través de las cuales los usuarios pueden acceder al servicio. Esta tecnología es independiente de la tecnología de acceso radio utilizada. En este caso, el border gateway mediante el cual la red CellularIP se conecta a Internet es el que juega el papel de Foreign Agent y, como tal, desencapsula los paquetes recibidos mediante tunnelling del Home Agent y los retransmite hacia las estaciones radio base. Del mismo modo los paquetes generados por los terminales móviles van hacia el gateway, y éste los transmite hacia Internet.

El terminal móvil, cuando se encuentra en una de estas redes CellularIP, mantiene como dirección IP su dirección original, es decir, la Home Address. El tráfico dirigido hacia el móvil viene direccionado mediante un protocolo ad-hoc. Cada router de la red posee una tabla de routing en la que se guarda información sobre la ruta a seguir a fin de alcanzar al nodo móvil. Las tablas son actualizadas por unos paquetes de routing cache o de paging cache. Estos paquetes son enviados por el terminal móvil de una manera periódica a la estación base más cercana, y ésta se ocupa de actualizar la tabla y enviárselos al siguiente router en el camino hacia el gateway, y así sucesivamente actúan todos los routers de la red CellularIP.

Se distinguen dos tipos de paquetes dependiendo de si el móvil está en estado activo (recibiendo paquetes o enviando tráfico de datos) o inactivo. En el primer caso se envían los paquetes de routing cache y en el segundo de paging cache. Se hace esta distinción porque los paquetes de routing cache son enviados con mucha más frecuencia

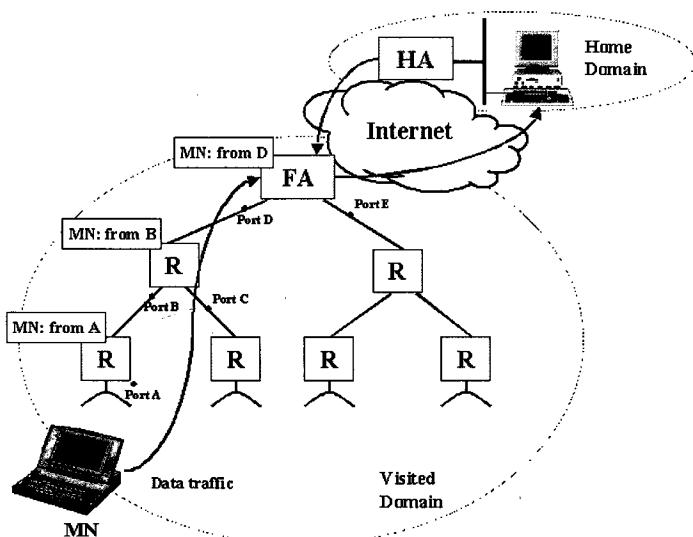


Figura 3. CellularIP

que los de paging, ya que cuando un móvil está activo es mucho más importante conocer su localización en todo momento.

Cuando el terminal móvil se desplaza, envía los paquetes a la nueva estación base y ésta inicia de nuevo el procedimiento de direccionarlo hacia el gateway. Las tablas actualizan los valores, y aquellas que no los actualicen borrarán el valor al cabo de un tiempo. Por tanto, los paquetes dirigidos al terminal móvil tomarán la nueva ruta de una manera automática. La Figura 3 muestra una posible estructura de CellularIP.

Cuando el gateway recibe un paquete dirigido al MN, consulta sus tablas de routing para ver si hay una entrada con esa dirección destino. Si la hay, va encaminando el paquete salto a salto consultando las tablas.

Hasta aquí hemos presentado como pueden gestionarse de forma independiente los dos niveles de movilidad. A continuación se verá una arquitectura híbrida.

LA ARQUITECTURA BASADA EN MOBILE IP/CELLULAR IP

Como una posible arquitectura de red para gestionar la movilidad de los usuarios a todos los niveles se propone el uso de MobileIP y CellularIP de manera conjunta. En la Figura 4 puede verse la arquitectura basada en ambos protocolos.

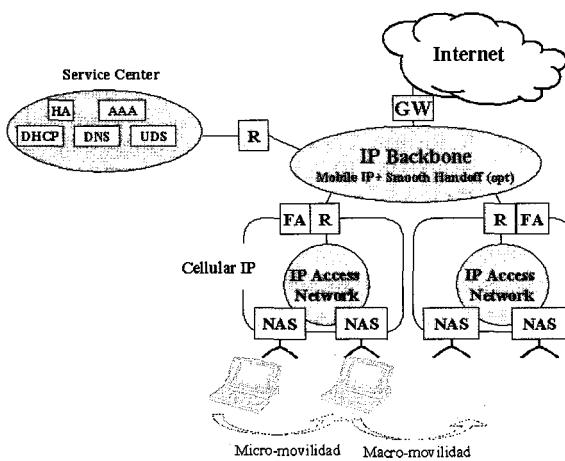


Figura 4. Arquitectura basada en MobileIP/CellularIP

La infraestructura de red comprende un backbone IP que interconecta un conjunto de redes de acceso wireless con tecnología IP. Esta red modela la estructura de la red de un ISP que ofrece servicios de acceso a Internet a particulares (por ejemplo, Teleline, Alehop, Ideo, Jazzfree, etc.), con la diferencia fundamental de que en este caso los NAS (Network Access Server) no terminan en conexiones dial-up, sino que están dotados de una o más interfaces

wireless. El backbone IP, por otro lado, puede interconectarse a Internet a través de un router que realiza las funciones de gateway, sobre el cual podría ser colocada también la funcionalidad de firewall o security gateway.

Una red construida de este modo podría ser:

- * La red conmutada de un operador radiomóvil que ofrece servicios de datos para usuarios móviles en el ámbito metropolitano o su escala nacional. En este caso las redes de acceso interconectadas son normalmente redes de router distribuidas sobre el territorio.
- * Una red corporativa con tecnología Wireless LAN. Las redes de acceso son normalmente las redes intranet con tecnología WLAN que cubren las sedes distribuidas o los diversos edificios de la empresa, mientras que el backbone IP es el que realiza la interconexión con la red backbone empresarial.

El terminal móvil se conecta mediante su interfaz wireless a unas redes de acceso IP que utilizan el protocolo CellularIP, el cual gestiona la movilidad en el interior de estas redes de acceso. Por otro lado, MobileIP se ocupa de la movilidad a nivel global, es decir, entre diferentes redes de acceso.

Esta es una solución basada totalmente en IP. Naturalmente hay otras alternativas para realizar una red wireless con estas características.

ESCENARIO FUTURO

Muy probablemente, la falta de regulación a la cual se está asistiendo en el mercado de las comunicaciones va a llevar a la presencia de diversos tipos de operadores con una única diferenciación en términos de servicios ofrecidos y de tecnologías utilizadas.

El escenario más probable en el futuro estará, por tanto, caracterizado por la coexistencia de varios tipos de redes wireless geográficamente superpuestas, pensadas para responder a necesidades específicas de los usuarios y por tanto, diferenciadas entre ellas en base a la banda disponible, a las zonas de cobertura, a las tecnologías de acceso radio y a los protocolos de interconexión. En un escenario de estas características, será bueno que el usuario pueda decidir cuándo moverse de un acceso wireless a otro, decisión que tomará en base a la disponibilidad o a las exigencias de calidad de servicio de una aplicación específica.

Actualmente no es posible que un usuario se mueva de modo transparente entre redes superpuestas, dado que a cada cambio de red wireless le viene asignada una dirección IP diferente.

Se espera que en el escenario futuro será posible realizar la interconexión entre redes wireless heterogéneas a través de un backbone IP común (por ejemplo Internet)



y en este escenario el protocolo MobileIP podrá ser reutilizado eficazmente para gestionar la movilidad del usuario entre una red y otra.

Por otro lado será necesario utilizar un protocolo común para gestionar la movilidad en el interior del dominio. Naturalmente pueden identificarse numerosas arquitecturas para redes de acceso wireless. Por una parte los ISP buscan una solución completamente basada en IP. En cambio, los operadores radiomóviles tienen diferentes alternativas, como el sistema GPRS simple, o un sistema GPRS evolucionado que será utilizado por el UMTS.

La solución basada totalmente en IP es sin duda la más revolucionaria y también la que presenta unas ventajas más interesantes, entre las cuales merece citar el routing óptimo a nivel IP, una integración más simple y completa con la red Internet y las otras redes IP fijas y el hecho de que no es necesaria la utilización de la señalización SS7. Por otra parte, la adopción de una solución arquitectural completamente IP deja abierta la posibilidad de ofrecer los servicios de voz y datos con una sola infraestructura de redes, aprovechando tecnologías emergentes como Voz sobre IP.

RESUMEN

En este artículo se ha presentado una solución para la gestión de movilidad en redes IP de datos. Para ello se han tratado los temas de MobileIP y CellularIP, como alternativas en desarrollo por los grupos de trabajo de la IETF.

En un futuro a medio plazo se espera una gran evolución del mercado de los servicios de datos para usuarios móviles, provocada por la alta demanda de estos servicios debida al desarrollo de los terminales portátiles. Esta evolución llevará muy probablemente a la coexistencia de un gran número de redes superpuestas con características diversas y gestionadas por entidades diferentes (ISP, operadores radiomóviles, etc.).

MobileIP es un protocolo propuesto por el IETF cuyo funcionamiento es óptimo para la gestión de la macromovilidad o movilidad poco frecuente. Es por tanto, un protocolo que puede gestionar la movilidad entre esas redes de acceso diversas de las que se ha hablado anteriormente.

Para la gestión de la micromovilidad o movilidad frecuente, es decir, el desplazamiento del terminal móvil en el interior de las redes de acceso hay diferentes alternativas. CellularIP es una de estas propuestas.

La arquitectura MobileIP/CellularIP se caracteriza por ser una solución basada totalmente en IP, lo cual permite una integración óptima con los servicios de Internet, entre otras ventajas. Por otra parte, la adopción de una solución arquitectural toda IP deja abierta la posibilidad

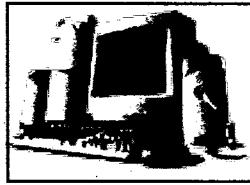
de ofrecer los servicios de voz y datos con una sola infraestructura de redes, aprovechando tecnologías emergentes como Voz sobre IP.

REFERENCIAS

- [CGW00] A. Campbell, J. Gomez, C-Y. Wan, S. Kim, Z. Turanyi, A. Valko. CellularIP. Enero 2000, <[draft-ietf-mobileip-cellularip-00.txt](#)>. Status: INTERNET DRAFT
- [GJP98] Eva Gustafsson, Annika Jonsson, Charles, E. Perkins. Mobile IP Regionalized Tunnel Management. [draft-ietf-mobileip-reg-tunnel-00.txt](#) Noviembre 1998. Status: INTERNET DRAFT.
- [MHW99] Pete McCann, Tom Hiller, Jin Wang, Alessio Casati, Charles E. Perkins, Pat R. Calhoun. Transparent Hierarchical Mobility Agents (THEMA). Marzo 1999. [draft-mccann-thema-00.txt](#). Status: INTERNET DRAFT.
- [RFC2002] C. Perkins. IP Mobility Support. Octubre 1996. Status: PROPOSED STANDARD
- [RFC2003] C. Perkins. IP Encapsulation within IP. Octubre 1996. Status: PROPOSED STANDARD
- [RFC2004] C. Perkins. Minimal Encapsulation within IP. Octubre 1996. Status: PROPOSED STANDARD
- [RFC2401] S. Kent, R. Atkinson. Security Architecture for the Internet Protocol. Noviembre 1998. Status: PROPOSED STANDARD
- [RFC2460] S. Deering, R. Hinden. Internet Protocol, Version 6 (IPv6) Specification. Diciembre 1998. Status: DRAFT STANDARD
- [RFC2661] Layer Two Tunneling Protocol «L2TP». W. Townsley, A. Valencia, A. Rubens, G. Pall, G. Zorn, B. Palter. Agosto 1999. (Status: PROPOSED STANDARD)
- [RFC791] J. Postel. Internet Protocol. Sep-01-1981. Status: STANDARD
- [RFC793] J. Postel. Transmission Control Protocol. Sep-01-1981. Status: STANDARD
- [RLT99] R. Ramjee, T. La Porta, Thuel, K. Varadhan, L. Salgarelli. IP micro-mobility support using HAWAII. [draft-ietf-mobileip-hawaii-00.txt](#). Junio 1999. Status: INTERNET DRAFT.

BIBLIOGRAFÍA

- * J. Solomon. Applicability Statement for IP Mobility Support ([rfc2005.txt](#)), October 1996
- * Charles E. Perkins. Mobile IP. Design Principles and Practices. Addison-Wesley. Wireless Communications Series. Reading, MA: Addison Wesley Longman, 1997
- * Andras G. Valko. CellularIP. A New Approach to the Internet Host Mobility
- * Andrea Calvi, Ivano Guardini. E01, Definizione delle attività su Mobile IP in CSELT. Documento Tecnico Informativo CSELT.
- * Paolo Fasano, Domenico Mazzei, Mobilità in reti IP. Documento Tecnico Informativo CSELT
- * David B. Johnson, David A. Maltz, Protocols for Adaptive Wireless and Mobile Networking. IEEE Personal Communications, 3, February 1996.



OBJETOS DISTRIBUIDOS SOBRE TCP/IP

José Luis Mateo Terrés¹, José Antonio Gonzalez Sanchez²

¹ E.T.S. Ingenieros de Telecomunicación Valencia, mateo@ieee.org

² Ingeniería Informática Universidad de Alicante, jagonzalez@ua.es

1. APPLICACIONES DISTRIBUIDAS

1.1 Introducción

Las aplicaciones distribuidas (AD de aquí en adelante) constituyen actualmente un campo de actuación muy interesante. Se componen de varios programas ejecutándose en diferentes computadoras y comunicándose entre ellos. Hasta ahora la mayoría de las AD eran bases de datos distribuidas, en las que un cliente accede a un servidor virtual (compuesto en realidad de varios servidores reales) y realiza una petición, obteniendo una respuesta del servidor real correspondiente, sin importarle de cual de ellos se trata. El cliente simplemente conoce el nombre del servidor virtual. Esta comunicación se realiza a través de los sockets o los RPC's (). Las AD vienen a solucionar los problemas de las centralizadas, como la total dependencia del servidor central, el elevado coste de éste frente al de unos cuantos pequeños ordenadores o el hecho de que la información ha de pasar siempre por el servidor aunque sea local. Hace unos años ha surgido, de la fusión de las AD con la programación orientada a objetos (POO) el concepto de Objetos Distribuidos, en los que cambia el concepto tradicional de cliente/servidor.

1.2 Problemas

Los principales problemas que encontramos en las aplicaciones distribuidas tienen su origen en la variedad de plataformas utilizadas tanto por el cliente como por el servidor o servidores. Las distintas plataformas pueden diferir en el lenguaje de implementación, el Sistema Operativo, la red utilizada para la interconexión y en el Hardware de la máquina. Esta falta de homogeneidad hace difícil el entendimiento entre las distintas partes de la aplicación, y éstas deben tener en cuenta todas y cada una de las particularidades de las otras partes.

Estos problemas han de ser resueltos satisfactoriamente para la implantación de aplicaciones distribuidas, y para ello surgen diversas plataformas las cuales trataremos más adelante, en especial la que mejores prestaciones ofrece, CORBA.

1.3 Objetos distribuidos

El concepto de objetos distribuidos (de ahora en adelante OD) surgen, como ya hemos dicho, de la unión de las AD y la POO. Esto viene a romper con el tradicional concepto cliente/servidor, ya que la frontera entre los mismos queda ahora mucho más confusa. Si trabajamos con objetos, puede que uno de ellos (cliente) realice una llamada a otro (servidor). A su vez, este último objeto puede utilizar los servicios de otro, con lo que se convierte en cliente. Con esto ya no tenemos un conjunto de clientes y servidores sino un completo sistema de información, como si tuviéramos todos los objetos en una misma máquina (ahora es la red) funcionando con el mismo programa (aplicación distribuida).

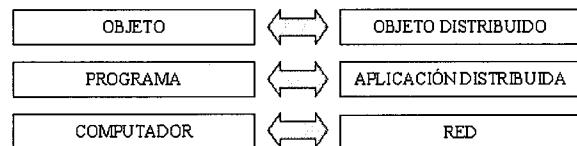


Figura 1

Si a las ventajas de la POO le unimos la posibilidad de tener cada uno de los objetos en una localización geográfica diferente, y además que estén implementados con el lenguaje, sistema operativo, hardware y red de interconexión más adecuado a la naturaleza del objeto y a su localización, tenemos las ventajas de las aplicaciones globales y locales al mismo tiempo. Para ello surgen las plataformas distribuidas, un *middleware* que intentan hacer transparente a cada uno de los objetos las características de implementación de los demás, así como su localización geográfica.

1.4 Diferentes plataformas para objetos distribuidos. Comparativa

Aquí comentamos brevemente diversas opciones a ahora de implementar nuestro sistema, a parte de CORBA que desarrollaremos más tarde:



Característica	CORBA/IOP	DCOM	RMI/RMP	HTTP/CGI	Servlets	Sockets
Nivel de abstracción	*****	*****	*****	**	**	*
Integración con Java	*****	*****	*****	**	**	**
Soporte de plataforma OS	*****	**	*****	*****	*****	*****
Implementación todo Java	*****	*	*****	**	*****	*****
Soporte de parámetros tipo	*****	*****	*****	*	*	*
Facilidad de configuración	***	***	***	***	***	***
Invocaciones del método distribuido	*****	***	***	*	*	*
Invocaciones remotas	*****	***	***	*	**	**
Soporte de metadatos	*****	***	***	*	*	*
Invocaciones dinámicas	*****	*****	*	*	*	*
Rendimiento (Pings remotos)	***	***	***	*	*	*****
	3.5 msecs	3.8 msecs	3.3 msecs	827.9 msecs	55.6 msecs	2.1 msecs
Seguridad	*****	*****	***	***	***	***
Transacciones	*****	***	*	*	*	*
Referencias a objetos persistentes	*****	*	*	*	*	*
Naming basado en URL	*****	**	**	*****	*****	***
Invocaciones a objetos multilenguaje	*****	*****	*	**	*	***
Protocolo de lenguaje neutral	*****	*****	*	*****	*****	*
Escala intergaláctica	*****	**	*	**	**	***
Estándar abierto	*****	**	**	***	**	***

· Java Sockets

Es la tecnología para la programación en red con Java. Hasta hace poco, era la única manera de escribir aplicaciones cliente/servidor en Java.

· HTTP/CGI

Es el modelo predominante para la creación de soluciones cliente/servidor para Internet hoy en día. HTTP proporciona semántica simple como RPC por encima de los Sockets. CGI proporciona un protocolo para entregar un comando HTTP a una aplicación de servidor.

· Servlets

Son los plug-ins de la parte servidor de Java.

· RMI

Es un ORB (Object Request Broker) nativo de Java, creado por JavaSoft. Introduce nueva semántica para objetos distribuidos en Java. Esta semántica forma la base del planteamiento CORBAJava-to-IDL.

· Caffeine

Creado por Netscape/Visigenic proporciona un medio natural de programación parecido a RMI por encima de VisiBroker para Java. Permite escribir objetos distribuidos CORBA sin IDL.

· DCOM

Es el otro destacado ORB. Es un gran competidor para CORBA porque el todopoderoso Microsoft lo está incluyendo con todos sus sistemas operativos. Es también la tecnología fundamental para ActiveX y Microsoft Internet. Microsoft's Visual J++ introduce DCOM para Java, y permite la invocación remota de un objeto Java por otro usando DCOM ORB.

2. CORBA (COMMON OBJECT REQUEST BROKER ARCHITECTURE)

2.1 Introducción

CORBA es un estándar desarrollado por el OMG (Object Management Group) para conseguir la interacción entre aplicaciones distribuidas independientemente de la

plataforma de ejecución y el lenguaje utilizados en cada una de ellas. Esta transparencia se consigue gracias a dos piezas fundamentales: el ORB (Object Request Broker), encargado de la traducción y el transporte, y el IDL (Interface Definition Language), interfaz universal a través del cual los objetos especifican qué servicios pueden ofrecer, de forma independiente a su propio lenguaje de programación.

2.1.1 OMG

El Object Management Group es un consorcio industrial internacional que promueve la teoría y la práctica del desarrollo de software orientado a objetos. Surge porque las grandes compañías toman conciencia de la gran aceptación, cada día mayor, que tiene la POO. El OMG fue fundado en 1989 por ocho compañías: 3Com Corporation, American Airlines, Canon Inc. Data General, Hewlett-Packard, Philips Telecommunications N.V., Sun Microsystems y Unisys Corporation. Actualmente aglutina a más de 500 compañías.

Este organismo ha publicado las especificaciones de CORBA en el Object Management Architecture (OMA) para lograr una compatibilidad entre los distintos fabricantes. Aquí la OMG no define la implementación de los elementos de la arquitectura CORBA, sino solamente sus interfaces y funciones. Esto evita la exclusión de cualquier tecnología o lenguaje.



Figura 2

2.1.2 Soluciones aportadas por CORBA

CORBA nos ofrece una serie de prestaciones que no nos ofrecen otras posibles soluciones, enfocadas en diferentes aspectos:

Integración: cada objeto de la aplicación estará representado por una interfaz IDL (Interface Description Language). Existe un compilador de IDL para cada lenguaje de programación diferente, pero el interfaz que creará será independiente del lenguaje utilizado. A partir de ahí el objeto y sus métodos serán invocados a través de su IDL, por aplicaciones escritas en cualquier lenguaje.

Comunicación: se realiza a través del ORB (Object Request Broker), un agente que se encarga de controlar las peticiones a los objetos. Para el diseñador es transparente el mecanismo de comunicación y la localización de los servidores de los objetos que precisa. Dichos servidores pueden ser de distinta clase y CORBA se encarga de repartir el trabajo entre ellos. Si uno cae, los clientes que lo utilizaban pasarán a ser servidos por otros servidores. En caso de no haber ninguno listo, la aplicación no se colgará.

Además de esto CORBA es un estándar aceptado y reconocido por OSI del que existen productos comerciales

2.1.3 OMA

Como ya hemos dicho la arquitectura CORBA está definida en la Object Management Architecture (OMA) en cuatro partes fundamentales:

Application Objects: se refiere a la implementación, mediante IDL, de un interfaz propio de cada objeto e independiente del lenguaje. Se utiliza para comunicarse, a través del ORB, con los demás elementos de la arquitectura o con otros objetos.

CORBA services: son los elementos que permiten la comunicación entre distintos objetos de forma transparente. Esto permite que no se incluyan en la implementación de un objeto detalles sobre los demás, sino que estos datos residan en la propia arquitectura CORBA. En principio se definieron en las llamadas Common Object Services Specification (COSS) a partir de los Request for Proposals (RFP). Los servicios básicos ofrecidos son los siguientes:

Naming Service: es el típico servicio de nombres de una red, que permite asociar un conjunto de caracteres con un objeto en particular.

Object Event Notification Service: permite el paso de mensajes entre objetos. El generador del mensaje se define como *supplier* y el destinatario como *consumer*.

Object Lifecycle Service: define unos mecanismos para crear, borrar, mover y copiar objetos.

Persistent Object Service: proporciona interfaces comunes a los mecanismos usados para la retener y gestionar el estado persistente de objetos en datos almacenados de manera independiente

Concurrency Control Service: este servicio define como un objeto simultáneamente accedido por uno o más clientes actúa para que sus accesos permanezcan consistentes y coherentes.

Externalisation Service: define estándares para el almacenamiento y al recuperación del estado de objetos en un *stream* de datos.

Object Relationship Service: permite crear, borrar y manejar relaciones entre objetos.

Object Transaction Service: proporciona soporte para asegurar que la computación que se compone de una o más operaciones o sobre uno o más objetos presenta propiedades de atomicidad, consistencia, aislamiento y durabilidad (ACID properties).

Object Security Services: proporciona para los objetos las propiedades de confidencialidad, integridad, responsabilidad y disponibilidad.

Object Time Service: proporciona un mecanismo para sincronizar relojes en sistemas distribuidos.

Object Licensing Service: proporciona un marco para la especificación y la gestión de licencias policía y para la tarificación.

Object Properties Service: asocia dinámicamente un objeto con sus propiedades.

Collection Object Service: proporciona una norma uniforme para la creación y manipulación de grupos de objetos tales como listas, pilas, etc.

Trading Service: permite a un cliente encontrar un objeto a partir de sus propiedades.

Startup Service: proporciona una inicialización correcta del ORB a través de los mensajes adecuados

Todos estos servicios están accesibles a través de interfaces IDL estandarizados por la OMG.

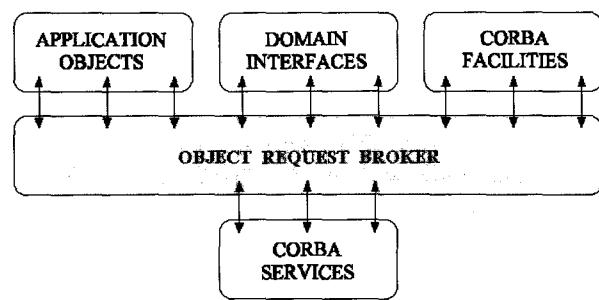


Figura 3. Arquitectura CORBA



CORBA facilities: definen los interfaces de aplicación que utilizan directamente los desarrolladores. Cubren servicios de más alto nivel como puede ser la interfaz de usuario, y es el área en la que la OMG está centrando sus esfuerzos actualmente. Son de carácter horizontal, e incluyen funciones de cobertura a varios sistemas, independientes de su contenido. Se dividen en cuatro grupos:

- User Interface: este grupo cubre todos los aspectos del interfaz de usuario que hace a un sistema de información accesible a sus usuarios y sensible a sus necesidades.
- Information Management: este es un grupo completo que incluye el modelado, almacenamiento, recuperación e intercambio de información.
- System Management: las facilidades de este grupo hacen a los sistemas de computadores responsables y gestionables para cumplir su función. Esto es conseguido mediante un conjunto común de interfaces de sistema operativo para herramientas, aplicaciones y recursos para la gestión.
- Task Management: este grupo proporciona infraestructura para la automatización del trabajo, incluida la automatización de los procesos de dos usuarios y los procesos del sistema.

Domain Interfaces: similares a las anteriores, pero con carácter vertical, pretenden dar servicio a un grupo concreto de usuarios de un determinado ámbito profesional. Esta área no será objeto de estandarización por el OMG, ya que sería una aplicación específica.

2.2 Conceptos Fundamentales

Aquí estudiaremos dos de los conceptos necesarios para comprender la arquitectura CORBA del OMG.

2.2.1 IDL

El Interface Definition Language (IDL) es un lenguaje diseñado por la OMG para definir el tipo de objetos mediante la especificación de sus funciones, independientemente del lenguaje en que estén implementados. El cliente por tanto sólo necesita la definición de funciones de los objetos servidores para invocarlas, sin necesidad de conocer detalles como el lenguaje de programación, la localización del objeto en la red o el sistema operativo sobre el que corre. Esta separación entre interfaz e implementación es similar a la utilizada en la POO.

El IDL es un lenguaje estándar cuya sintaxis está derivada de C++, pero con nuevas sentencias específicas para la programación de objetos distribuidos. Con dicha sintaxis se pueden especificar los atributos de los componentes, las excepciones que se lanzan, los componentes de

los que se hereda, etc. Mediante un traductor se puede transformar en un lenguaje de programación de alto nivel, haciendo la correspondencia entre los tipos de datos CORBA y los del lenguaje en cuestión. Actualmente hay compiladores para Java, C, C++, COBOL y prácticamente para el resto de lenguajes. Por ejemplo, el JDK 1.2 de Sun incluye el compilador para IDL y ORB de forma gratuita.

Este lenguaje permite definir para un interfaz los tipos y constantes exportadas por el objetos (1), las excepciones activadas en caso de error (2), los atributos definidos para ese objeto (3) y las operaciones que ofrece con sus parámetros y tipo de resultado (4). Así mismo permite agrupar interfaces de objetos y definiciones de tipos de datos IDL en módulos (5).

Veamos un ejemplo:

```
// ejemplo de especificación OMG IDL:
// CuentaBanco.idl
module BANCO { // módulo 5)

interface CuentaBanco {
    // types (1)
    enum clase_cuenta {corriente, ahorro};

    // exceptions (2)
    exception cuenta_no_disponible {string razon;};
    exception pin_incorrecto {};

    // attributes (3)
    readonly attribute float saldo;
    attribute clase_cuenta mi_clase_cuenta;

    // operations (4)
    void acceso (in string cuenta, in string pin)
        raises(cuenta_no_disponible, pin_incorrecto);
    void deposito (in float f, out float nuevo_saldo)
        raises (cuenta_no_disponible);
    void reintegro (in float f, out float nuevo_saldo)
        raises (cuenta_no_disponible);
}; // fin de interface CuentaBanco
} // fin de módulo BANCO (5)
```

A continuación, para transformar el anterior código IDL en código Java hay que ejecutar:

idltojava -fno-cpp CuentaBanco.idl

Esto creará cinco ficheros en el directorio Calculadora:

- *CuentaBancoImplBase.java*: implementa el interfaz de la cuenta. Es el *skeleton* CORBA en la parte del servidor.
- *CuentaBancoStub.java*: implementa el interfaz de la cuenta. Es el *stub* CORBA en la parte del cliente.
- *CuentaBanco.java*: es la versión Java del interfaz que habíamos definido mediante IDL.

· *CuentaBancoHelper.java*: permite convertir los tipos de datos CORBA a tipos Java (función *narrow*). Se utiliza en el cliente.

· *CuentaBancoHolder.java*: mantiene una clase pública CuentaBanco de tipo *final*. Se utiliza por clientes y servidores para pasar objetos de tipo CuentaBanco como parámetros.

Una vez que disponemos de estos ficheros, hay que crear el cliente, que guardaremos en un fichero llamado *Cliente CuentaBanco.java*:

2.2.2 ORB

El Object Request Broker (ORB) es la parte fundamental y la que da nombre a la arquitectura CORBA. Éste es un componente software que proporciona soporte al envío y recepción de mensajes por parte de los objetos distribuidos, ocultando al programador la compleja comunicación en la red y la ubicación de los objetos. El ORB de CORBA es muy superior en prestaciones a los de sus competidores. Entre sus características podemos resaltar las siguientes:

- Llamadas estáticas o dinámicas: las llamadas a métodos pueden ser definidas en compilación o en ejecución.
- Independencia del lenguaje de implementación: separación total entre interfaz e implementación a través de IDL
- Localización transparente: una vez localizado un objeto remoto, se comporta igual que si fuese local.
- Autodescripción del sistema: en el *Interface Repository* el ORB contiene información sobre los servidores (funciones y parámetros) para que un cliente los pueda invocar en tiempo real.
- Seguridad y transacciones integradas.
- Coexistencia con sistemas existentes: es posible encapsular código para que parezca un objeto en el ORB

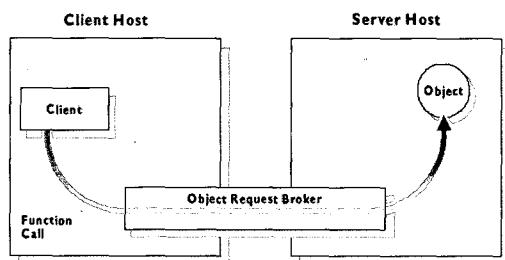


Figura 4. ORB

El ORB no puede recibir llamadas directas de los clientes, sino que éstas deben ser traducidas a través del

interfaz IDL para que el ORB las comunique al servidor, nuevamente a través de IDL. Por tanto el ORB desconoce los objetos. La implementación de un ORB concreto es asunto de fabricantes de software, Iona(Orbix), Visigenic (VisiBroker), Digital (Objectbroker), HP (ORB Plus) y Sun con Neo son solo algunos ejemplos concretos . Desafortunadamente, los más potentes son comerciales.

2.2.3 Estructura de una aplicación CORBA

Vamos a ver como se implementa una aplicación y los componentes necesarios para ello. Comenzaremos por el lado del cliente:

IDL Stub: implementa el enlace entre el cliente y el ORB para el método de invocación estática. De esta manera el cliente sólo puede llamar a los objetos cuyos interfaces son conocidos en tiempo de compilación. La llamada efectuada sobre un objeto es transferida a través del Stub al ORB, y si es la primera vez que accede al mismo deberá obtener una referencia del mismo. Para ello se utiliza el anteriormente comentado *Interface Repository*, que contiene la información sobre el objeto. Un ejemplo de Stub es el que aparece en el apartado referente a IDL.

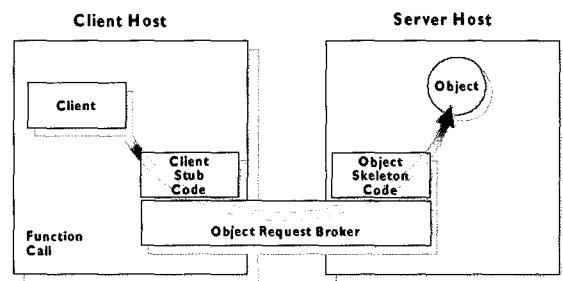


Figura 5. Método de invocación estática.

Dynamic Invocation Interface (DII): implementa el enlace entre el cliente y el ORB para el método de invocación dinámica. De esta manera el cliente puede llamar a objetos de servidores que no nos pertenecen, es decir, de los que no poseemos Stub. Lo que hace este método es crear dicho Stub en tiempo de ejecución, con ayuda del *Interface Repository*. Además permite aislar al cliente de los cambios en la implementación del objeto (programa o localización). La especificación dinámica describe dos modos de invocación, síncrono en el que el cliente se bloquea en espera de la respuesta del servidor, y un modo asíncrono, en el que el cliente no espera la respuesta del servidor y puede pedirla mas tarde. El servidor no establece diferencias entre las invocaciones estáticas que utilizan el stub en el lado del cliente y las dinámicas que no lo utilizan. Una invocación dinámica se construye en cuatro etapas: identificar el objeto destinatario de la petición, generarnos nosotros la invocación al método, generarnos las variables que le pasamos a la llamada, así como el valor que nos



devolvería la función, ejecutamos la invocación, y recuperar los datos o las excepciones.

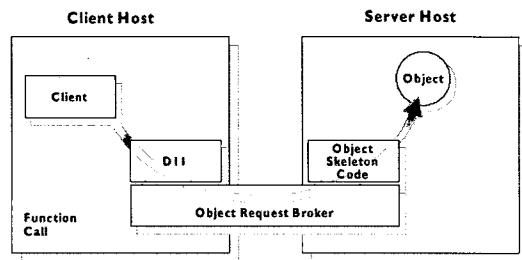


Figura 6: Invocación dinámica-estática.

ORB Interface: permite que las funciones del ORB sean accedidas directamente por el código de la aplicación. En su especificación normal esta interfaz proporciona sólo una pocas funciones, y se implementa tanto por parte del cliente como por parte de la implementación.

Ahora procedemos igual con la parte del servidor, es decir, en la que se encuentra la implementación del objeto:

IDL Skeleton: es la parte equivalente del Stub pero a este lado de la aplicación, es decir, la interfaz del objeto es conocida en el tiempo de compilación por parte del servidor.

Dinamic Skeleton Interface (DSI): equivale al DII del cliente, es decir, permite la programación dinámica. CORBA puede recibir llamadas a funciones de objetos que no existen. El servidor puede examinar la estructura de esas llamadas e implementar un interfaz en tiempo de ejecución. Para utilizar el DSI, el programador debe implementar un procedimiento que reciba todas las peticiones, sean cuales sean los objetos destinatarios. Este procedimiento, la Dinamic Implementation Routine, recibe el nombre del objeto, de la operación y de sus parámetros para cada uno de los mensajes, y es luego responsable de la ejecución de estas operaciones en los objetos correspondientes.

Esta función es imprescindible para la interoperabilidad entre ORB's. Cuando un cliente invoca un objeto presente en un servidor de otro ORB, el DSI transmite la petición al ORB destino, y entonces el puente usa la DII para invocar el objeto destino en el segundo ORB. El DSI puede funcionar con invocaciones del cliente estáticas y dinámicas.

Object Adapter: este interfaz definido por la OMG es sustituido por la mayoría de los ORB, cumple funciones de generar referencias a objetos, registrar implementaciones, autenticación, etc. para el servidor. CORBA define un adaptador estándar llamado Basic Object Adapter (BOA).

ORB Interface: este interfaz es compartido con la parte del cliente.

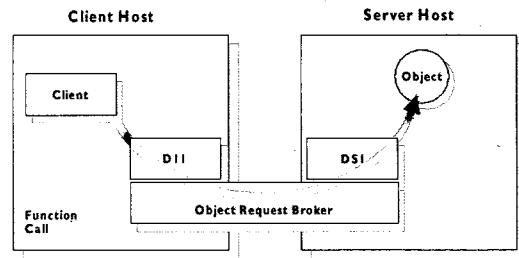


Figura 7. Invocación dinámica-dinámica.

Por tanto disponemos de métodos estáticos y dinámicos tanto por parte del cliente como del servidor, lo que nos da cuatro posibles modos de funcionamiento: estático-estático, estático-dinámico, dinámico-estático y dinámico-dinámico.

2.3 CORBA sobre TCP/IP

En este apartado veremos la implantación de CORBA sobre el protocolo de transporte más utilizado hoy en día, el TCP/IP.

2.3.1 GIOP (General Inter-ORB Protocol)

El protocolo GIOP especifica un estándar de transferencia de datos y un formato de mensaje para la comunicación entre ORB. GIOP es independiente del protocolo de red elegido para la comunicación. Los objetivos de GIOP son los siguientes:

- Simplicidad: esto le asegura su implantación ya que ahorra trabajo a los vendedores de ORB, manteniendo sus costes bajos.
- Escalabilidad: supone soportar todos los ORB y redes de ORB, hasta llegar a toda la Internet.
- Generalidad: el hecho de definir GIOP para cualquier protocolo de transporte orientado a conexión le hace implementable en casi todas las redes. Es particular, su utilización sobre Internet le otorga la mayor disponibilidad posible.

El método que define GIOP codifica los tipos de datos de IDL en una representación de bajo nivel que pueden entender todas las partes. Este método recibe el nombre de Common Data Representation (CDR) y define representaciones para los tipos primitivos (por ejemplo, int, long, double) y los construidos (struct, union, array y string). También permite codificar pseudo-objetos como la información de contexto de una invocación y las excepciones. Las referencias a objetos son un pseudo-objeto muy importante ya que representan a un objeto una vez que sus métodos han sido invocados. Con un solo ORB, estas referencias pueden ser un simple puntero, pero cuando han de ser mandadas a otro ORB se deben codificar como un Interoperable Object Reference (IOR). Esta estructura contiene la referencia a dicho objeto y la información de su

localización, que depende del protocolo de transporte que soporta la conexión.

Para cumplir con su primer objetivo, la simplicidad, GIOP define solamente ocho tipos de mensajes. Tres se originan desde los clientes (sistemas iniciando la conexión), tres desde los servidores (sistemas aceptando la conexión) y otros dos bidireccionales. Todos los mensajes se componen de tres partes: cabecera general, cabecera específica del mensaje y cuerpo del mensaje. A continuación describimos los ocho tipos:

- *Request* (cliente a servidor): es el método básico de invocación. Incluye información sobre el objeto destino, el método a invocar con sus parámetros, si el cliente espera un *Reply*, y un identificador *Request ID*. Éste es generado por el cliente y posibilita a las dos partes a identificar únicamente un *Request*.

- *Reply* (servidor a cliente): es la respuesta al anterior mensaje. Contiene el *Request ID*, el *status* y el cuerpo del *Reply*. Este último incluye los valores de retorno del método invocado si el *status* no indica que se han producido interrupciones. Si ha ocurrido alguna, el cuerpo del mensaje indica cual es. Finalmente, si el *status* es *Location Forward* indica que el objeto requerido ha sido transferido a otro ORB, cuya localización se indica en el cuerpo del mensaje. Entonces el cliente emitirá un nuevo *Request* hacia la nueva localización.

- *LocateRequest* (cliente a servidor): permite al cliente conocer si una referencia a un objeto es conocida, si el servidor en cuestión puede procesar un *Request* hacia ese objeto, o en caso contrario el servidor al que debe dirigirlo. Esta información se puede obtener con un simple *Request*, pero así se evita en caso de *Location Forward* tener que enviar todos los parámetros de una función innecesariamente.

- *LocateReply* (servidor a cliente): es la respuesta al anterior. Contiene el ID correspondiente y un *status* indicando si el objeto es conocido, localizado en el servidor, o localizado en cualquier otra parte, en cuyo caso se incluye su nueva localización.

- *CancelRequest* (cliente a servidor): su único parámetro es un *Request ID*. Indica al servidor que ya no espera más tiempo a un *Reply* de un *Request* o *LocateRequest* pendiente.

- *CloseConnection* (servidor a cliente): notifica al cliente el cierre de una conexión. Todos los *Request* pendientes se pierden y por tanto deben ser reenviados en otra conexión.

- *MessageError* (bidireccional): es enviado en respuesta a un mensaje desconocido o de formato incorrecto.

- *Fragment* (bidireccional): permite a los ORB a fragmentar mensajes y enviarlos en varios separados de tipo GIOP. Si la cabecera de un *Request* o de un *Reply*

indica que le siguen más fragmentos de mensaje, estos serán del tipo *Fragment*.

2.3.2 IIOP (Internet Inter-ORB Protocol)

Este es el GIOP propio para la conexión mediante TCP/IP. Lo que esto significa es bien claro: la posibilidad de tener acceso a objetos presentes en cualquier parte de Internet. Las condiciones que el OMG impone para las capas de transporte que soportan a GIOP parecen estar hechas a medida para TCP/IP: protocolo orientado a conexión, fiable, que pueda enviar paquetes y protección robusta contra errores en caso de fallo en la conexión. Mientras que GIOP define la forma y el contenido de los mensajes, IIOP codifica la información necesaria para la invocación de métodos sobre objetos en sus propios perfiles IOR (Interoperable Object Reference). Éstos están compuestos de un número de versión, el host y el puerto al que dichos mensajes deben ser enviados, una referencia al objeto (*object key*) y una serie de componentes contenido información usada cuando invocamos métodos en el objeto, como el tipo de ORB originario o parámetros de seguridad.

La distinción entre cliente y servidores es muy grande en IIOP, y ninguno debe mandar mensajes propios del otro. En la figura Flujos de mensajes IIOP podemos ver la forma en que se comunican clientes y servidores. Cada una de las flechas representa un socket. En la primera un *Request* de un cliente es respondido por un *Reply*, pero dividido en dos fragmentos. En la segunda vemos que los diferentes *Requests* o *Replies* pueden incluso solaparse. Esto implica un mayor aprovechamiento de los recursos. Algunos ORB permiten elegir entre una conexión multiplexada o una dedicada, para ganar en rapidez. En la tercera secuencia vemos que un cliente puede mandar una serie de *Request* sin esperar respuesta, simplemente para recordarle su presencia a un servidor (*ping*).

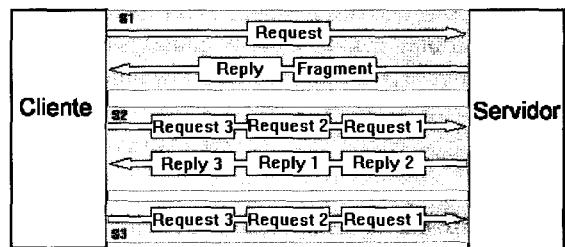


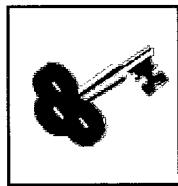
Figura 8. Flujo de mensajes IIOP.

El protocolo IIOP puede ser utilizado también fuera de la arquitectura CORBA. En junio de 1997 JavaSoft anunció que desarrollaría una implementación de RMI para Java, recordemos que es uno de los competidores de CORBA, sobre IIOP. Esto permite usar Java con los métodos de los ORB a través de RMI. Algunos incluso pronostican que puede sustituir al HTTP en algunas funciones.

BIBLIOGRAFÍA Y DIRECCIONES DE INTERÉS:

[1] Object Management Group: www.omg.org





SEGURIDAD A NIVEL DE IP: IPSEC

Pedro Antonio Mur Siles

Estudiante de la ETSETB, UPC y socio colaborador de BJT.

pedro@bjt.upc.es

INTRODUCCIÓN

La utilización de Internet, cada vez más amplia y entre usuarios más diversos, ha provocado la necesidad de proteger todo tipo de información que viaja por la red. Existen diversas propuestas y alternativas para garantizar la seguridad y autenticación de todos los paquetes que vayan por la red. La ampliación del mundo de Internet, junto con los mecanismos de cifrado disponibles, animará, sin duda, al desarrollo de nuevas aplicaciones como comercio electrónico o cualquier actividad desde casa que necesite seguridad como por ejemplo acceso a datos bancarios, etc.

En este artículo se pretende analizar una serie de mecanismos de seguridad que son aplicables a la capa IP y que se han denominado IPsec: la cabecera de autenticación (AH: Authentication Header), el encapsulado de seguridad de carga útil (ESP: Encapsulating Security Payload) así como su combinación, además de los protocolos, la negociación y el intercambio de claves para ambos mecanismos de seguridad (asociaciones de seguridad e intercambio de claves)

También es cierto que existen otros mecanismos de seguridad que se pueden aplicar en otros niveles que no sean el IP, pero no serán motivo del presente documento.

IPSEC

IPsec proporciona servicios de seguridad en la capa IP mediante un sistema para seleccionar los protocolos de seguridad requeridos, determinar los algoritmos usados para los servicios, y determinar unas claves criptográficas necesarias para proporcionar los servicios pedidos. IPsec puede ser utilizado para proteger uno o más "caminos" entre dos hosts, entre dos pasarelas o entre pasarela segura (p.e. encaminador o cortafuegos con IPsec) y host.

El conjunto de servicios de seguridad que puede proporcionar IPsec incluye control de acceso, integridad, autenticación del origen, reenvío de paquetes, confidencialidad (encriptación) y confidencialidad de flujo de tráfico limitado. Debido a que estos servicios

son proporcionados por la capa IP, también pueden ser utilizados por las capas superiores (TCP, UDP, ICMP, BGP, etc.)

IPsec utiliza dos protocolos para proporcionar seguridad: la cabecera de autenticación (AH) y el encapsulado de carga útil (ESP). El primero proporciona integridad, autenticación y un servicio opcional de no-repudio. Mientras que el segundo puede proporcionar confidencialidad (encriptación) y confidencialidad de flujo de tráfico limitado. También puede proporcionar integridad, autenticación y un servicio opcional de no-repudio. Ambos (AH y ESP) son vehículos de control de acceso basados en la distribución de claves criptográficas y la administración de flujos de tráfico relativos a este tipo de protocolos de seguridad. Estos protocolos pueden ser aplicados solos o uno en combinación con el otro y tanto en IPv4 como en IPv6. Ambos protocolos soportan dos modos de uso: el modo transporte y el modo túnel. En modo de transporte tendremos confidencialidad en los datos pero las cabeceras IP estarán al descubierto, es decir que si alguien quiere obtener el flujo de información y así saber con quien nos comunicamos, en cambio el modo túnel encripta la cabecera IP y crea una nueva cabecera con la dirección del encaminador con lo cual sabríamos a que Intranet va la información pero no a que usuario. La elección entre modo transporte y modo túnel depende de si la comunicación es entre hosts o pasarelas.

IPsec permite al usuario (o administrador) controlar el tipo de seguridad ofrecido. Es decir, en cada paquete se puede especificar: qué servicios de seguridad usar y con qué combinaciones, en qué tipo de comunicaciones usar una protección determinada y por último los algoritmos de seguridad utilizados. Otro punto importante es que debido a que estos servicios de seguridad son valores secretos compartidos (claves), IPsec confía en una serie de mecanismos para concretar este tipo de claves (IKE, SA) que se explicarán a continuación.

ASOCIACIONES DE SEGURIDAD (SA)

Para suministrar confidencialidad en las comunicaciones es imprescindible el uso de claves (siempre

y cuando el medio por donde viajan no sea seguro). Estas claves deben ser conocidas tanto por el emisor como por el receptor pero por nadie más. También es conveniente no usar siempre las mismas claves para dificultar la faena de posibles criptoanalistas. Por tanto, nos encontramos con el primer problema: el emisor y el receptor deben decirse que clave usaran antes de empezar la comunicación. Pero no basta con eso, también deben acordar que nivel de seguridad quieren y que algoritmos utilizaran. Del hecho de esta necesidad de acuerdos surge la idea de Asociación de seguridad.

Una SA es una relación entre dos o más usuarios que describe como estos usuarios van a utilizar los servicios de seguridad para comunicarse entre ellos. Toda comunicación que requiera IPSec debe establecer una SA (o varias como veremos más adelante) antes de enviar datos. Todos los datos que van por una SA tienen el mismo tipo de seguridad, los mismos algoritmos y la misma clave de sesión, por tanto los problemas antes mencionados se transforman en la creación de la SA. Existen multitud de protocolos que se encargan de realizar lo anteriormente citado aunque el más extendido es el ISAKMP (Internet Security Association and Key Management) que se comentará posteriormente.

Una asociación de seguridad viene unívocamente identificada por una dirección IP y un índice de parámetro de seguridad (SPI), pero en ella son diversos los parámetros que se pueden definir para darle a la comunicación la seguridad que nos convenga:

- Algoritmo de autenticación y modo de utilización del algoritmo con la cabecera de autenticación de IP (requerido para AH).
- Claves utilizadas con el algoritmo de autenticación en uso con la cabecera de autenticación (requerido para AH).
- Algoritmo de encriptado, modo del algoritmo y transformación que se está utilizando con el encapsulado IP de la carga de seguridad útil (requerido para ESP).
- Claves usadas con el algoritmo de encriptado en uso con el encapsulado de seguridad de la carga útil (requerido para ESP).
- Presencia/ausencia y tamaño de la sincronización de criptografía o inicialización del campo vector para el algoritmo de encriptado (requerido para ESP).
- Claves de autenticación usadas con el algoritmo de autenticación que es parte de la transformación ESP, si hay alguna. (requerido para ESP)
- Tiempo de vida de la clave o tiempo en el que se debería cambiar la clave (recomendado para todas las implementaciones)

- Tiempo de vida de la SA (recomendado para todas las implementaciones)
- Dirección origen de la SA, debería ser una dirección comodín, si existe más de un sistema que envía datos que comparten la misma asociación de seguridad con el destino. (recomendado para todas las implementaciones)
- Nivel de seguridad (por ejemplo, secreto o no clasificado) de los datos protegidos.

ISAKMP (INTERNET SECURITY ASSOCIATION AND KEY MANAGEMENT)

En el ISAKMP diferenciamos dos fases. En la primera se definen los parámetros de seguridad que se van a usar durante la negociación y en la segunda fase se definen los parámetros que se usarán durante la comunicación. Es decir, primero se crea una SA llamada ISAKMP SA que será usada exclusivamente para la negociación y cuyos parámetros vienen definidos en la fase 1. Luego, en la fase 2 se negociará la seguridad posterior encriptando por medio de la ISAKMP SA de modo que nadie sabrá como encriptaremos ni que seguridad vamos a usar dando lugar a la creación de la SA definitiva.

A) Fase 1

Se pueden usar dos modos en la fase 1: el modo principal y el modo agresivo. En el modo principal los dos primeros mensajes negocian los parámetros de seguridad (la del ISAKMP SA), los dos siguientes sirven para intercambiarse los valores públicos de Diffie Hellman, y los dos últimos sirven para autenticarse. En el modo agresivo los dos primeros paquetes negocian la seguridad e intercambian los valores públicos de Diffie Hellman, el tercer paquete sirve para identificar al receptor y el cuarto para autenticar al emisor.

En la fase 1 el punto clave es al autenticación, de nada nos sirve encriptar un mensaje si no estamos seguros de que la persona a la que le llega no es un impostor. Hay diversos métodos de autenticación pero todos ellos se basan en el uso de claves asimétricas, por ello es primordial conocer la clave pública de la persona con quien nos vamos a comunicar. Para evitar que un impostor nos diera su clave pública diciendo que es la de otra persona (con lo cual le enviaríamos información creyéndonos que es otro), existen las Autoridades de certificación (CA). Estas nos garantizan que una clave pública es de un usuario concreto. La figura 1 nos ilustra como se realiza este intercambio.



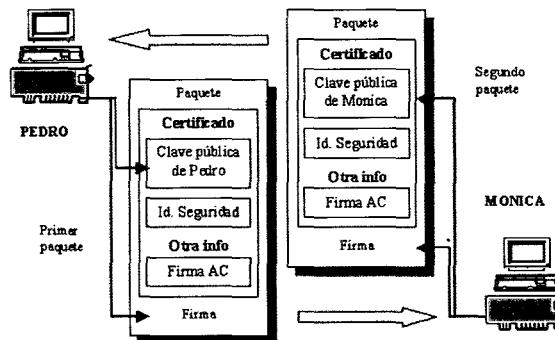


Figura 1: Intercambio de claves públicas

Una vez los usuarios se han autenticado entonces ya pueden generar una clave de sesión mediante los valores públicos requeridos por Diffie Hellman. Esto acontece tal como se ilustra en la figura 2.

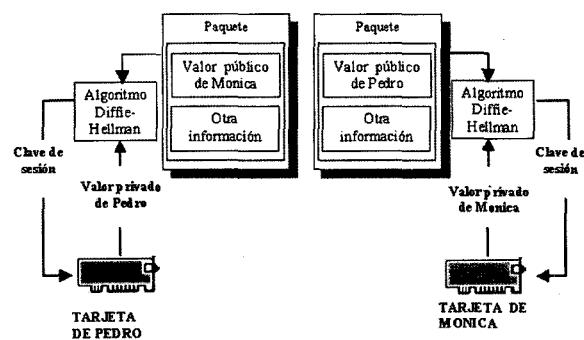


Figura 2: Generación de clave de sesión

B) Fase 2

Ahora que ya disponemos de una comunicación segura (el ISAKMP SA) ya podemos empezar a negociar los parámetros de la SA, esto sucede en la fase 2. En esta fase se usa el “modo rápido” que consiste en que un usuario presenta una serie de alternativas al otro quien o bien elige una de ellas o presenta otra serie de alternativas, y así hasta llegar a un acuerdo. También se debe especificar el SPI (Índice de parámetros de seguridad) que se le asignará a la SA pues será su identificador.

Después de unos campos de control, se propone el tipo de seguridad (AH por ejemplo) y a continuación se presentan diversas alternativas (transforms) al otro usuario, quien debería responder con el mismo mensaje pero con una sola transform, la que haya elegido.

Resumiendo, hay dos fases de negociación. La primera es llevada a cabo por las tarjetas de red de los usuarios o por las pasarelas (cortafuegos) si se diera el caso, y es cuando se decide como proteger el tráfico de la negociación estableciendo un ISAKMP SA. La segunda fase se usa para crear la SA que servirá para crear los parámetros de seguridad para el intercambio posterior de datos; aquí son los usuarios (la aplicación) quienes fijan las características de la SA. El hecho de tener dos fases es ventajoso pues varios SA pueden ser definidos por una misma ISAKMP SA además así proporcionamos confidencialidad a las características de las SA creadas en la fase dos, es decir que nadie sabe con que algoritmo ciframos la información

C) Funcionamiento

Cuando requerimos de una comunicación segura con IPsec (AH, ESP) entonces necesitamos tener una SA. Lo primero que debemos hacer es elegir el tipo de seguridad que queremos y fijar los algoritmos. Una SA puede tener o AH o ESP pero no ambos, si queremos tener ambas protecciones deberemos usar una combinación de SAs llamada “SA bundle”. Otro parámetro que debemos elegir es si queremos modo transporte o modo túnel.

Una vez negociados y aceptados los parámetros, todos los paquetes IP que viajen entre el primer usuario y el segundo (recordar la unidireccionalidad de las SA) irán con dicha seguridad hasta que, o bien acabe la comunicación, o bien finalice el tiempo de vida de la SA. Si ocurre esto último habrá que hacer el proceso desde el principio de nuevo pasando exactamente por los mismos pasos que antes.

D) Algoritmos disponibles

Los algoritmos que podemos usar son varios aunque en la figura 3 se presentan los más utilizados.

Algoritmos para Encriptación	Algoritmos de Hash	Métodos de autenticación
DES-CBC	MD5	Clave pre-compartida
IDEA-CBC	SHA	Firmas DSS
BLOWFISH-CBC	Tiger	Firmas RSA
CAST-CBC		Encriptación con RSA
3DES-CBC		Encriptación revisada con RSA
RC5-R16-B64-CBC		

Figura 3: Algoritmos disponibles actualmente

E) SAD Y SPD

Para elegir entre todas estas posibilidades a la hora de dar seguridad a un mensaje debemos ir a consultar al SPD (Security Policy Database). Allí tene-

mos todas las posibles combinaciones que nos son permitidas. Una vez hayamos escogido una el sistema verificará el SAD (Security Associations Database). Allí se indican todos las SA que hay abiertas. Si hay alguna SA abierta que se corresponde con nuestras necesidades (igual seguridad, algoritmos y destino) nos será asignada y sino se abrirá una nueva SA con tales especificaciones. El resultado es que se nos devolverá un valor que será el SPI (Índice de parámetros de seguridad) de la SA que nos ha sido asignada. A partir de entonces en todos los paquetes IP destinados a esta comunicación deberemos ponerles el SPI y también decir si se trata de una seguridad AH o ESP. Esto debemos indicarlo porque en realidad hay dos SAD, uno para las comunicaciones con AH y otro para las que usan ESP, es también por esta razón que AH y ESP no pueden estar combinados en una única SA

A partir de aquí la gestión del flujo de información es muy sencilla. Para el flujo saliente en cada paquete IP se siguen los siguientes pasos:

- Se comprueba si va a usar AH o ESP o ambos, y se mira el SPI.
- Se busca en la adecuada SAD (la de AH o la de ESP) la SA que corresponda con el SPI
- Una vez encontrado se codifica la información tal como indique la SA y se envía el paquete.

Para el flujo entrante el método es muy similar:

- Se mira la dirección IP destino del paquete y si no coincide con la nuestra se descarta
- Se comprueba si el paquete está codificado con AH o ESP y se anota el SPI
- Se busca en la adecuada SAD (la de AH o la de ESP) la SA que corresponda con el SPI.
- Se decodifica el paquete según se indica en la SA.

Si usamos modo túnel los 2 primeros pasos son los mismos que los anteriores pero además la pasarela (que es quien se encarga de la gestión de la seguridad en este caso) deberá, para flujo saliente codificar la cabecera IP (además de la parte de datos) y construir una nueva cabecera IP con la dirección de la pasarela destino (no con la del host) y enviar el paquete. Y para el flujo entrante decodificar el paquete de datos y la cabecera IP encriptada y enviar el paquete al destinatario final.

En el caso de que usáramos AH más ESP tendríamos un SA bundle que no es mas que la concatenación de dos SA. En este caso el paquete antes de ser enviado pasaría por dos SAs una de ESP y otra de AH, y a la hora de recibirla sucedería lo mismo.

AH (Authentication header)

Como se ha dicho anteriormente, proporciona integridad, autenticación y protección anti-repudio. Este último parámetro es opcional y puede ser escogido cuando se establece la asociación de seguridad. AH proporciona autenticación a tanta información como sea posible de la cabecera, además de los datos de capas superiores. Sin embargo, algunos campos de la cabecera IP pueden cambiar en tránsito y el valor de estos campos, cuando llegan a destino no pueden ser predichos por el que los envía. Estos valores no pueden

0	1	2	3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1	RESERVADO		
Siguiente cabecera	Longitud carga útil		
Índice de parámetros de seguridad			
Campo de número de secuencia			
Datos de autenticación (variable)			

Figura 4: Formato del paquete AH

ser protegidos por la AH. Se puede utilizar en modo transporte o en modo túnel.

Todos los campos descritos en la figura 4 son obligatorios, es decir, que siempre están presentes en el formato AH y están incluidos en el cálculo del valor de comprobación de integridad (ICV). La cabecera AH tiene que ser múltiple de 64 bits.

El índice de parámetros de seguridad es un número arbitrario de 32 bits que, en combinación con la dirección IP destino y el protocolo AH, únicamente identifica la asociación de seguridad para este paquete. En el caso de que el SPI tome el valor 0 significará que no existe ninguna SA. El número de secuencia contiene un número creciente de contador. Es obligatorio y siempre está presente incluso si el receptor no elige habilitar el servicio anti-repudio. Los contadores de emisor y receptor se inicializan a 0 cuando se establece la SA. El emisor lo incrementa para esa SA e inserta el nuevo valor en este campo. Por tanto, el primer paquete enviado usando una determinado SA tiene como número de secuencia el 1. Si está activado el anti-repudio (por defecto), el emisor comprueba para asegurarse que el contador no ha pasado un ciclo antes de insertar el nuevo valor en el campo. En otras palabras, el emisor no enviará otro paquete en la SA si haciéndolo causa que el número de secuencia pase un ciclo.

El ICV está dentro del los datos de autenticación. El algoritmo de autenticación empleado para el cálculo del ICV se especifica en la asociación de seguridad establecida anteriormente. En comunica-



ción punto a punto, algoritmos apropiados son los denominados MACs (Keyed Message Authentication Codes) basados en algoritmos de encriptación simétrica (p.e. DES) o en funciones hash (p.e. MD5 o SHA-1). Para comunicaciones multicast son apropiados los algoritmos de hash combinados con algoritmos de firma asimétrica. De todas formas se pueden utilizar otros algoritmos.

Para finalizar, el ICV del AH se calcula sobre los campos de la cabecera IP que son o constantes en tránsito o son un valor predecible en destino, la cabecera AH (todos los campos, aunque los datos de autenticación, donde se encuentra este valor, se suponen 0) y los datos de niveles superiores, que se asumen como constantes durante el trayecto.

Una vez recibido el paquete, el receptor calcula el ICV sobre los campos apropiados del paquete, usando el algoritmo de autenticación especificado y verifica que es el mismo que el ICV incluido en el campo de datos de autenticación. Si coinciden, entonces el paquete es válido y se acepta. Si el test falla, el receptor debe descartar el paquete recibido.

En las figuras 5 y 6 se pueden observar los paquetes IPv6 en modo transporte y túnel después de aplicar AH.

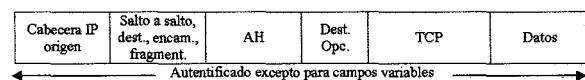


Figura 5: Paquete IPv6 después de aplicar AH en modo transporte

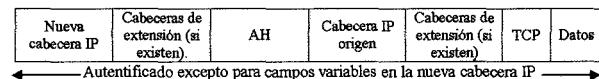


Figura 6: Paquete IPv6 después de aplicar AH en modo túnel

ESP (ENCAPSULATING SECURITY PAYLOAD)

Los servicios de seguridad que nos puede ofrecer ESP son confidencialidad, autenticación, anti-repudio, integridad y confidencialidad parcial en el flujo de tráfico. Hay que tener en cuenta que: la confidencialidad y/o autenticación deben estar activos; no puede haber anti-repudio ni integridad sin autenticación y que para que haya confidencialidad en el flujo de tráfico debe seleccionarse el modo túnel.

El formato del paquete ESP variará según la seguridad elegida así como también variara su localización dentro del paquete IP conforme al modo selec-

cionado (modo transporte o túnel). Pero en general el formato del paquete ESP se puede definir como la figura 7.

En el campo de datos de carga útil es donde se encuentra la información propiamente dicha. Consiste

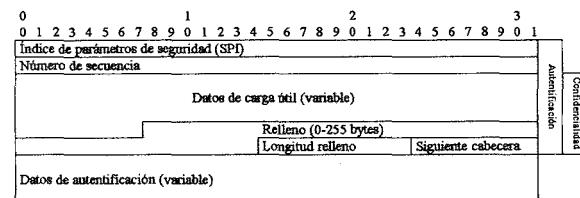


Figura 7: Formato del paquete ESP

en un número indefinido de bytes (mucho mayor que el mostrado en la figura 7) que contienen la información encriptada que queremos transmitir. Muchos algoritmos de encriptación necesitan un vector de inicialización que les sirve de sincronización. Cuando este vector es requerido va incluido en los datos de carga útil. En este caso es importante que el algoritmo especifique la exacta localización y el tipo de estructura del vector en cuestión. En el caso de que hayamos escogido autenticación en este campo tendremos un ICV, muy similar al que utiliza el AH que nos servirá para cerciorarnos de si alguien ha modificado el paquete después de haber sido enviado. Hay que darse cuenta que esta autenticación es sobre la información encriptada es decir que para comprobar que la información que nos llega es auténtica deberemos codificarla, lo cual puede no sernos útil dependiendo del sistema que se utilice.

En las figuras 8 y 9 se pueden observar los paquetes IPv6 en modo transporte y túnel después de aplicar ESP.

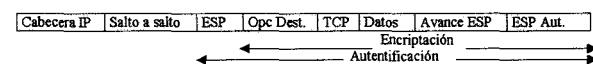


Figura 8: Paquete IPv6 después de aplicar ESP en modo transporte

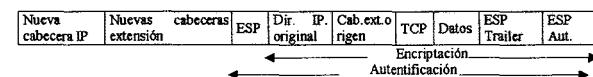


Figura 9: Paquete IPv6 después de aplicar ESP en modo túnel

COMBINACIÓN: AUTENTIFICACIÓN MÁS PRIVACIDAD

Los dos mecanismos de seguridad de IP se pueden combinar para transmitir un paquete IP que tenga autenticación y privacidad. Existen dos técnicas que

se puedan utilizar, diferenciadas por el orden en que se apliquen los dos servicios.

A) Encriptado antes de autentificación

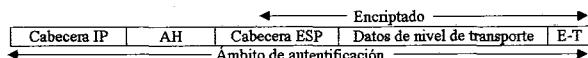


Figura 10: Encriptado antes de autentificación (modo transporte o túnel) en IPv6

En este caso, el paquete IP entero transmitido se autentifica, incluyendo ambas partes, la encriptada y la no encriptada. Primero se aplica ESP a los datos que se van a proteger, después incorpora al principio la cabecera de autentificación y la(s) cabecera(s) IP en texto nativo. De hecho, existen dos casos:

- ESP en modo transporte. La autentificación se aplica al paquete IP entero pero sólo el segmento de la capa de transporte se protege por el mecanismo de privacidad.
- ESP en modo túnel. La autentificación se aplica al paquete IP entero entregado a la dirección IP destino externa y la autentificación se lleva a cabo en el destino. El paquete IP interno se protege por el mecanismo de privacidad para su entrega al destino IP interno.

B) Autentificación antes del encriptado

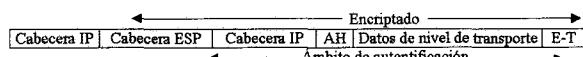


Figura 11: Autentificación antes de encriptado (modo túnel) en IPv6

Esta técnica sólo es apropiada para ESP en modo túnel. En este caso la cabecera de autentificación se sitúa dentro del paquete IP interno. Este paquete interno es autenticado y protegido por el mecanismo de privacidad.

Cabe destacar que este método puede ser preferible por las siguientes razones. Primero, ya que el AH se protege por el ESP, es imposible que cualquiera intercepte el mensaje y altere el AH sin ser detectado. Segundo, puede ser deseable almacenar la información de autentificación con el mensaje y el destino para una referencia posterior. Es más conveniente hacer esto si la información de autentificación se aplica a un mensaje no encriptado; de la otra forma, el mensaje tendría que ser reencriptado para verificar la información de autentificación.

CONCLUSIÓN

Las cabeceras AH y ESP están definidas tanto en IPv4 como IPv6. En el caso de IPv4 las nuevas cabeceras son añadidas al paquete como opciones adicionales. En el caso de IPv6, el protocolo ya está diseñado para incorporar estas cabeceras y se disponen en el orden óptimo para no entorpecer el tráfico de manera ostensible.

Los mecanismos de seguridad que se han citado son, en principio, suficientes para asegurar que toda información que viaje por la red vía protocolo IP será segura, fiable y autenticada. Estamos en el comienzo de una era donde Internet va a jugar un papel muy importante en la sociedad y por consiguiente la seguridad que pueda tener va a ser uno de los principales condicionantes ya que la mayoría de aplicaciones que se van a poder llevar a cabo necesitarán llenar datos personales, algunos de ellos muy importantes o hacer transacciones bancarias de gran importancia y saber que existe un protocolo que va a proteger este tipo de transmisión de información va a empujar a desarrollar nuevas aplicaciones e impulsar las ahora poco existentes que necesitan una seguridad y fiabilidad del 100%.

IPSec está todavía en fase de desarrollo y por tanto los mecanismos de seguridad pueden ser modificados. El futuro de dichos mecanismos de seguridad es bastante incierto, cabiendo la posibilidad de que nunca sean realmente utilizados ya que se tiende a ofrecer seguridad a nivel de aplicación, donde la seguridad se ofrece de modo transparente al usuario siendo esta de punto a punto.

BIBLIOGRAFÍA

- [1] William Stallings. 1995. Network and Internetwork Security, principles and practice. Ed. Prentice Hall.
- [2] Cisco systems: <http://www.cisco.com>
- [3] Softpro: <http://www.softpro.com/softpro>
- [4] SSH: <http://www.ipsec.com>
- [5] <http://www.cs.ucl.ac.uk/staff/J.Crowcroft/mmbook/book/node345.html>
- [6] <http://www.web.mit.edu/network/isakmp>
- [7] <http://www.whatis.com/IPSec.htm>
- [8] Kent, S., and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, Noviembre 1998
- [9] Kent, S., and R. Atkinson, "IP Authentication Header", RFC 2402, Noviembre 1998.
- [10] Kent, S., and R. Atkinson, "IP Encapsulating Protocol", RFC 2406, Noviembre 1998.
- [11] D. Piper, "The Internet IP Security Domain of Interpretation for ISAKMP", RFC 2407, D.





COMUNICACIÓN EN TIEMPO REAL SOBRE INTERNET

Felipe Moreno Strauch

*Estudiante de la ETSETB, UPC y socio colaborador de BJT.
felipe@bjt.upc.es*

INTRODUCCIÓN

En los primeros 20 años de su existencia, Internet era básicamente utilizada para el intercambio de mensajes de correo electrónico y para la transferencia de ficheros y casi exclusivamente por personal técnico o de investigación de las universidades, de instituciones del gobierno o laboratorios de investigación de la Industria. Pero en los últimos años, con el crecimiento exponencial, la aparición de nuevos servicios y una transición hacia una red comercial la situación ha cambiado radicalmente.

El significativo aumento del ancho de banda y de la capacidad de proceso de los ordenadores son dos factores que, juntamente con la evolución de las tecnologías de acceso nos ha permitido avanzar hacia el concepto de Integración de Servicios. Desarrollar una sola red, capaz de soportar todos los servicios que actualmente viajan por distintas redes de transporte: la difusión de televisión y radio, la telefonía, la transmisión de ficheros, etc. Pero no solo esto, sino además, asegurar para cada uno de ellos, la calidad de servicio requerida, lo que supone ser capaz de tratar los distintos tipos de tráfico generado de forma óptima.

Empiezan a aparecer nuevos servicios que funcionarán en Internet y que van más allá que la simple distribución de páginas web o envío de mensajes de correo electrónico. Nuevos servicios como la difusión de vídeo y audio (tanto desde ficheros como retransmisión en directo de eventos), Radios y Televisión vía web, vídeo bajo demanda, telefonía, videoconferencia, presentaciones multimedia, simulaciones en tiempo real, etc.

Aunque por ahora algunos aún no se pueden implementar en la práctica debido a la escasez de ancho de banda, otros ya empiezan a ponerse en marcha y los resultados son muy interesantes. La industria se ha interesado: Microsoft ha creado una nueva división para desarrollar aplicaciones multimedia (Windows Media Player) y ha entrado de lleno en la batalla con Real Networks (RealAudio y RealVideo) por convertirse en el standard para la transmisión multimedia en tiempo real. Las grandes empresas del sector de las comunicaciones como la NBC o CNN ven Internet como un mercado potencial muy interesante y ya han empezado a explorarlo (en 1997 60.000 usuarios siguieron en directo por Internet un capítulo de la serie ER). La posibilidad de incluir publici-

dad es un atractivo añadido ya que se puede enviar los anuncios en función del perfil de cada usuario.

Pero para que todo ello pueda realmente evolucionar es necesario estudiar la transmisión de información en tiempo real sobre una red de conmutación de paquetes como Internet, preparada para transmitir datos.

Hay que buscar una forma de adaptar este nuevo tráfico, que tiene unas características muy singulares a la infraestructura disponible, estudiando los posibles problemas y desarrollando nuevos protocolos.

TRANSMISIÓN EN TIEMPO REAL SOBRE REDES DE CONMUTACIÓN DE PAQUETES

La comunicación en tiempo real tiene como característica importante el hecho de que el valor de la comunicación depende del momento en que los mensajes llegan al destino. Al contrario de lo que ocurre con la transmisión de datos a las que estamos acostumbrados (por ejemplo una transmisión de fichero), los paquetes que integran una transmisión de vídeo o audio deben llegar al destino en el momento adecuado, o como mucho dentro de un cierto margen ya que el sistema los necesita en aquel momento para reproducir la señal. Existe una cuota para el retardo con que un mensaje se ha de entregar correctamente al destino. Este retardo máximo aceptable fija una especie de "fecha de caducidad" para la información o tiempo límite para la llegada del mensaje. Si un mensaje llega una vez ya ha caducado, es decir sobrepasa el tiempo límite, el valor de la información que contiene disminuye y muchas veces acaba siendo completamente inútil (como una muestra de señal de audio que llegue con un retraso de 2 segundos) y acaban por ser descartados y se consideran, desde el punto de vista de la conexión, como paquetes perdidos.

En función de la tolerancia a que los paquetes lleguen con un retardo mayor que el aceptable se puede clasificar la comunicación en tiempo real como "hard real-time" o "soft real-time". En el primer caso el servicio no es capaz de tolerar perdidas (por ejemplo un sistema de control remoto, cuando ha de reaccionar frente a una emergencia), mientras que en el segundo es aceptable una cierta tasa de perdida (por ejemplo un sistema de transmisión de audio).

El retardo acumulado desde un extremo a otro de la comunicación se llama latencia. A la latencia contribuyen el origen, la red y el destino. El origen contribuye con el tiempo que transcurre desde que muestrea la señal hasta que envía las muestras y el destino con el tiempo que tarda antes de analizarlo. Estos retardos en general no son significativos frente a los que impone la red. La red contribuye de diferentes formas al retardo total:

- Retardo de propagación: es el tiempo que tarda la información en viajar desde un extremo a otro sobre el medio de transmisión utilizado. En entornos reducidos este retardo es siempre despreciable, pero al hablar de sistemas que funcionen sobre Internet, a escala global se ha de tener en cuenta. Por ejemplo, considerando la velocidad de la luz como velocidad de propagación una señal tardaría unos 134 ms en dar la vuelta a la Tierra sobre el ecuador. Teniendo en cuenta que las restricciones impuestas rondan las centenas de ms este factor puede ser importante.
- Retardo de transmisión: es el tiempo en que el origen tarda en poner el paquete en el medio de transmisión. Viene determinado por la velocidad de transmisión y por el tamaño del paquete.
- Retardo "store-and-forward": es el tiempo que se pierde debido a que los routers intermedios deben recibir el paquete completamente antes de retransmitirlo. Depende del número de nodos que atraviesa el paquete.
- Retardo de proceso: debido a que los routers deben analizar la cabecera y decidir la ruta que debe seguir. Además es necesario cambiar algunos campos y recalcular cálculos de checksum (por ejemplo al cambiar el campo time-to-live de un paquete IP).

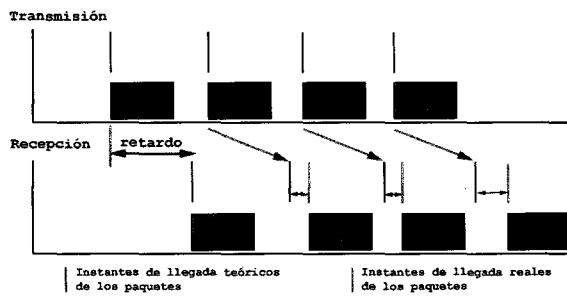


Figura 1. Efecto de la transmisión de un flujo constante de información sobre una red de conmutación de paquetes.

Otro parámetro que hemos de tener en cuenta al analizar los servicios que requieren tiempo real es el jitter, que se podría definir como la máxima variación del retardo extremo a extremo que sufren los paquetes de una misma conexión. Como acabamos de ver, muchos facto-

res contribuyen al retardo extremo a extremo, básicamente retardos introducidos por la red. Lógicamente estos retardos no son deterministas y dependen del estado actual de la red lo que hace que los distintos paquetes lleguen al destino con diferentes retardos o lo que es lo mismo, que el tiempo entre la llegada de paquetes consecutivos será una variable aleatoria. Esto causará una "perdida de sincronismo" en el receptor que debe ser corregida para que sea posible reconstruir el flujo de datos original.

Este efecto es especialmente malo, por ejemplo, en transmisiones de audio ya que produce una serie de chasquidos que resultan muy molestos en la comunicación. La solución que permite corregirlo es sencilla y se basa en la utilización de un buffer en recepción (con política FIFO) que se suele denominar "playout buffer". En lugar de procesar los paquetes a medida en que llegan, el receptor los almacena en el buffer y cuando este se ha llenado los empieza a procesar. Como en media la tasa de paquetes que entra es igual a la que sale del buffer el sistema es estable.

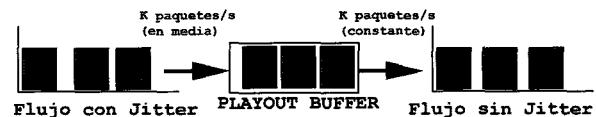


Figura 2. Utilización del Playout Buffer para corregir el

La utilización del playout buffer reduce el efecto del jitter de la misma forma que el sistema de amortiguación de un coche reduce el efecto de las irregularidades de la carretera. Las variaciones que son menores que el tamaño total del buffer no son observables a la salida. El cálculo del tamaño que debe tener el buffer se ha de hacer basándose en una estimación de la estadística del retardo que introduce la red y la tasa de perdidas que se acepta como tolerable y se hace mediante el percentil.

Hay que observar sin embargo que para compensar el jitter de esta forma estamos aumentando la latencia ya que ahora los paquetes han de estar almacenados en el buffer durante un tiempo antes de ser procesados. Aunque en general siempre será necesario controlar el jitter, la latencia sólo es importante en aquellos servicios que requieren interacción entre las partes implicadas. Por ejemplo para telefonía la tolerancia en el retardo total en ir y volver (round-trip delay) es de 400 ms (un retardo mayor degradaría la calidad por encima de lo aceptable) mientras que en el caso de la difusión de vídeo podríamos hablar de 500 ms o más ya que la información circula solo en un sentido. El problema en las aplicaciones que requieren interactividad es que el tiempo entre que el usuario



actúa y percibe el resultado de su actuación es un factor que influye mucho en la calidad percibida. En cambio en un sistema sin interactividad, el hecho de que haya una diferencia entre el momento en que se envía una muestra de señal y el receptor la recibe no es detectable por el usuario.

Solucionar el problema de la latencia no es tan sencillo ya que hemos de buscar mecanismos que permitan reducir el retardo que sufre un paquete en la red. Alternativas como la asignación de prioridades o la reserva de recursos podrían contribuir a ello. Desde el punto de vista del procesado de señal el desarrollo de mejores sistemas de codificación y compresión de la información (reduciendo la tasa en bits por segundo que es necesario enviar) también contribuiría.

Lo primero que hemos de hacer basándonos en estos requerimientos es verificar si las herramientas que disponemos actualmente son suficientes o es necesario algo más.

PORQUÉ TCP O UDP NO SON SUFICIENTES

Aunque podría ser posible utilizar el TCP para transportar los datos de una transmisión en tiempo real hay una serie de motivos que no aconsejan su utilización:

TCP es un protocolo que realiza comprobación de errores y retransmisión de paquetes. El sistema de retransmisión de TCP se basa en esperar durante un cierto tiempo (el time-out) la llegada de una confirmación de que el destinatario ha recibido correctamente los datos. Si esta confirmación no llega dentro del tiempo se retransmite el paquete. El problema es que cuando el emisor se da cuenta de que debe retransmitir suele ser demasiado tarde, la información ya ha perdido su valor. Para los requerimientos de tiempo real un sistema de retransmisiones resulta inútil dadas las fuertes restricciones temporales.

TCP utiliza un control de congestión que decrementa el tamaño de la ventana de transmisión cuando hay perdidas de paquetes para evitar sobrecargar a la red. Sin embargo en transmisiones en tiempo real hay una tasa de transmisión (la que genera la fuente) que debe llegar siempre al destino y no se puede recortar.

TCP no dispone de un sistema de distribución multicast lo que es una seria restricción en los servicios multipunto a gran escala: enviar un paquete por cada destinatario representaría una utilización ineficaz del ancho de banda.

Como alternativa podríamos pensar en UDP ya que este protocolo no utiliza ningún sistema de control de errores y permite utilizar IP multicast pero tampoco

resulta adecuado ya que es demasiado sencillo y no contiene toda la información necesaria: momento de generación de los datos (necesario para reordenar muestras y recuperar el sincronismo con otros flujos de datos), información sobre la codificación utilizada, etc.

Como hemos visto, ni TCP ni UDP resultan ser útiles para el transporte de datos con requerimientos de tiempo real. Si la infraestructura de transporte disponible es insuficiente hemos de desarrollar nuevas herramientas. Necesitamos un nuevo protocolo que complete la funcionalidad de UDP.

REAL TIME TRANSPORT PROTOCOL (RTP)

La IETF (Internet Engineering Task Force) ha desarrollado un nuevo protocolo especialmente pensado para el transporte de este tipo de información, el Real Time Transport Protocol (RTP). Este nuevo protocolo incluye una serie de funciones que facilitan esta tarea: identificación del tipo de información (tipo de codificación), números de secuencia, sincronismo (timestamp) y monitorización de la calidad de servicio. Sin embargo, RTP no posee ninguna función que garantice la entrega de los paquetes dentro del periodo adecuado ni ningún otro tipo de garantía de la calidad de servicio. Conviene resaltar también que aunque sea un protocolo de "transporte", porque regula el envío de datos de un extremo a otro, este protocolo funciona realmente en el nivel de aplicación.

RTP prácticamente no hace ninguna suposición sobre el nivel inferior sobre el que funciona excepto que el protocolo en cuestión defina el tamaño de los paquetes ("framing") ya que el RTP no incluye en su cabecera ningún campo de longitud de los datos. No se asume la existencia de ningún tipo de control de errores, ni se asume la existencia de una conexión, ni ningún mecanismo de reordenación de paquetes. En general, en una arquitectura TCP/IP el RTP se utiliza sobre UDP, aunque el protocolo es lo suficientemente general como para utilizarse en otros tipos de red, como ATM por ejemplo.

El protocolo RTP consta de dos partes muy relacionadas: la parte que transportan los datos (a la que en general se refiere como RTP) y la parte de control, referida como RTCP (Real Time Transport Control Protocol) que realiza las funciones de monitorización de la calidad de servicio y control de los participantes de una determinada sesión. El control realizado no es estricto, es decir, no se verifica si los usuarios tienen o no derecho a participar. Esta función podría ser realizada por algún otro protocolo de control de sesión como SIP (Session Initiation Protocol) por ejemplo. Cuando RTP se utiliza sobre UDP se necesitan dos puertos: uno para la transmisión de información útil y otro para la transmisión de información de control.

FORMATO DE LA CABECERA DE RTP: TRANSPORTE DE LA INFORMACIÓN ÚTIL

El paquete RTP que transporta la información incluye la cabecera básica de RTP que consta de 96 bits (12 bytes), una lista de identificadores de fuentes (CSRC, solo utilizada cuando se agregan varios flujos de datos en uno solo) y los datos. Según el tipo de datos se puede añadir una cabecera adicional con mas información, cuyo formato esta especificado en documentos diferentes para cada tipo y publicados como RFCs.

Los campos más importantes en la cabecera son:

- Payload Type (7 bits): indica el tipo de datos que lleva el paquete. Es lo que servirá a la aplicación que recibe los datos para saber como debe interpretarlos (que CODEC utilizar).
- Timestamp (32 bits): representa el instante de generación del primer octeto del campo de datos. Este instante se deriva de un reloj (reloj en el sentido de contador) que se incrementa monotónicamente y de forma lineal con el tiempo para permitir la sincronización y la estimación del jitter. La resolución de este reloj debe ser suficiente para la precisión que se quiera conseguir en la sincronización y en el calculo del jitter. Su frecuencia depende del tipo de datos transportado y se indica en la especificación del tipo en cuestión. Por ejemplo para la transmisión de audio, el reloj de timestamp se incrementa en uno por cada muestra de señal. Si se envían 160 muestras por paquete el timestamp se incrementaría en 160. El valor inicial se elige aleatoriamente. Es posible que paquetes consecutivos tengan el mismo timestamp si fueron generados en el mismo instante de tiempo (por ejemplo información de un mismo frame en un transmisión de video). También es posible que paquetes consecutivos lleven timestamps que no sean monotónicos ya que es posible que, según el formato de codificación utilizado, la información no se transmita en el mismo orden en que fue muestreada.
- Sequence Number (16 bits): indica el numero de secuencia del paquete dentro del flujo de datos. El valor inicial es aleatorio y se incrementa en uno para cada paquete enviado. Se utiliza para detectar perdidas de paquetes y reordenar paquetes con el mismo timestamp (no se utilizan para ordenación de las muestras por lo comentado en el punto anterior sobre el envío en un orden diferente al de muestreo).
- SSRC (32 bits): especifica un identificador para la fuente del flujo de datos. Todos los paquetes identificados con el mismo SSRC forman parte de un mismo espacio de números de secuencia y de timestamp. El receptor agrupa los paquetes por el SSRC cuanto

recibe información de mas de una fuente diferente a la vez. El valor de este campo, que ha de ser único entre todos los participantes de una misma sesión, se elige aleatoriamente. Si un mismo host genera múltiples flujos de datos en la misma sesión (una sesión multimedia por ejemplo, con audio y vídeo) necesita un identificador diferente para cada uno.

EL PROTOCOLO DE CONTROL: RTCP

El protocolo de control RTCP consiste en una serie de paquetes, cada uno con un formato y función especificado que se envían periódicamente los participantes en una sesión para transmitir información sobre la calidad de la recepción. Si la sesión es multipunto (funciona sobre IP multicast) los paquetes se envían a la misma dirección multicast utilizada para enviar los paquetes de datos. De esta forma el envío de los paquetes RTCP sirve además como indicador de actividad para los participantes aunque estos sean miembros pasivos (solo reciben y no emiten).

Las cuatro funciones básicas que realiza la parte de control son:

- Monitorizar la calidad de servicio: la principal función del RTCP es informar al emisor sobre la calidad de la información recibida. Como esta información se envía a todos los participantes es posible detectar si los problemas generados son locales o globales.
- Relacionar cada fuente de datos con un identificador persistente (ya que el SSRC es diferente para cada flujo de datos) llamado CNAME (Canonical Name), que se utiliza para identificar el host durante una sesión y sirve para relacionar distintos flujos de datos generados por la misma fuente. El CNAME utiliza un formato del tipo: usuario@host.
- Permitir que todos sepan el numero total de participantes en una sesión. Esto es importante ya que para que el sistema no saturé la red con paquetes de control cuando el numero de participantes crece, la tasa de generación de paquetes de control debe disminuir a medida que aumente el numero de usuarios para que entre todos ocupen siempre un ancho de banda fijo. Para que cada uno pueda calcular la tasa con que debe emitir es necesario estimar el numero total de participantes.
- Distribuir información necesaria para que se puedan sincronizar los distintos flujos de datos en una sesión multimedia (por ejemplo sincronizar el audio y vídeo). Cada fuente indica la relación entre su timestamp y su reloj global (wallclock), que representa el tiempo real, para cada flujo de datos que envía. Si el sincronismo de flujos se debe lograr entre fuentes situadas en maquinas diferentes los relojes de ambas máquinas deben estar sincronizados.



En la figura 3 se puede ver un esquema que resume los identificadores utilizados por un host dentro de una sesión multimedia: el CNAME que identifica el host en la sesión, los identificadores de fuente SSRC que identifican cada flujo de datos generado y los puertos RTP y RTCP para cada flujo de datos.

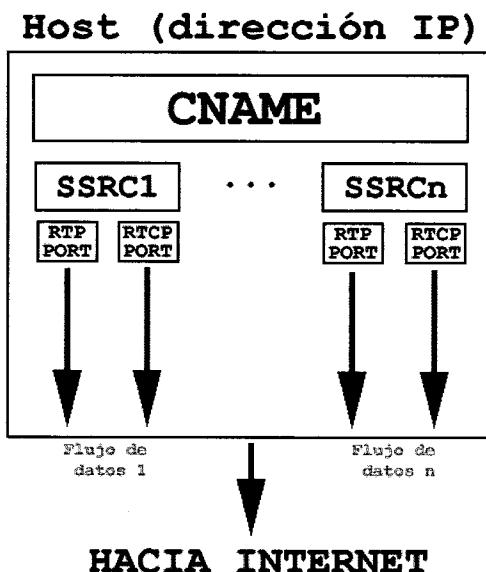


Figura 3. Esquema de identificadores de una fuente con varios flujos de datos en una sesión multimedia.

La información sobre la calidad de la recepción que se envía indica, entre otros parámetros:

- La fracción de paquetes de datos perdidos sobre el total de paquetes enviados desde el último informe.
- El número total de paquetes perdidos desde el inicio de la comunicación.
- El número de secuencia del ultimo paquete recibido correctamente.
- Una estimación del jitter, medido en las unidades del timestamp.

Esta información es enviada por todos los participantes que han recibido datos recientemente y se envía a todos los participantes (hacia la dirección del otro componente en el caso punto a punto o hacia la dirección multicast en el caso de multipunto). Se genera un informe por cada una de las fuentes de las que se ha recibido datos. En función de estos parámetros enviados el emisor puede detectar si existen problemas, si estos son locales o globales y actuar en consecuencia, por ejemplo cambiando el tipo de codificador utilizado para disminuir la tasa de información que envía a cambio de perder calidad.

CONCLUSIONES

Aunque tengamos especificadas algunas herramientas para tratar el tráfico con características de tiempo real y ya existan sistemas que funcionen actualmente en Internet, todavía nos quedan problemas que solucionar. Problemas como la garantía de la calidad de servicio (se está desarrollando una solución basada en el protocolo RSVP – Resource Reservation Protocol – que permitirá reservar ancho de banda en los routers) o como integrar el contenido Multimedia en las páginas web (se está desarrollando un nuevo lenguaje basado en XML llamado SMIL – en desarrollo por el W3C – que permite definir distintos tipos de datos y una relación temporal entre ellos).

Los sistemas que funcionan actualmente en Internet son básicamente dos (se indican entre paréntesis el nombre del software cliente y el servidor): el RealSystem G2 (RealPlayer + RealServer) de Real Networks, que prácticamente fue un standard de-facto desde la creación de RealAudio en 1995 y el Windows Media (Windows Media Player + Windows Media Services) de Microsoft, creado hace tan solo un año y medio pero que ya ha empezado a ganar terreno.

La ultima versión de RealSystem utiliza el RTP como protocolo de transporte de datos y el RTSP como protocolo de control, en sustitución al RTCP. RTSP o Real Time Streaming Protocol ofrece mas opciones a la hora de monitorizar la calidad de servicio.

BIBLIOGRAFÍA

- [1] Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V., "RTP: A Transport Protocol for Real-Time Applications", RFC 1889, January 1996.
- [2] Schulzrinne, H., "Internet Services: from Electronic Mail to Real-Time Multimedia", KIVS'95, February 1995.
- [3] Aras, Ç., Kurose, J., Reeves, D., Schulzrinne, H., "Real-Time Communication in Packet-Switched Networks".
- [4] IETF Audio/Video Transport (avt) charter, <http://www.ietf.org/html.charters/avt-charter.html>
- [5] RTP: About RTP and the Audio-Video Transport Working Group, <http://www.cs.columbia.edu/~hgs/rtp/>
- [6] Windows Media Technology en [microsoft.com](http://www.microsoft.com/windows/windowsmedia/), <http://www.microsoft.com/windows/windowsmedia/>
- [7] RealNetworks Documentation Library, <http://service.real.com/help/library/>



COMPUTER TELEPHONY INTEGRATION

David Roldán Martínez

Ingeniero de Telecomunicación

droldan@tissat.es

INTRODUCCIÓN

La utilización de las nuevas tecnologías da a las empresas modernas oportunidades que no pueden ignorarse. El sistema telefónico de la empresa siempre ha sido considerado como la herramienta principal de comunicación y esta es la razón del por qué la inversión en PaBXs ha sido siempre una obligación. Pero, ¿se requieren las mismas funciones en un sistema de comunicaciones de empresa hoy en día que hace 10 años? Por supuesto que no. Las comunicaciones de empresa ya no están limitadas únicamente al ámbito telefónico, sino que es necesario aprender cómo aprovecharse de toda la potencialidad ofrecida por los nuevos canales de comunicación (fax, E-mail, Web, SMS, Teletext, Wap,...). El nuevo entorno en el que las empresas se mueven se dirige a la diversificación de los canales de comunicación con el cliente y la gestión eficiente de dichos canales se convierte en un objetivo principal.

Entre los beneficios derivados de la adopción de un sistema inteligente para gestionar toda la información sobre los contactos de la empresa, se encuentran los siguientes:

- Gestión simple.
- Aumenta la disponibilidad de la empresa hacia el público.
- Aumenta la productividad y el nivel de satisfacción del personal.
- Disponibilidad de datos y estadísticas sobre la calidad servicio que se está ofreciendo en tiempo real.
- Posibilidad de definir varios niveles de servicio dependiendo del perfil del cliente.

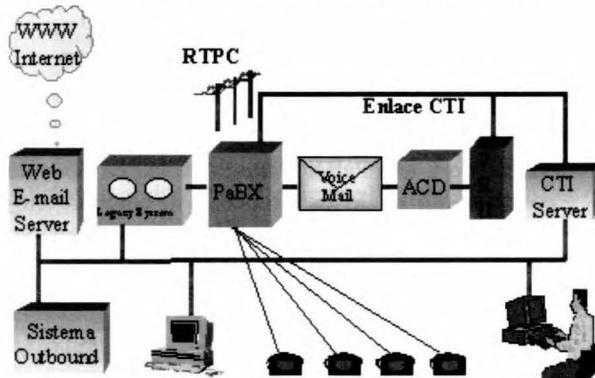
ARQUITECTURA

¿Quiénes son los actores principales de un *Centro de Atención al Cliente*? En una configuración convencional existen muchos equipos con diferentes funciones, interconectados a través de una red de área local (LAN) y/o una red de voz:

- **PaBX (Private automatic Branch eXchange).** Es el elemento básico de toda la infraestructura. Su misión es conectar la Red Telefónica Pública (RTPC) y gestionar las extensiones corporativas internas.
- **ACD (Automatic Call Distributor).** Este equipo permite gestionar grupos con distintas tareas y compe-

tencias así como crear colas de tamaño variable para gestionar los clientes de la lista de espera. Todo ello tiene como objetivo fundamental el incrementar la eficiencia y la productividad, ya el trabajo es repartido equitativamente entre los operadores.

Aproximación Tradicional



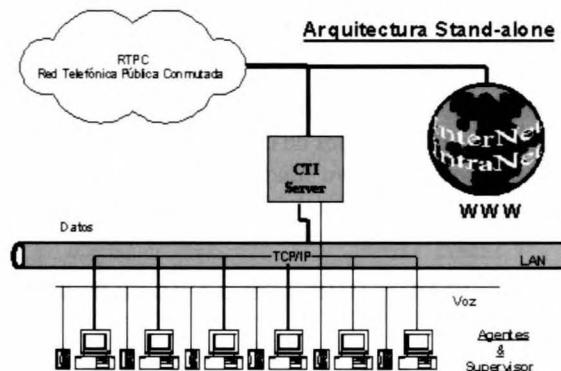
- **VRU (Voice Response Unit).** Sus funciones son muy diversas y abarcan desde el ofrecer información a través de mensajes simples (telephone notice board) hasta aplicaciones interactivas (telephone orders). Resulta clave para desarrollar servicios automáticos sin sobrecargar a los operadores (por ejemplo cuando el centro de servicios no está vigilado).
- **VMS (Voice Mail System).** Soporta funcionalidades de contestador avanzado y la posibilidad de DDI (Direct Dialing Inward).
- **CTI server.** El Servidor CTI (Computer Telephony Integration) une la infraestructura informática corporativa y la telefónica. Cuando la llamada llega a la operadora adecuada, en la pantalla aparece la plantilla con toda la información del cliente. Este automatismo libera al operador de tareas repetitivas (identificación del cliente) y le permite centrarse en el objetivo establecido con el cliente.

Hay que hacer notar que, por una parte, la conexión a la red LAN garantiza la coordinación y sincronización con las aplicaciones de los agentes, mientras que, por otra, el empleo de redes WAN permite controlar, mantener y reconfigurar el sistema incluso desde una estación de trabajo remota. Además, el soporte del estándar H.323 ofrece la posibilidad de implementar un Web Call Center.

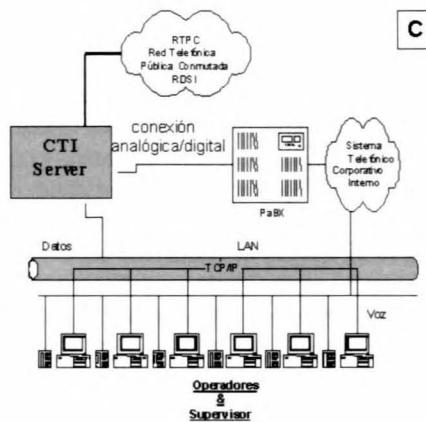
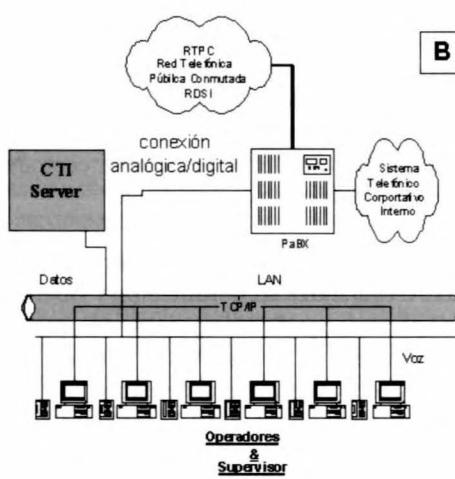
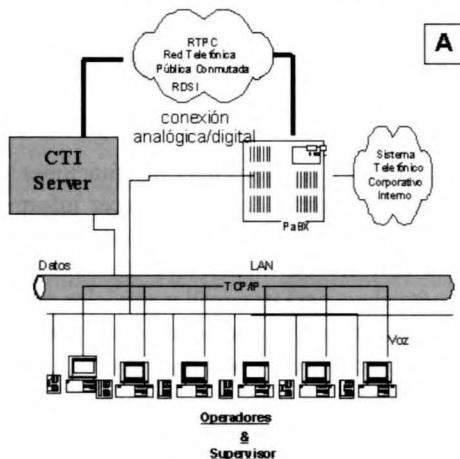


Existen varios modos de realizar un Centro de Atención al Cliente, dependiendo de si la PaBX está disponible o no:

- **Configuración “Stand alone”:** No necesita una PaBX externa y se caracteriza por maximizar el rendimiento de la inversión ofreciendo todas las funciones de un Centro de Atención al Cliente avanzado.



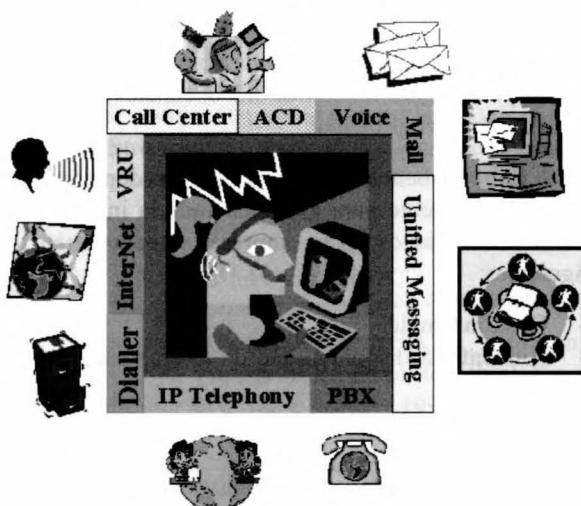
- **Configuración «with PaBX»:** el CTI-server está conectado a una PaBX a través de varios tipos de conexión utilizando los protocolos de comunicaciones telefónicas. Según la topología de la conexión, podemos distinguir tres casos (ver figura).



Arquitecturas conectadas a la PaBX

- A En paralelo a la PaBX
 - B En cascada a la PaBX
 - C Frente a la PaBX

FUNCIONALIDADES



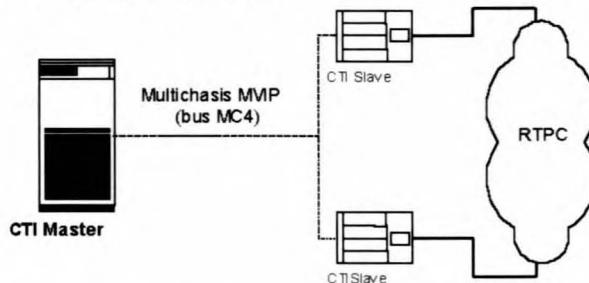
Call Center Distribuido

El soporte del estándar H.323 para la telefonía IP (VoIP) y del protocolo de intercambios de señalización SS7 permite realizar el Web Centro de Atención al Cliente o el Virtual Centro de Atención al Cliente.

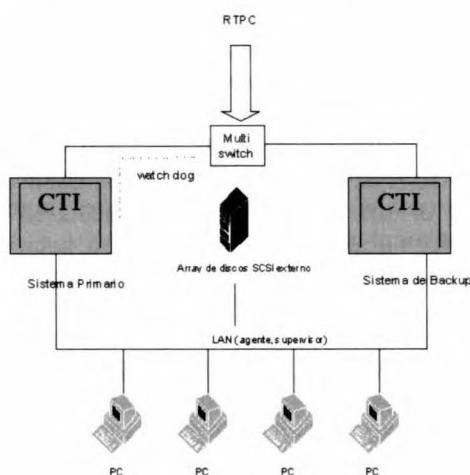
Los operadores pueden estar dispersos por diferentes lugares del territorio y, a la vez, trabajar juntos como operadores virtuales de un único e «invisible» Centro de Atención al Cliente gestionado de forma centralizada.

Para evitar el cese del servicio en caso de bloqueo, es posible equiparse con un sistema de *backup* que reemplace al sistema primario de manera transparente (*clustering*). Un servicio especial realinea los datos varia-

Configuración Multichasis



bles del sistema primario en el sistema de backup (configuración y estadísticas) y una señal de control especial, *watchdog*, permite encontrar el fallo o el bloqueo. En este caso, un dispositivo basado en retardos (*multi-switch*), conmuta todas las líneas de una máquina a otra con el fin de garantizar el mismo servicio que la maquina primaria, quizás con un número de líneas de operadores menor.



PaBX de última generación

Un servicio completamente software (*soft-PBX*) actúa como matriz de conmutación entre las extensiones y las líneas externas, que no hace mucho tiempo, se implementaban a través de máquinas específicas muy caras.

Todas las operaciones telefónicas son realizadas a través del módulo PBX, el cual interactúa con los otros módulos de software, particularmente con aquellos que gestionan el hardware y los canales de comunicación. Las funciones principales son:

- **Transferencia Ciega (flash hook).** El operador selecciona el número al que la llamada actual tiene que ser transferida y cuelga sin comprobar si hay una respuesta.
- **Transferencia comprobada.** En este caso, el operador comprueba no sólo si hay respuesta si no también la persona a la que la llamada ha sido transferida, por ejemplo, para informarle de la razón de la llamada.

- **Respuesta por ausencia.** Esta función es para contestar directamente desde el teléfono propio una llamada entrante de otro teléfono, por ejemplo, de un colega que se ausenta momentáneamente.
- **Música de espera.** El módulo de PBX puede coger música de una fuente externa (grabadora, wire broadcasting, etc.) y ponerla en línea mientras el cliente espera.
- **«No molestar».** Cuando esta función está activa, en vez de sonar el teléfono el sistema dirige directamente la llamada a un mensaje por defecto o bien permite dejar un mensaje.
- **Desvío mientras la línea está ocupada.** Cuando la línea está ocupada, la siguiente llamada es enviada a otro número de teléfono, bien sea interno o externo.
- **Desvío fijo.** Cualquier llamada a un teléfono se desvía a otro número de teléfono fijado de antemano.
- **Conferencia.** Esta función es utilizada para crear o mantener conferencias. Cada teléfono, dependiendo de la accesibilidad autorizada, puede crear o mantener conferencias.
- **(Silencio) Inclusión.** Esta función es utilizada normalmente por el supervisor del Call Center mientras los operadores se están formando, y consiste en que el supervisor puede escuchar la conversación que está siendo mantenida.
- **Menú personalizado.** Es posible, para usos especiales, asociar códigos numéricos a aplicaciones particulares desarrolladas en el sistema.

ACD Multimedia

Nunca más será necesario distinguir entre los distintos canales de comunicación interactivos entre clientes y empresas, puesto que un nuevo concepto de Call Center es capaz de gestionar todas las maneras posibles de pensar y entender el soporte a clientes. Se trata de un servicio optimizado que cubre todo lo relacionado con el cliente: con una sola herramienta, el operador es capaz de gestionar todos los canales de comunicación, reduciendo el tiempo de respuesta a los problemas y aventajando a la competencia.

Este paradigma aumenta dramáticamente las optimizaciones que pueden ser alcanzadas pues desde un mismo punto se puede gestionar todos los centros de contacto por medio de canales interactivos.

Uno de los aspectos más críticos a la hora de evaluar el funcionamiento de un Centro de Atención al Cliente es la división adecuada del trabajo entre los diferentes agentes. El servicio de ACD reparte los contactos entre los agentes uniformemente, supervisa la gestión de los grupos dependiendo de las capacidades del personal y la localización de los agentes. Todo ello permite ajustar el Centro de Atención al Cliente a las necesidades específicas de los clientes.



Cada agente del centro puede trabajar en diferentes productos o campañas, y recibe información del nuevo contacto por medio de un mensaje en la pantalla (pantalla pop-up), así que sabe de qué tema trata. Además, el servicio ACD multimedia redistribuye entre los agentes disponibles las llamadas entrantes (*inbound*) procedentes de la PBX, la WEB y el correo electrónico, y las llamadas salientes (*outbound*) son generadas automáticamente por el proceso automático de marcación. Esta característica permite al supervisor decidir si algún grupo de operadores deben recibir más llamadas de otro grupo, suspendiendo o reduciendo la prioridad de las acciones que llegan del módulo principal antes que la emergencia se acabe.

Además, todos los datos estadísticos relativos al tráfico telefónico están almacenados en una base de datos y pueden ser visualizados o impresos en tiempo real.

VRU

El servicio VRU (Unidad de Respuesta de Voz) es responsable de todos los servicios que tienen que ser ofrecidos sin el soporte de ningún operador.

El cliente llamante es recibido en un entorno con el que interactúa de manera organizada gracias al tonos DTMF o a comando de voz. De acuerdo con las elecciones realizadas por el cliente llamante, el sistema escoge el mensaje apropiado; el usuario puede interrumpir cada mensaje mandado, para acortar la interacción.

La aplicación VRU soporta un número amplio de aplicaciones modulares, algunas de las cuales son las siguientes:

- **Síntesis de números y fechas dinámicos.**
- **Reconocimiento de voz multilingüe (ASR, Automatic Speech Recognition).** Este servicio permite reconocer, independientemente de la voz del usuario, números, palabras e incluso frases especificadas en un diccionario definido durante la etapa de configuración.
- **Reconocimiento del cliente (SR, Speaker Recognition).** Permite reconocer únicamente al cliente que llama basándose en su tono de voz.
- **Síntesis de texto multilingüe (TTS, Text To Speech).** El mensaje de voz es sintetizado directamente de un texto.

Mensajería Unificada

El servicio de Mensajería Unificada es capaz de dar uniformidad a todos los mensajes, para utilizarlos con una única herramienta. Cualquier mensaje puede ser recibido en un formato electrónico manejable en el escritorio como una cuenta única. Por ejemplo, un fax aparecerá como mensaje con un fichero gráfico TIFF, donde se puede encontrar el documento que nos han mandado, y reenviarlo muy fácilmente vía fax o mail.

Pero qué pasa cuando estamos fuera de la oficina. Gracias al servicio de Mensajería Unificada, se pueden escuchar todos nuestros mensajes por un teléfono móvil. Teniendo acceso a un VRU especial y navegando por un menú interactivo, es posible escuchar un mensaje de voz del e-mail sintetizado gracias al servicio de Texto a Voz, e incluso el contenido de un fax si está equipado con un servicio de OCR.

Fax

Es posible aprovecharse de un servicio de fax potente e inmediato para transmitir directamente en un formato electrónico. Así, la compañía podrá crear aplicaciones de fax de la siguiente manera:

- **Fax Back.** Más allá de una campaña de marketing o a una solicitud explícita del cliente, el documento es mandado al número de fax definido en los datos del cliente o en la solicitud del cliente, vía una aplicación VRU. En cualquier caso, el coste del teléfono del fax mandado es soportado por el sistema.
- **Fax bajo demanda.** El documento es enviado utilizando la llamada del cliente, este más tarde deberá llamar desde un fax o una línea conectable a un fax. En este caso el cliente pagará los costes de teléfono relativos al fax mandado.

VoIP

La revolución de Internet es tan significativa porque no está limitada sólo a campos como el correo electrónico u otros, que a veces se centran más en la apariencia que en el contenido. Este nuevo límite es la Voz sobre IP (VoIP), que amplía increíblemente las fronteras de las aplicaciones y reduce drásticamente los costes de las comunicaciones.

Las tecnologías CTI se aprovechan de esta nueva oportunidad para ampliar aún más su capacidad multimedia y, sobre todo, sus capacidades del sistema multicanal. El Centro de Atención al Cliente es capaz de recibir todas las llamadas de teléfono vía Internet (Microsoft Netmeeting y otros...) y debe estar preparado para gestionarlos junto a todos los otros contactos, utilizando las mismas herramientas y procedimientos.

CONCLUSIONES

Tal y como hemos visto en los párrafos anteriores, la integración de los servicios de telefonía y de datos es ya una realidad. Por otra parte, la tendencia actual del mercado apunta a nuevos avances en materias como Centros Multimedia de Atención al Cliente, interacción e integración de la infraestructura telefónica corporativa en la web y, relacionado con la anterior, el desarrollo de aplicaciones de Telefonía IP.



LA MISIÓN DE OPORTUNIDAD SMOS DE LA SERIE EARTH EXPLORER.

RADIOMETRÍA POR SÍNTESIS DE APERTURA PARA LA MEDIDA DE LA HUMEDAD DEL SUELO Y LA SALINIDAD DEL OCÉANO

A. Camps, I. Corbella, J. Bará, F. Torres, N. Duffo, M. Vallllossera

camps@tsc.upc.es, corbella@tsc.upc.es, bara@tsc.upc.es, xtorres@tsc.upc.es,
duffo@tsc.upc.es, merce@tsc.upc.es

Universitat Politècnica de Catalunya, Campus Nord, D3, tel (34) 934016849,

ABSTRACT

Desde mediados de los años 80, diversas Agencias Espaciales han prestado una atención a los llamados radiómetros interferométricos por síntesis de apertura. Estos instrumentos ofrecen por primera vez un salto cuantitativo importante en resolución espacial como para permitir monitorizar la superficie terrestre a frecuencias bajas de microondas (banda L). En esta banda de frecuencias (1.4 GHz) existe la máxima sensibilidad de la temperatura de brillo tanto a la humedad del terreno, como a la salinidad del océano.

En los radiómetros clásicos, la resolución espacial viene dada por el ancho de haz de la antena que, al ser escaneada, forma la imagen de temperatura de brillo. Por ello, para alcanzar la resolución espacial deseada (30-50 km como máximo, 10-20 km ideal) desde un satélite en órbita baja, las antenas requeridas tienen unas dimensiones inaceptablemente grandes: entre 10 y 20 metros de diámetro.

Durante los años 90, la Agencia Europea del Espacio (ESA) llevó a cabo una serie de estudios tecnológicos con vistas a desarrollar un radiómetro por síntesis de apertura bidimensional en banda L. A este proyecto se le llamó **MIRAS** (Microwave Imaging Radiometer by Aperture Synthesis). En Noviembre de 1998, la misión **SMOS** (Soil Moisture and Ocean Salinity) basada en el concepto derivado de los estudios del proyecto MIRAS, fue propuesta como respuesta a un anuncio de «Misiones de Oportunidad Earth Explorer» lanzado por la ESA [1]. En Mayo de 1999, después de un proceso de selección de 27 propuestas, la ESA aprobó la misión SMOS en segundo lugar para una fase A extendida.

Este artículo describe brevemente la motivación de esta misión, los principios de funcionamiento de dicho instrumento y las actividades en las que ha participado y participa un grupo de profesores del

Departament de Teoria del Senyal i Comunicacions, de la Universitat Politècnica de Catalunya.

1. INTRODUCCIÓN

El progreso en la predicción del tiempo, monitorización del clima y predicción de desastres naturales pasa por una mejor cuantificación de la humedad del suelo (Soil Moisture, SM) y de la Salinidad del Mar (Sea Salinity, SS). Recientemente, los resultados de varios grupos de trabajo concluyen que nuevos progresos dependen en estos momentos de la disponibilidad de información global y periódica de la SM y de la SS.

Hoy en día es bien conocido que sobre la tierra, los flujos de agua y energía en la interfaz entre el suelo y la atmósfera dependen fuertemente de la SM. La evaporación, la infiltración y la escorrentía están regulados por la SM en la superficie, mientras que en la zona de las raíces regula la cantidad de agua que es capaz de absorber la vegetación. Por lo tanto, la variación espacio-temporal de la SM aparece como una variable clave en el ciclo hidrológico, y en consecuencia, en los modelos climáticos, de predicción del tiempo y de monitorización de la vegetación.

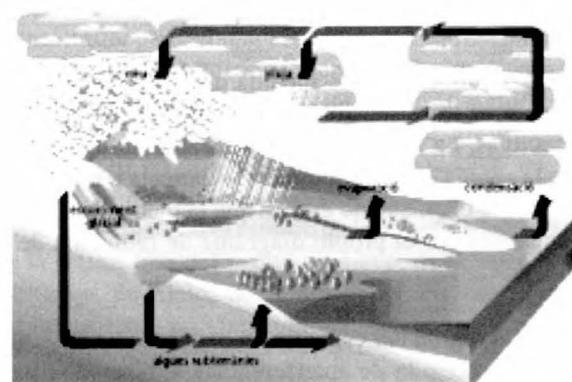


Figura 1. El ciclo hidrológico.



En el océano, la SS juega un papel importante, por ejemplo, en la zona subpolar del Atlántico Norte, donde intrusiones de agua de mar de baja salinidad influyen la circulación termohalina profunda y el transporte de calor meridional. Las variaciones de la salinidad también afectan la dinámica superficial en los océanos tropicales, donde la lluvia modifica la densidad de la capa superficial y los flujos de calor en la interfaz entre la superficie del océano y la atmósfera. Las variaciones espacial y temporal (anual e interanual) de la SS son, pues, un indicador del ciclo del agua e imponen las condiciones de contorno en los modelos que rigen el acoplamiento entre el océano y la atmósfera.

Aunque tanto la SM como la SS se utilizan habitualmente en los modelos atmosféricos, oceanográficos e hidrológicos, hoy en día no existe la capacidad de medir directa y globalmente estas dos variables claves. Como la realización de medidas in situ dista mucho de ser global, la única solución es la de una misión espacial dedicada.

2. RADIOMETRÍA DE MICROONDAS: CONCEPTOS BÁSICOS

La radiometría de microondas es la técnica más eficiente y precisa conocida para monitorizar la SM y la SS. Radiometría es la rama de la teledetección consistente en la medida de la radiación espontánea emitida por los cuerpos, en el caso que nos ocupa, en la banda de microondas. Esta medida viene caracterizada por la temperatura de brillo (T_B), proporcional a la temperatura física (T_{ph}) y a un parámetro llamado emisividad que depende de las características eléctricas de la superficie en cuestión, su rugosidad, la polarización p y los ángulos de observación en elevación (θ) y azimuth (ϕ).

$$T_{B,p}(\theta, \phi, f) = e_p(\theta, \phi, f) T_{ph} \quad (1)$$

Habitualmente la medida de T_B se realiza con receptores muy sensibles conectados a antenas muy directivas. La medida de la potencia de ruido captada por la antena es proporcional a la llamada temperatura de antena (T_A), que no es más que la media de $T_B(\theta, \phi)$ ponderada por el propio diagrama de radiación de la antena.

La región del espectro en banda L (1.400-1.427 MHz), protegida de interferencias y reservada para observación pasiva, ofrece una posibilidad única de medir estos dos parámetros, que no pueden ser detectados en ninguna otra banda de frecuencias.

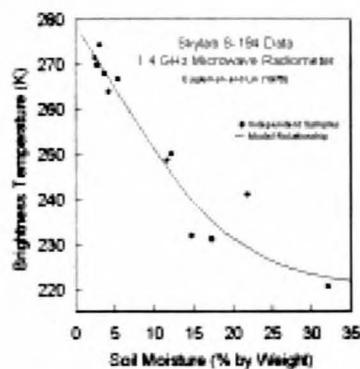


Figura 2. a) Temperatura de brillo del suelo en función de la humedad del terreno.

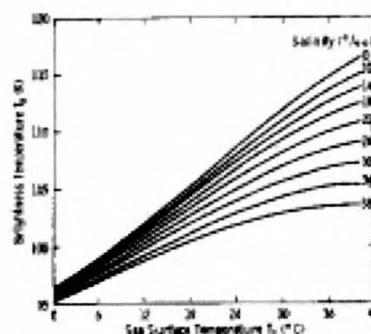


Figura 2.b) Temperatura de brillo del agua en función de la temperatura física y de la salinidad. [2].

Uno de los mayores inconvenientes de la radiometría en banda L es que requiere enormes antenas para conseguir una resolución espacial aceptable, lo que supone un reto tecnológico muy importante en una misión espacial. Por ello, aunque este concepto fue demostrado por algunos experimentos pioneros en banda L llevados a cabo en el SKYLAB en los 70, ninguna misión espacial dedicada fue lanzada posteriormente: para conseguir una resolución espacial aceptable ($\leq 50-60$ km) se requerían antenas de tamaños prohibitivos (radio del reflector ≥ 4 m). Las investigaciones que siguieron en años posteriores se centraron pues en medidas terrestres o en radiómetros embarcados en avión, con resultados muy esperanzadores.

La figura 3 muestra un mapa de SS y otro de SM obtenidos con los radiómetros en banda L SLFMR y ESTAR, respectivamente.

Gracias al desarrollo reciente de la llamada radiometría interferométrica por síntesis de apertura es posible conseguir estas prestaciones. La radiometría por síntesis de apertura está inspirada en el concepto de radioastronomía mediante interferometría de gran línea de base VLA (Figura 4).

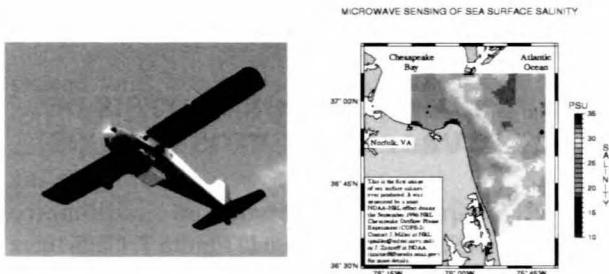


Figura 3.a) El sensor SLFMR en los bajos de un avión DeHavilland Beaver

Figura 3.b) Mapa de salinidad en la bahía de Chesapeake obtenida con el sensor SLFMR [3]

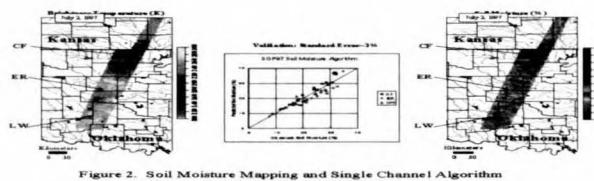


Figura 3.c) Mapa de temperatura de brillo obtenida con el sensor ESTAR y mapa de humedad del terreno asociado [4].

La idea consiste en colocar en una estructura desplegable un conjunto de pequeños receptores, para después resconstruir la temperatura de brillo de la escena con una resolución comparable a la de una antena de radio cuyo tamaño fuese igual a la separación entre los receptores más alejados.

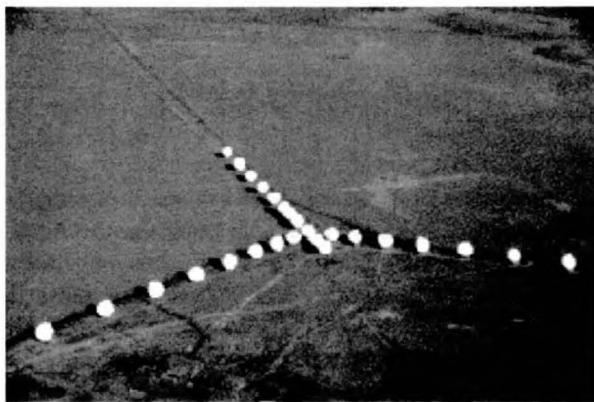


Figura 4. El Very Large Array (VLA) en Socorro, Nuevo Méjico, USA.

Esta idea fue propuesta inicialmente por D. M. LeVine (NASA Goddard) et al., en los años 80 con el proyecto ESTAR (Electrically Steered Thinned Array Radiometer) y validado con un sistema aerotransportado desarrollado en la Universidad de Massachusetts en Amherst. Este sistema obtenía reso-

lución angular en una dirección mediante apertura real (antenas tipo bastón) y en la otra dirección mediante síntesis de apertura (Figura 5).

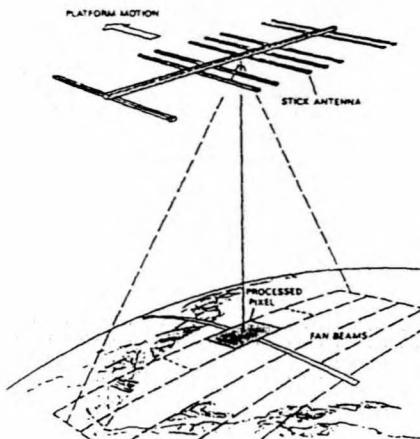


Figura 5. El radiómetro por síntesis de apertura ESTAR(Electrical Steered Thinned Array Radiometer

En Europa, un concepto mejorado del mismo fue estudiado por la Agencia Europea del Espacio. Es el proyecto MIRAS (Figura 6). MIRAS capitaliza parte del diseño de ESTAR, y a la vez representa mejoras substanciales. Al realizar síntesis de apertura bidimensional se puede medir TB en un margen de ángulos de incidencia mayores y en las dos polarizaciones (vertical y horizontal, o paralela y perpendicular al plano de incidencia). Además, el instrumento mide una escena completa en sólo 0.3 s, lo que corresponde a un emborronamiento («blurring») de la imagen de 2.2 km, menos de un 10% del tamaño de píxel. A medida que el satélite avanza, cada píxel se ve bajo diferentes ángulos de incidencia en cada una de las imágenes, lo que permite recuperar parámetros de la superficie con mucha mayor precisión [5].

Con esta perspectiva la misión SMOS fue propuesta a la Agencia Europea del Espacio [2]. Es una misión con unos objetivos científicos amplios y ambiciosos, a la vez que puede considerarse como un demostrador para allanar el camino de futuros sistemas que utilicen estas mismas técnicas. Se prevé que la misión SMOS genere además información significativa del contenido de agua de la vegetación, lo que puede ser muy útil en la estimación de producción de las cosechas. Finalmente, se espera poder realizar un progreso importante en el estudio de la círosfera, mejorando la estimación del manto de nieve, de la estructura multicapa de hielo, del hielo en el mar etc. Estos parámetros son igualmente importantes en el estudio del cambio climático.

La misión SMOS pretende obtener, sobre el océano abierto, mapas globales de salinidad con una precisión mejor que 0.1 PSU (aproximadamente 1

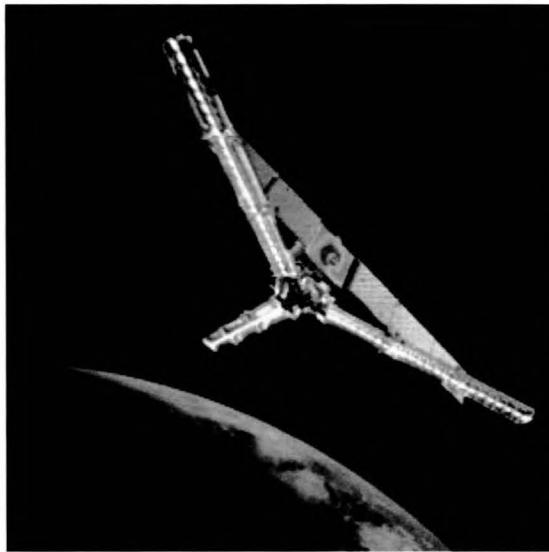


Figura 6. Geometría de observación del instrumento MIRAS (misión SMOS). Nótese el parecido con el VLA (Figura 4), aunque en el VLA las antenas están espaciadas logarítmicamente y en MIRAS están equiespaciadas 0.89λ .

PSU = 1 parte por mil) cada varios días, con una resolución espacial de 200 km; y sobre la tierra mapas globales de humedad del suelo con una precisión de $0.035 \text{ m}^3/\text{m}^3$ cada 3 días, con una resolución espacial mejor que 60 km, así como contenido de agua en la vegetación con una precisión de 0.2 kg m^{-2} . La plataforma estará en órbita heliosíncrona (6 a.m.) a una altura de 757 km. La Tabla 1 resume las características principales.

Tabla 1 Principales parámetros de la misión.

Parámetro	Valor
Tamaño*	~ 4.5 m cada brazo (Y)
Peso	175 kg
Consumo	220 W
Swath	620 km
Resolución espacial	30 – 90 km
Sensibilidad radiométrica	0.8 – 2.2 K
Precisión radiométrica	< 3 K
Fecha de lanzamiento prevista	2005

*Nótese que, aunque el tamaño del brazo sea comparable al del radio del reflector requerido para alcanzar la misma resolución espacial, la masa y volumen del radiómetro interferométrico son mucho menores.

SMOS es una misión ambiciosa, basada en un concepto de instrumento novedoso que ha requerido y requerirá en los próximos años de una cantidad de trabajo muy considerable tanto en el diseño de la misión y como del propio instrumento. El presente

La misión SMOS pretende obtener, sobre el océano abierto, mapas globales de salinidad con una precisión mejor que 0.1 PSU

artículo pretende describir los conceptos básicos del instrumento y de la misión.

3. RADIOMETRÍA POR SÍNTESIS DE APERTURA: CONCEPTOS BÁSICOS

El esquema básico de un radiómetro interferométrico se presenta en la Figura 7. Cada línea de base mide la correlación cruzada entre las señales $b_1(t)$ y $b_2(t)$ captadas por cada par de antenas que forman la agrupación dispersa (Figuras 5 y 6).

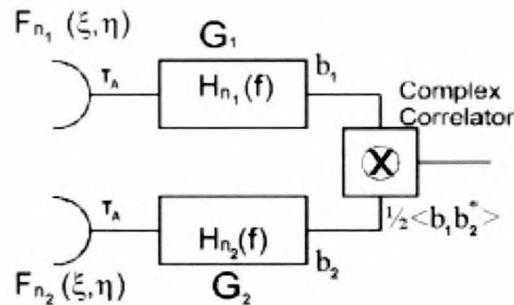


Figura 7. Esquema básico de una línea de base en un radiómetro interferométrico.

En el caso ideal, antenas y receptores idénticos y ancho de banda relativo despreciable, suponiendo que las antenas de la agrupación están situadas sobre el plano XY en las posiciones (x_1, y_1) y (x_2, y_2) , cada correlación cruzada es una muestra de la llamada función de visibilidad $V(u,v)$ (unidades: Kelvin) [6,7]

$$\begin{aligned} V_{12}(u_{1,2}, v_{1,2}) &= \frac{1}{k_B B G} \frac{1}{2} \langle b_1(t) b_2^*(t) \rangle = \\ &= \frac{1}{\Omega} F \left[\frac{T_B(\xi, \eta)}{\sqrt{1 - \xi^2 - \eta^2}} |F_n(\xi, \eta)|^2 \right], \end{aligned} \quad (2)$$

donde $(u_{1,2}, v_{1,2}) = (x_2 - x_1, y_2 - y_1)/\lambda$ es el espaciado entre antenas normalizado a la longitud de onda, $(\xi, \eta) = (\sin \theta \cos \phi, \sin \theta \sin \phi)$ son los cosenos directores respecto de los ejes X e Y, y F representa la transformada de Fourier entre los dominios (ξ, η) y (u, v) . $T_B(\xi, \eta)$ es la temperatura de brillo (Kelvin), $|F_n(\xi, \eta)|^2$ es el diagrama de radiación normalizado de las antenas (sin unidades, y supuesto igual para todas), k_B es la constante de Boltzman, Ω es el ángulo sólido de antena, G es la ganancia en potencia de cada cadena receptor y B es el ancho de banda ruido.

Como la temperatura de brillo es una función en dos dimensiones limitada al círculo unidad ($\xi^2 + \eta^2 = \sin^2 \theta \leq 1$), su transformada de Fourier (función de visibilidad) es muestreada de manera óptima sobre una malla hexagonal en el plano (u, v) . La gran ventaja de los

sistemas interferométricos sobre los de apertura real reside en su mayor resolución angular. En concreto, se puede demostrar que ésta es la misma que la de un phased array con una antena situada en cada punto (u,v). La Figura 8 muestra los puntos de muestreo correspondientes a una agrupación en Y con 23 antenas por brazo.

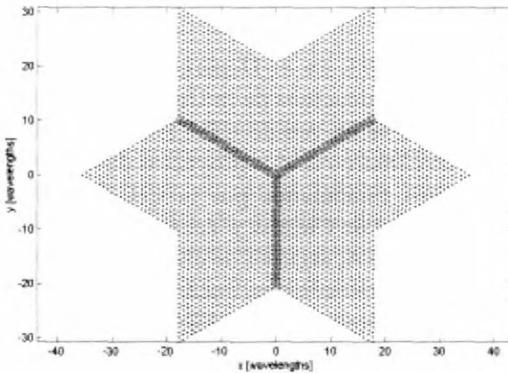


Figura 8. Cobertura de frecuencias espaciales (u,v) asociadas a una agrupación en Y con 23 antenas por brazo separadas $d=0.89$ longitudes de onda.

Respecto el muestreo rectangular, el muestro hexagonal permite una reducción del 13.4 % de las muestras de visibilidad y del hardware asociado [8,9]. El único problema es que debido a limitaciones tecnológicas (tamaño de las antenas y acoplamientos mútuos entre ellas) la mínima línea de base no se puede hacer igual a longitudes de onda (criterio de Nyquist para muestreo hexagonal, en vez de 1/2 para muestreo rectangular). En este caso, tendremos aliasing y el campo de visión (FOV, Field Of View) estará limitado por la repetición periódica del círculo unidad centrado en $(1/(3d), 1/d)$, $(1/(3d), -1/d)$, $(-1/(3d), 1/d)$, $(-1/(3d), -1/d)$, $(2/(3d), 0)$, y $(-2/(3d), 0)$. Sin embargo, como una gran parte del círculo unidad está ocupada por el cielo, con una temperatura de brillo conocida y mucho menor que la temperatura de brillo de la tierra, se pueden aplicar algunas técnicas de pre-procesado [9] a fin de ensanchar el FOV libre de alias hasta la frontera Tierra-cielo (Figura 9).

Finalmente, la Figura 10 presenta el FOV libre de alias tal y como aparece sobre la superficie de la Tierra [1]. Las Figuras 9 y 10 son parecidas, pero como la agrupación está inclinada hacia delante, el FOV parece ensancharse en la parte superior, y encogido en la parte inferior.

Se presentan tres familias de curvas: ángulo de incidencia constante (i), ángulo respecto de la dirección perpendicular al plano de la agrupación (θ_a), y resolución espacial (Δs). Estas tres familias son distin-

tas ya que el plano de la agrupación está inclinado, esto es $\beta \neq 0$.

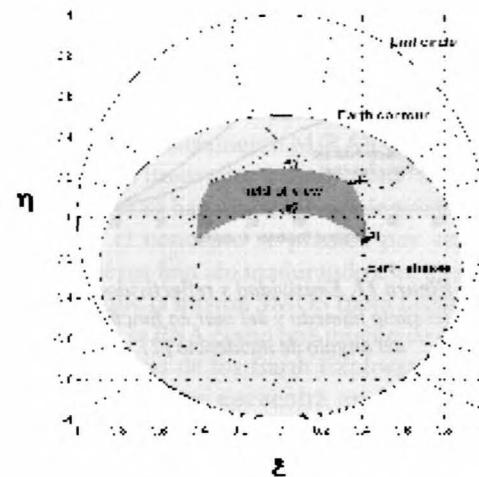


Figura 9. Contorno de la Tierra (línea continua) y alias de la Tierra (líneas discontinuas) limitando el FOV libre de alias para un instrumento como MIRAS, en la siguiente configuración altura =639Km, cabeceo $\beta=34^\circ$, y separación entre antenas $d=0.89 \lambda$. En esta configuración la anchura del FOV libre de alias es de 725 Km para ángulos de incidencia sobre la tierra entre 40° y 55° .

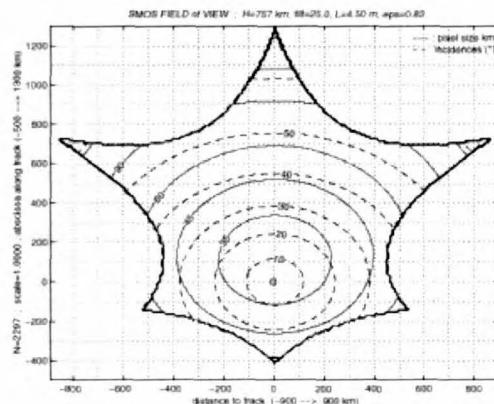


Figura 10. FOV de una configuración de SMOS. El eje y es paralelo a la dirección de avance del satélite [1]

A la hora de evaluar las prestaciones de la misión SMOS la resolución espacial juega una especial relevancia. El ángulo θ_a no sólo influye en la resolución espacial, sino en la sensibilidad radiométrica ya que el diagrama de radiación de la antena elemental debe ser tenido en cuenta. Por otra parte, la temperatura de brillo emitida por la Tierra depende del ángulo de incidencia para cada polarización.



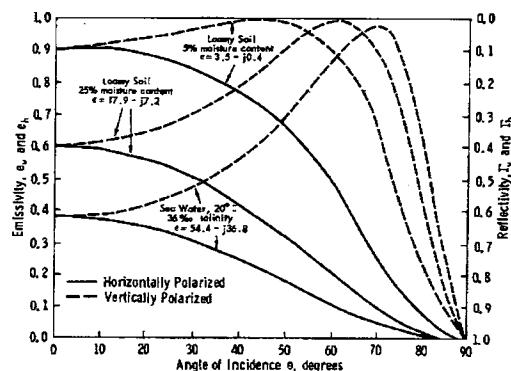


Figura 11. Emisividad y reflectividad del suelo húmedo y del mar en función del ángulo de incidencia [2]

Como se comentó anteriormente, en cada periodo de medida se obtiene un mapa completo bidimensional de temperaturas de brillo, T_B . A medida que el satélite avanza, cada píxel sobre la Tierra se ve varias veces, con ángulos de incidencia variables. Esto es, cada píxel se mueve de arriba a abajo a lo largo de un segmento vertical en la Figura 10. A medida que avanza, el ángulo de incidencia, el tamaño del píxel, y el ángulo respecto de la normal al array varían.

MIMOSA es un proyecto a tres años actualmente en proceso de evaluación

En términos de recuperación de la humedad de suelo este hecho es de considerable interés [5]. En primer lugar, proporciona un número mayor de muestras independientes; y en segundo lugar, permite mejorar la discriminación entre parámetros de la superficie desconocidos (espesor óptico de la vegetación, etc.), ya que introducen distintas variaciones en T_B con el ángulo de incidencia.

WISE es un proyecto a un año financiado por la ESA que prevé una campaña de medidas de un mes de duración en la plataforma petrolífera Casablanca de Repsol, en las costas de Tarragona

En términos de recuperación de la salinidad del mar también se espera que la información multivista

sea importante a la hora de descartar píxeles contaminados con radiación proveniente del Sol a través de reflexiones en la superficie de la mar, y a la hora de tener en cuenta la variación azimutal de la emisividad del mar debida a la rugosidad introducida por el viento.

4. LA UNIVERSITAT POLITÈCNICA DE CATALUNYA Y EL PROYECTO MIRAS

Desde 1993, en el Departament de Teoria del Senyal i Comunicacions de la Universitat Politècnica de Catalunya se ha venido trabajando en el proyecto MIRAS en el marco de dos proyectos de investigación de tres años financiados por la CICYT (Comisión Interministerial de Ciencia y Tecnología) y ocho convenios con la ESA, algunos de ellos como consultores de MIER Comunicaciones, encargada de la construcción de los receptores, y CASA, encargada de las antenas, estructura mecánica etc. y contratista principal del llamado Proyecto Piloto del Demostrador de MIRAS. La mayor parte del trabajo realizado se ha centrado en el estudio del instrumento en sí mismo desde el punto de vista de ingeniería: análisis de errores, técnicas de calibración e inversión, etc.

La radiometría por síntesis de apertura bidimensional ofrece una oportunidad única para obtener dos variables geofísicas de vital importancia en el estudio del clima y su variación, como son: la humedad del terreno y la salinidad del mar

En la actualidad, las dos grandes líneas de actividad se centran en los proyectos WISE - MIMOSA y el desarrollo del simulador de MIRAS.

Los proyectos WISE y MIMOSA- Estos dos proyectos pretender estudiar los efectos de la salinidad, del viento (intensidad y dirección del oleaje y espuma) y de la temperatura superficial del mar en la temperatura de brillo del mar en banda L, para diferentes ángulos de incidencia y de azimut. Se pretende con ello tener modelos fiables para ser incluidos en los algoritmos de recuperación de salinidad que se aplicarán a los datos generados por la misión SMOS.

WISE es un proyecto a un año financiado por la ESA que prevé una campaña de medidas de un mes de

duración en la plataforma petrolífera Casablanca de Repsol, en las costas de Tarragona (Figura 12). Además de UPC, cuenta con la participación del Institut de Ciències del Mar y LODYC (Francia). Sus objetivos se focalizan básicamente en el modelado de los llamados parámetros de Stokes en banda L. Los parámetros de Stokes son las temperaturas de brillo en polarización vertical y horizontal, y las partes real e imaginaria de la correlación cruzada de los campos en polarizaciones vertical y horizontal.

MIMOSA es un proyecto a tres años actualmente en proceso de evaluación. Sus objetivos son más ambiciosos y, caso de ser aprobado, incluirán una campaña de medidas de 4 meses de duración en la misma plataforma, con vuelos simultáneos de otros radiómetros en banda L etc. Su objetivo último es ya el desarrollo de los algoritmos que, partiendo de los datos de SMOS, lleguen a la recuperación de la SS. Además, se estudiará el impacto de la salinidad, su variación con la profundidad etc. en los modelos de circulación oceánica y de interacción océano-atmósfera.



Figura 12. La plataforma petrolífera Casablanca de Repsol en las costas de Tarragona.

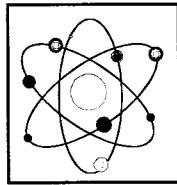
El Simulador de SMOS. El simulador de SMOS pretende optimizar y completar un programa ya existente que permite estudiar el comportamiento del instrumento: desde la generación de la órbita de la plataforma hasta el análisis de errores en las temperaturas de brillo recuperadas, pasando por un generador de escenas de temperatura de brillo a partir de parámetros geofísicos, un modelado de los errores instrumentales, técnicas de calibración etc. Este trabajo se llevará a cabo dentro de un proyecto financiado por la CICYT, y probablemente con apoyo de la ESA.

5. CONCLUSIONES

La radiometría por síntesis de apertura bidimensional ofrece una oportunidad única para obtener dos variables geofísicas de vital importancia en el estudio del clima y su variación, como son: la humedad del terreno y la salinidad del mar. A pesar de haberse utilizado desde hace varias décadas en radioastronomía, el concepto del instrumento MIRAS para observación de la Tierra es innovador y ambicioso. Gran parte de los desarrollos se han realizado desde principios de los 90, cuando el concepto se planteó por vez primera. Estos progresos han ido madurando esta técnica, hasta el punto de que la misión SMOS fuera factible, y fuera aprobada por la ESA como una misión de oportunidad dentro de la serie de los Earth Explorer. En la actualidad el proyecto se encuentra en fase A, donde se llevará a cabo un análisis detallado de toda misión, y si no hay imprevistos, el lanzamiento está previsto para el año 2005. Este artículo describe las principales motivaciones de dicha misión, los principios básicos de las técnicas de síntesis de apertura en radiometría y las principales contribuciones de la Universitat Politècnica de Catalunya a dicho proyecto.

REFERENCIAS

- [1] Kerr et al., 1998, MIRAS on RAMSES: radiometry applied to soil moisture and salinity measurements, Full proposal, A.O. Earth Explorer Opportunity Missions, ESA, 1998. SMOS web site: <http://www-sv.cict.fr/cesbio/smoss>
- [2] Ulaby et al., Microwave Remote Sensing, Vol I, ed. Artech House, Boston MA, 1981
- [3] <http://www.vims.edu/facilities/BeaverSLFMR.htm>, <http://www.quad-eng.com/terra/lb/lb.html>
- [4] <http://maximus.ce.washington.edu/~tempcm/Post2002/smm3.html>
- [5] Wigneron, J.P., P. Waldteufel, A. Chanzy, J. C. Calvet, O. Marloie, Hanocq, and Y. Kerr, «Retrieval capabilities of L-Band 2-D interferometric radiometry over land surfaces (SMOS Mission)», 6th Specialist Meeting on Microwaves Radiometry, VSP, Zeist, The Netherlands, in press, 1999.
- [6] Thompson, A. R., J. M. Moran, and G. W. Swenson, Interferometry and Synthesis in Radio Astronomy, John Wiley and Sons, 1986.
- [7] Ruf, C. S., C. T Swift, A. B. Tanner, D. M. LeVine, «Interferometric Synthetic Aperture Radiometry for the Remote Sensing of the Earth», IEEE Trans. on Geoscience and Remote Sensing, Vol. 26, N° 5, pp 597-611, September 1988.
- [8] Mersereau, R. M., «The Processing of Hexagonally Sampled Signals», Proceedings of the IEEE, Vol 67, N° 6, pp 930-949, June 1979.
- [9] Camps, A., «Application of Interferometric Radiometry to Earth Observation», Tesis Doctoral, 1996, Universitat Politècnica de Catalunya.



COMPUTACIÓN CUÁNTICA: NUEVAS PERSPECTIVAS EN EL TRATAMIENTO DE LA INFORMACIÓN

Pedro J. Salas Peralta*, Ángel L. Sanz Sáenz**

(*) Profesor Titular del Dpto Tecnologías Especiales Aplicadas a la Telecomunicación

(**) Profesor Titular del Dpto. Física Aplicada a las Tecnologías de la Información

Universitat Politècnica de Madrid Ciudad Universitaria s/n, 28040 Madrid

INTRODUCCIÓN

El concepto de ordenador se pierde quizás en la historia del ser humano. Los primeros se construyeron para propósitos muy concretos y eran poco versátiles. Un ejemplo son los monumentos megalíticos como las estructuras de Stonehenge, en Inglaterra, que servían para predecir eventos astronómicos. Evidentemente si se hubiera deseado utilizar tales construcciones para otros cometidos habría sido necesario cambiar las piedras de sitio o incluso de tamaño. El proceso de programación sería realmente difícil.

A lo largo de la historia, las máquinas de cálculo se fueron complicando a la vez que se iban volviendo más versátiles, pasando desde el abaco hasta las máquinas calculadoras de Pascal o Leibnitz. Sin embargo, la escasa utilidad de tales máquinas para la vida diaria de la época, las relegó al olvido durante cien años.

A medida que los ordenadores han aumentado su velocidad de funcionamiento, su tamaño ha ido disminuyendo debido a que la velocidad de la luz es finita. La tecnología de los ordenadores ha evolucionado siguiendo un proceso de miniaturización que lleva de los relés, válvulas, transistores hasta los circuitos integrados... Parece que el próximo nivel será el molecular. Sin embargo, a este nivel no sólo hay que tener en cuenta la Mecánica Cuántica (MC) para conseguir que los dispositivos funcionen correctamente, sino que la MC participa activamente en el comportamiento global. En la actualidad se está aprendiendo a captar y aprovechar las ventajas derivadas de considerar a la información como un ente cuántico, lo que abre una serie de nuevas posibilidades que darían vértigo al propio Bohr.

TEORÍA CLÁSICA DE LA COMPUTACIÓN

Cabe, quizás, situar el origen moderno de la teoría de la computación en la respuesta a un problema

esbozado por David Hilbert en 1900. Hilbert planteó la necesidad de preguntarse no sobre la veracidad de ciertas proposiciones matemáticas, sino acerca de la capacidad real de las matemáticas para responder a tales cuestiones. Desplazó así el interés acerca de las soluciones dadas por las matemáticas a determinadas sentencias hacia la demostración de la posibilidad de tal cosa.

La tecnología de los ordenadores ha evolucionado siguiendo un proceso de miniaturización... parece que el próximo nivel será el molecular

Turing se enfrentó a este desafío y empezó a pensar en una posible solución "mecánica" del problema. Introdujo el concepto de Máquina de Turing (MT), dispositivo formado por un elemento de lectura/escritura y una memoria en forma de cinta con capacidad ilimitada, que permite estudiar la computación desde dos puntos de vista:

a) Posibilidad de computar ciertas funciones

Las MT realizan cálculos mecánicos que ejecutan un algoritmo o procedimiento efectivo. Podemos preguntarnos si el concepto de máquina de Turing engloba a todas las operaciones matemáticas representadas por algoritmos. La respuesta se conoce como la tesis de Church-Turing: el concepto de máquina de Turing define lo que entendemos por procedimiento algorítmico. Cualquier función que sea computable, se puede computar mediante una máquina de Turing constituida por un dispositivo físico "razonable". Se trata de una hipótesis no demostrada que ha superado todos los intentos de encontrar contraejemplos. Las MT no tienen la pretensión de llegar a ser construidas real-

mente, sino que sólo pretenden captar lo esencial del comportamiento de un ordenador.

Las MT no tienen que ser necesariamente deterministas (MTD): a un estado le sigue precisamente otro. Existe la posibilidad de que, partiendo de una configuración determinada, el siguiente estado (existen varias posibilidades) pueda alcanzarse con una cierta probabilidad. Estamos ahora ante una Máquina de Turing Probabilista (MTP). Su funcionamiento puede de representarse mediante un esquema en árbol. La computación sigue una única ruta que se produce con cierta probabilidad. El resultado de estas MTP puede no ser correcto, sólo lo es con cierta probabilidad, aunque el procedimiento podría llegar a ser más efectivo que en el caso determinista. Se demuestra que todo lo que es computable mediante una MTP también lo es mediante una MTD.

Uno de los problemas de los actuales ordenadores de alta velocidad, es la eliminación del calor producido durante su funcionamiento

Desgraciadamente, la respuesta dada por las MT a la pregunta planteada por Hilbert acerca de la posible existencia de un método general que permita averiguar si una proposición es verdadera o falsa, tiene una respuesta clara: no, no existe tal método. Es imposible predecir si una determinada MT se detendrá o no, en otras palabras, si podrá o no proporcionar una solución al problema planteado (Problema de la Parada).

b) Eficiencia

El segundo aspecto interesante de las MT es que permiten estudiar la eficiencia de los algoritmos. La eficiencia se establece clasificando los algoritmos en determinadas Clases de Complejidad en función de cómo escalan los recursos de un cálculo con el tamaño de los datos. Si el tamaño de los datos se mide a través de los bits necesarios para su representación (L), los algoritmos se pueden clasificar en tratables (o de clase P), si el número de pasos temporales (p) escala como $O(\text{polinomio de } L)$, y en no tratables (clase NP), si escalan como $p \sim O(\text{exponencial de } L)$. Esta forma de clasificación no depende de la velocidad real de los ordenadores actuales (siempre cambiante). Durante mucho tiempo se creyó que este tipo de clasificación no tenía ninguna relación con el tratamiento físico de la información, de ahí su importancia (la situación iba a cambiar con la computación cuántica).

El modelo de la máquina de Turing es una forma de describir un ordenador en abstracto. Otra forma de hacerlo es construir un circuito mediante elementos (puertas) que realicen operaciones lógicas sobre un conjunto de variables booleanas. Ambas aproximaciones son polinómicamente equivalentes. De la misma forma que existe una MT Universal (que permite simular cualquier otra), existe un conjunto de puertas que son universales. Por ejemplo, la puerta NAND por sí sola es universal. Utilizando esta puerta se puede reproducir el comportamiento de cualquier MT, con recursos polinómicos.

LIMITACIONES DEL PROCESO DE CÓMPUTO

Hacia el inicio de la década de los 60, Rolf Landauer¹ comenzó a preguntarse si las leyes físicas imponían algunas limitaciones al proceso de cómputo. En concreto se interesó sobre el origen del calor disipado por los ordenadores, y si este calor era algo inherente a las leyes de la física o se debía a la falta de eficiencia de la tecnología disponible. Este tema parece realmente interesante si recordamos que uno de los problemas de los actuales ordenadores de alta velocidad es la eliminación del calor producido durante su funcionamiento. Estas reflexiones iban a ser el germen de las actuales ideas acerca de los ordenadores cuánticos.

La pregunta era: ¿se podría idear una puerta que funcionara de forma reversible, y que por tanto no disipa energía?. La respuesta no estaba clara, ya que la lógica clásica se basaba en puertas no reversibles, es decir que no permitían obtener los bits de partida después de realizada la operación. Esto es lo que sucede, por ejemplo en la puerta NAND.

La idea de computación clásica reversible la introdujo matemáticamente Yves Lecerf en 1963 y la desarrolló Bennett en 1973 demostrando que, desde un punto de vista teórico, es posible la existencia de una máquina de Turing reversible. En la representación de circuitos se plantearon puertas clásicas reversibles como la puerta CNOT (control not). Esta puerta actúa sobre pares de bits, donde se realiza una operación NOT sobre el segundo bit sólo si el primero es "1". Es posible obtener una única puerta universal reversible tal como lo es NAND para la lógica irreversible: se trata de la puerta que introdujo Toffoli y que lleva su nombre; no es mas que una "controlled-controlled-NOT" (CCNOT).

La existencia de tales máquinas de Turing reversibles nos indica que no hay una cantidad mínima de energía que haya que poner en juego para efectuar un cómputo concreto.



COMPUTACIÓN Y FÍSICA

La teoría clásica de la computación habitualmente no hacía referencia a la física del dispositivo, y se suponía que los fundamentos de tal teoría eran independientes de la realización física de los mismos. Hubieron de pasar 20 años antes de que Deutsch, Feynman y otros pusieran de manifiesto que esta idea era falsa, mostrando la conexión entre las leyes de la física y la información, en concreto con la computación. A partir de aquí se produjo una más de tantas uniones entre ideas distintas que han aparecido en la física: computación y MC. De esta unión surgió la Computación Cuántica.

De forma general podemos decir que la computación es la creación de conjuntos de símbolos (resultados) a partir de ciertos conjuntos de símbolos iniciales (o datos). Si interpretamos los símbolos como objetos físicos, la computación correspondería a la evolución de los estados de los sistemas. Por tanto, dicha evolución es un ejemplo de computación. *Si la evolución es cuántica, tenemos la Computación Cuántica.*

MÁQUINAS DE TURING CUÁNTICAS

Ya que el sentido común deja de ser correcto cuando descendemos a los reductos cuánticos, podría suceder que el paradigma de la MT no fuera todo lo independiente de la física que se deseaba. La pregunta surgió con cierta timidez: ¿serían las MT basadas en la Mecánica Cuántica equivalentes a las clásicas?. Dado que la MC permitía nuevas formas de evolución a través de estados coherentes, la respuesta se adivinaba negativa.

La posibilidad de que una máquina de Turing cuántica pudiera hacer algo genuinamente cuántico fue planteada por Richard Feynman² (1982), demostrando que ninguna máquina de Turing clásica (probabilista o no) podía simular algunos comportamientos cuánticos sin incurrir en una ralentización exponencial; sin embargo una máquina de Turing cuántica sí podía hacerlo. Este comportamiento surge del hecho de que la dimensión del espacio de Hilbert accesible al sistema aumenta de forma exponencial (2^n) con el número de amplitudes (n) a manejar y guardar. Feynman describió un "simulador cuántico universal" que simulaba el comportamiento de cualquier sistema físico finito. Desafortunadamente, Feynman no diseñó este simulador y su idea tuvo poco impacto.

El siguiente paso se dio en 1985, cuando David Deutsch³ describió la primera máquina de Turing cuántica (MTC). Esta MTC podía realizar tareas que una clásica no podía. Los procesos totales del ordena-

dor cuántico deben ser unitarios y por tanto no disipativos y usa una lógica reversible. Las MTC dieron lugar a una modificación de la hipótesis de Church-Turing, en el siguiente sentido: "existe (o puede construirse) un ordenador universal que puede programarse para simular cualquier sistema físico finito operando con unos recursos limitados".

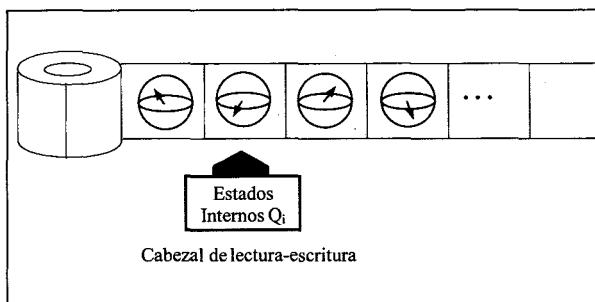


Figura 1. Máquina de Turing Cuántica

El funcionamiento de esta MTC es simple. Está formada por un cabezal de lectura-escritura que recorre la cinta y puede adquirir un conjunto finito de estados. Dependiendo del estado que lee en cada casilla (qubit, cuya representación aparece en la figura 1), ejecuta una instrucción (borra el qubit, lo modifica o lo deja igual) y se desplaza a una nueva casilla. Existen dos características cuánticas que proporcionan la potencia a la computación cuántica:

1.- Paralelismo:

Esta propiedad surge de la propia representación de la información cuántica. El fragmento de información clásico fundamental es el bit, entendiéndose como tal un sistema material que puede adoptar uno de los dos posibles estados distintos que representan dos valores lógicos (0 y 1 o sí y no, verdadero y falso, etc.). Sin embargo si la codificación de la información es cuántica, y se hace a través de, por ejemplo, dos estados de un sistema microscópico, ahora también es posible un estado del sistema que sea una superposición coherente de estos 0 y 1. Ello implicaría que, por ejemplo, un átomo descrito por este estado, estaría en «ambos estados a la vez». Este estado no sería ni un 0 ni un 1 clásicos. La existencia de estos estados "esquizofrénicos" nos indica que el nuevo ordenador cuántico tiene que poder tratar estos estados: generarlos y trabajar con ellos. En este sentido los ordenadores cuánticos serán algo distintos a los clásicos. De forma general, a un sistema cuántico con dos estados ($|Q\rangle$, lo llamaremos bit cuántico o simplemente qubit), de forma que estará representado por el estado general:

$$|Q\rangle = a|0\rangle + b|1\rangle$$

donde $|0\rangle$ y $|1\rangle$ son los dos estados en los que puede estar el sistema y los coeficientes a y b son, en

general, números complejos, y si el qubit está normalizado se cumplirá $|a|^2 + |b|^2 = 1$.

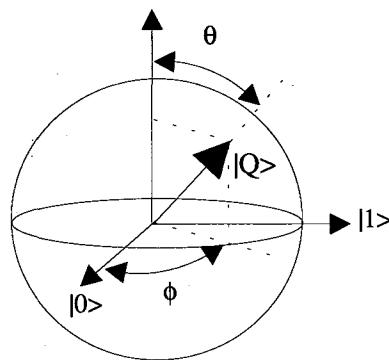


Figura 2. Representación de un qubit

$$|Q\rangle = \cos(\theta)|0\rangle + e^{i\phi}\sin(\theta)|1\rangle$$

En general, es posible representar a un qubit mediante tres coordenadas polares, tal como aparece en la figura 2. Notemos que debido a que no existen restricciones acerca de los posibles valores de estos coeficientes (salvo quizás la condición de normalización del vector de estado), un solo qubit contiene, en realidad "infinita información". Esto no representa un problema conceptual, pues para extraerla, necesitaríamos medir sobre el estado, lo cual implicaría su colapso (Postulado de Proyección) en uno de los dos bits clásicos $|0\rangle$ ó $|1\rangle$, y no sería posible extraer esa infinita información del qubit. De esta medida sólo podríamos extraer un bit clásico de información, lo que causa una complicación adicional en el planteamiento de la extracción de la información en los algoritmos típicamente cuánticos.

La primera potencia de los ordenadores cuánticos radica, precisamente, en la posibilidad de usar este tipo de superposiciones coherentes de qubits para realizar los cálculos. Consideraremos que queremos calcular una función de la variable x definida por:

$$f: x \in \{0, 1, \dots, 2^m - 1\} \rightarrow \{0, 1, \dots, 2^n - 1\}$$

Un ordenador clásico haría 2^m cálculos, obteniendo $f(0), f(1) \dots f(2^m - 1)$. En un ordenador cuántico el cálculo es ligeramente distinto. Para realizar el cálculo cuántico debemos usar dos registros cuánticos: $|x\rangle$ y el resultado $f(x)|f(x)\rangle$, y la evolución debe hacerse mediante un operador unitario U_f que actúe sobre un registro cuántico total:

$$U_f\{|x\rangle \otimes |0\rangle\} = |x\rangle \otimes |f(x)\rangle$$

Inicialmente necesitamos un registro en el estado "cero", $|0\rangle$, donde se va a colocar el resultado después de la operación.

En realidad, podemos hacer algo más que un cálculo uno a uno de los valores de f , ya que estamos usando un ordenador cuántico. Podemos preparar una superposición de todos los registros clásicos en un solo estado $|\Psi\rangle$.

$$|\Psi\rangle = \frac{1}{\sqrt{2}}[|0\rangle + |1\rangle] \otimes \Lambda \otimes \frac{1}{\sqrt{2}}[|0\rangle + |1\rangle] = \frac{1}{2^{m/2}} \sum_{x=0}^{2^m-1} |x\rangle$$

(este estado se podría conseguir preparando un estado $|0, \dots, 0\rangle$ de longitud m y aplicando una puerta de Hadamard (H) a cada qubit, tal como definimos más adelante) y realizar *una sola operación* sobre este ket para generar todos los resultados en solo paso:

$$|f\rangle = U_f\{|\Psi\rangle \otimes |0\rangle\} = U_f\left\{\frac{1}{2^{m/2}} \sum_{x=0}^{2^m-1} |x\rangle \otimes |0\rangle\right\} = \frac{1}{2^{m/2}} \sum_{x=0}^{2^m-1} |x\rangle \otimes |f(x)\rangle$$

(hemos usado la notación de x formada por la representación decimal de las cadenas de m bits).

Ahora nos encontramos con el problema de extraer la información codificada en el ket $|f\rangle$. Si medimos sobre este ket, obtendremos cualquiera de los resultados posibles con la misma probabilidad ($1/2^m$), colapsando el estado y obteniendo un solo resultado de todos los 2^m valores. Sin embargo existen otras formas más sutiles de obtener información acerca de alguna propiedad global de los valores de $f(x)$, por ejemplo de su periodicidad, de las cuales podremos extraer más información.

2.- Interferencia:

El proceso de computación cuántica podría representarse mediante un diagrama de árbol donde todas sus ramas (a diferencia del caso clásico) se producirían a la vez y estarían caracterizadas por números complejos (amplitudes de probabilidad), cuyos cuadrados son los que nos dan la probabilidad de que al medir sobre el estado final de la MTC, obtengamos un cierto estado concreto al final de una rama. La posibilidad de usar superposiciones coherentes para la representación de la información permitiría, por ejemplo, que si una determinada configuración final de una MTC puede alcanzarse a través de dos caminos con amplitudes de probabilidad α y $-\alpha$, la probabilidad final de alcanzar dicha configuración es $|\alpha - \alpha|^2 = 0$. Es decir, que el resultado de la computación cuántica puede surgir de una adecuada interferencia entre los distintos caminos posibles. De esta forma se pueden codificar varios datos de un problema y tratarlos de forma simultánea y, provocando su interferencia, hacer que algunos de ellos tengan una probabilidad grande, mientras que otros desaparezcan.



PUERTAS Y CIRCUITOS CUÁNTICOS

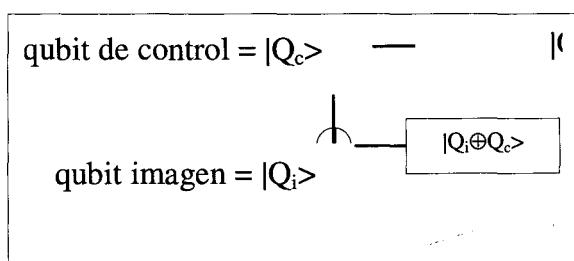
Análogamente a los ordenadores clásicos, podemos representar la evolución de los cuánticos mediante circuitos formados por puertas cuánticas que realizan las operaciones sobre los qubits. En 1989 Deutsch describió los circuitos cuánticos como formados por puertas cuánticas conectadas mediante hilos, demostrando que existía una puerta cuántica universal⁵ y reversible análoga a la de Toffoli clásica. Una forma de obtener puertas cuánticas es la cuantización de las puertas clásicas, que pasa por reinterpretar los bits como qubits. El propósito de los hilos es transmitir estados cuánticos de una a otra puerta y su forma concreta dependerá de las realizaciones tecnológicas concretas de los qubits.

Una puerta típicamente cuántica es la descrita por la matriz de generación de superposiciones del tipo:

$$U_H \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Es la puerta de Hadamard, que genera una rotación o cambio de base. Una puerta de gran importancia (en realidad la de mayor importancia) en los ordenadores cuánticos es la versión cuántica de la CNOT clásica. Su representación matricial es:

$$U_{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$



expresada en la base de computación $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$. En la representación de la puerta CNOT cuántica, el primer qubit ($|Q_c\rangle$) es el de control, mientras que el segundo es el imagen ($|Q_i\rangle$). La puerta ejecuta una operación que es la suma mod 2 ((\oplus)) de ambos qubits. Esta puerta, junto con otras puertas que afectan a un sólo qubit, forman un conjunto universal. De ahí su importancia. Con su ayuda se puede construir

un *programa cuántico*, a través de un circuito cuántico, junto con algún tipo de *puertas de medida*, todos ellos unidos mediante "cables". En este sentido podemos decir que la potencia de la computación cuántica no reside en la rapidez de aplicación de las puertas cuánticas, sino más bien en que debe de usarse un número exponencialmente menor que las necesarias en el caso clásico para realizar la misma tarea.

ALGORITMOS CUÁNTICOS

De momento no se ha demostrado que un ordenador cuántico pueda hacer cosas que uno clásico no pudiera hacer con recursos suficientemente grandes. El problema clásico puede estar encerrado en la frase "con recursos suficientemente grandes", que quizás en la realidad hagan al proceso de cálculo totalmente inviable. El problema consiste en construir algoritmos que aprovechen las características cuánticas *para cambiar la clase de complejidad* de un problema tratado clásicamente. A pesar del esfuerzo que se ha dedicado a la obtención de algoritmos que aprovechen el comportamiento cuántico, en la actualidad su número es reducido. Para programar con ordenadores cuánticos se requieren algunas técnicas nuevas. Dos técnicas básicas son: extracción de una propiedad *global* de una función y los métodos de *amplificación de las amplitudes* para aumentar la probabilidad de sucesos deseables.

Ya se ha mencionado que, aunque mediante apropiadas combinaciones lineales es posible manejar un número exponencial de estados, ello no supone que esta información esté disponible. En realidad, debido a que para acceder a esa información debemos medir sobre el estado colapsándolo, la información se pierde casi en su totalidad. Para aprovechar los aspectos cuánticos, debemos combinar la posibilidad del *parallelismo cuántico* con la *interferencia*. Tal posibilidad permite aprovechar la interferencia destructiva para cancelar los términos no deseables y, por otro lado la interferencia constructiva aumenta los términos deseables. De esta forma al medir obtendremos resultados deseables con mayor probabilidad.

A continuación indicamos algunos de los algoritmos cuánticos existentes:

• Generación de números aleatorios

La necesidad de disponer de secuencias de números aleatorios está ampliamente extendida en campos como criptografía, algoritmos aleatorios (Monte Carlo), simulaciones de evolución física de sistemas, etc. Los ordenadores clásicos sólo pueden calcular funciones, y por tanto cualquier secuencia de números resultante no es completamente aleatoria.

La MC, sin embargo, está fundamentada en leyes indeterministas. Esta indeterminación es básica, a diferencia de la imposibilidad de predicción clásica debida a una incompleta especificación de las condiciones iniciales del problema (caos determinista). Por ejemplo, partiendo de un qubit $|0\rangle$ y aplicando una puerta de Hadamard, obtenemos el qubit $\{ |0\rangle + |1\rangle \} / \sqrt{2}$. Si realizamos una medida, colapsamos el vector de estado del qubit en los estados $|0\rangle$ o $|1\rangle$ con un 50% de probabilidad cada uno. Este método podría ser ampliado (por lo menos en principio) a la generación de números aleatorios entre 0 y 2^{N-1} , sin mas que preparar una superposición de todos los estados de un sistema de N qubits. La medida del vector de estado causaría su colapso aleatorio en uno de sus términos. Trasladando de notación binaria a decimal el qubit, tendríamos el número aleatorio requerido.

• Algoritmo de Deutsch

El problema de Deutsch-Jozsa⁶ fue el primer ejemplo de problema que podía resolverse exponencialmente más rápido en un ordenador cuántico que en una MT clásica. Consideremos un conjunto de funciones booleanas del tipo $f: \{0,1\} \rightarrow \{0,1\}$, en concreto hay cuatro de ellas: dos constantes $f(0)=f(1)=0$ y $f(0)=f(1)=1$ y otras dos balanceadas, $f(0)=0, f(1)=1$ y $f(0)=1, f(1)=0$. Este es el problema que surge si deseamos averiguar si una moneda es falsa (con dos caras o dos cruces) o auténtica (con una cara y una cruz). En realidad no estamos interesados en saber los valores concretos de las funciones, sino únicamente en una característica *global* de la función; averiguar si la función f es constante o balanceada. Desde un punto de vista clásico tenemos que hacer al menos *dos* cálculos de la función (necesariamente) para averiguarlo. Sin embargo, la información obtenida de si la función es constante o balanceada corresponde a un *solo bit*, luego deberíamos ser capaces de obtener el resultado en un *solo cálculo*. Esto podemos conseguirlo mediante un algoritmo cuántico.

El método sería el siguiente. Preparamos dos estados $|0\rangle$ y $|1\rangle$, los rotamos mediante una transformación de Hadamard, realizamos el cálculo de la función mediante una puerta habitual U_f (que dado que $f: \{0,1\} \times \{0,1\} \rightarrow \{0,1\}$, se trata de una *f-controlled-not*):

$$(|0\rangle + |1\rangle)(|0\rangle - |1\rangle) \equiv |00\rangle + |10\rangle - |01\rangle - |11\rangle \rightarrow U_f \rightarrow$$

$$U_f |x, y\rangle = |x, y \oplus f(x)\rangle$$

$$|0,0 \oplus f(0)\rangle + |1,0 \oplus f(1)\rangle - |0,1 \oplus f(0)\rangle - |1,1 \oplus f(1)\rangle$$

y volvemos a rotar el primer qubit:

$$|0\rangle \{ |f(0)\rangle + |f(1)\rangle - |1 \oplus f(0)\rangle - |1 \oplus f(1)\rangle \} + \\ |1\rangle \{ |f(0)\rangle - |f(1)\rangle - |1 \oplus f(0)\rangle + |1 \oplus f(1)\rangle \}$$

Después de esto, medimos el primer qubit, y si obtenemos $|0\rangle$, la función es constante, y si obtenemos $|1\rangle$ la función es balanceada.

Por tanto con una sola medida hemos conseguido el objetivo. Este fue el primer algoritmo que mostró este comportamiento.

• Periodicidad de una función

Ya hemos indicado que si medimos sobre una superposición de valores como los indicados en $|f\rangle$, colapsamos el estado y perdemos el resto de la información. Pero de la misma forma que sucede en un experimento de interferencia, el estado final tiene información de todos sus componentes. Este tipo de superposiciones alberga cierto tipo de información acerca de las propiedades globales del estado, como es la periodicidad de la función. El cálculo del periodo de una función es complejo, involucrando un tipo de transformación que se ha convertido en una herramienta fundamental: la Transformada Discreta de Fourier Cuántica. Esta TDFC se define como la operación unitaria y reversible UTDFC en q ($0 \leq q \leq 2^N$) dimensiones:

$$U_{TDFT} |x\rangle = \frac{1}{q^{1/2}} \sum_{k=0}^{q-1} e^{i2\pi kx/q} |k\rangle$$

Aplicando esta TDFC en determinado momento del desarrollo del algoritmo, se consigue un proceso de interferencia que da lugar a que ciertos términos indeseables de la superposición coherente desaparezcan, mientras que otros deseables se potencien.

• Algoritmo de Shor

Sabemos que mientras el algoritmo de multiplicación es muy rápido, el mejor algoritmo inverso, es decir la factorización, es muy lento. En realidad, clásicamente este último está dentro de la clase de complejidad NP de algoritmos no tratables. El número más grande que se ha factorizado hasta hoy tiene 129 cifras y para hallar sus factores fue necesario el concurso de unos 1600 ordenadores en todo el mundo trabajando sin parar durante unos 8 meses. El crecimiento de los recursos necesarios para la factorización es exponencial, sin embargo, la cantidad de términos que podíamos mantener en una superposición cuántica es también exponencial. Esta es la razón de que algunos algoritmos cuánticos puedan transformar problemas de tipo NP (como la factorización clásica) en P, es decir, tratables o polinómicos.

En 1994 Peter Shor⁷ puso a punto el primer algoritmo de interés práctico, ya que logró plantear un



algoritmo eficaz para la factorización, usando los recursos de un ordenador cuántico. La importancia de esta posibilidad radica en que la dificultad de la factorización está en la base de los códigos criptográficos (mediante un ordenador cuántico, romper la clave RSA 140 sería cuestión de segundos) usados ampliamente por ejemplo en transacciones bancarias o en secretos militares. Romper estos códigos significaría acceder a una gran cantidad de información, al mismo tiempo que destrozar estas claves.

El algoritmo se basa en encontrar el periodo de cierto tipo de funciones relacionadas con el número a factorizar. Desgraciadamente, desde un punto de vista clásico no hay un algoritmo que lo calcule de forma eficiente. Para calcular el periodo se usa el algoritmo cuántico indicado anteriormente. De esta forma, el algoritmo de Shor, es un híbrido entre el cálculo cuántico del periodo de una función y el uso de algoritmos clásicos (en concreto para calcular el m.c.d.) eficientes. Este algoritmo usa $O((\log N)^3)$ pasos⁸, lo que demuestra *cómo este algoritmo cuántico es capaz de cambiar la clase de complejidad clásica de la factorización de NP a P*. A pesar de todo nadie ha demostrado todavía que no exista un algoritmo clásico para la factorización que calcule con eficiencia polinómica.

Hughes ha analizado las previsiones de factorización de números, comparando los resultados de un conjunto de 1000 estaciones de trabajo con los de un ordenador cuántico.

Número de bits	1024	2048	4096
año 2006	10^5 años	$5 \cdot 10^{15}$ años	$3 \cdot 10^{29}$ años
año 2024	38 años	10^{12} años	$7 \cdot 10^{25}$ años
año 2042	3 días	$3 \cdot 10^8$ años	$2 \cdot 10^{22}$ años
En un ordenador cuántico			
número de qubits	5124	10244	20484
número de puertas	$3 \cdot 10^9$	$2 \cdot 10^{11}$	$2 \cdot 10^{12}$
tiempo	4.5 minutos	36 minutos	4.8 horas

Este análisis evidencia la potencia de cálculo de un ordenador cuántico.

• Algoritmos de búsqueda

Un tipo interesante de problemas son los de búsqueda. El algoritmo de Grover trata este problema. Para ello usa un algoritmo cuántico que aprovecha la posibilidad de superposición coherente⁹. Por ejemplo, consideremos que tenemos una lista con N datos (que puede ser una lista de 106 nombres de un listín telefónico, ordenado por orden alfabético), nos planteamos encontrar un elemento concreto (es decir, dado un teléfono, encontrar su dueño). Clásicamente para encontrar un dato concreto con una probabilidad 1/2, deberíamos buscar, en promedio, unos $N/2$ ($5 \cdot 10^5$

búsquedas) elementos hasta dar con el buscado, por tanto el algoritmo escala como $N=2L$, y por tanto no es tratable. El algoritmo de Grover¹⁰ demuestra que para una búsqueda en una base de datos sin estructura, se necesitan sólo $O(N^{1/2})$ pasos temporales (con una probabilidad acotada), con lo que sigue siendo no tratable (no se cambia la clase de complejidad), sin embargo se aumenta su eficiencia, lo que es importante cuando se trata una gran cantidad de datos. En el caso del listín se necesitarían sólo ¡1000! búsquedas.

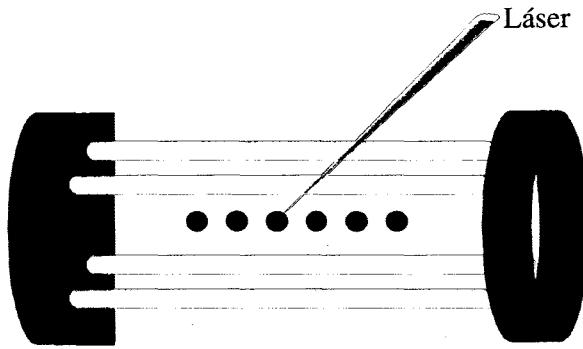
Ciertos problemas se pueden plantear como problemas de búsqueda. Por ejemplo el problema de ordenar un conjunto de números por orden creciente, búsqueda de extremos, etc. Brassard ha apuntado una posible aplicación en criptografía donde se usan claves de 56 bits. La búsqueda de la clave correcta entre las 2^{56} (~7.1016) posibles tardaría más de 2000 años si se realizara clásicamente (a una velocidad de 10^6 búsquedas por segundo), mientras que mediante el algoritmo de Grover tardaría menos de cinco minutos.

PROCESOS DE CONTROL DE ERRORES

Quizás uno de mayores problemas a la hora de construir un ordenador cuántico es el control de los posibles errores, errores que provienen de la inexorable interacción del ordenador con el entorno, proceso denominado *decoherencia*. Desde un punto de vista clásico, los errores se corrigen mediante procesos que implican cierta disipación, siendo estos métodos inviables en el caso cuántico, ya que esta disipación implicaría una pérdida irremediable de coherencia. Este hecho, unido a que además de los errores de bit clásicos, también aparecen *nuevos errores*, típicamente cuánticos, como los relacionados con la variación de las fases relativas en las superposiciones coherentes, hizo que durante algún tiempo se pensara que no podían existir métodos para el control de errores cuánticos.

...la tecnología actual que implementa puertas y circuitos cuánticos está sólo en su infancia

Afortunadamente, dos recientes contribuciones (ahora ya clásicas) debidas a Shor y a Steane¹¹, han cambiado este panorama. Estos autores han mostrado cómo es posible contener los errores mediante códigos cuánticos correctores de errores. Tales códigos detectan y corren los errores usando sofisticadas técnicas cuánticas que involucran un tercer estado de apoyo, hacia donde se copia la información sólo del error.



Iones confinados en una trampa de Paul lineal.
Se perfila como el elemento básico de los futuros ordenadores cuánticos

Midiendo este estado de apoyo, podemos saber el error *sin colapsar el estado*. Aplicando ahora la transformación inversa al error, al estado con error, logramos su corrección.

HARDWARE CUÁNTICO

Debido al problema de la creación, control y corrección de errores en las superposiciones coherentes de estados cuánticos, la tecnología actual que implementa puertas y circuitos cuánticos está sólo en su infancia. Aunque se han construido puertas CNOT de dos qubits experimentales, y se han usado algunas técnicas simples de corrección de errores, no se espera que antes de unos 30 años existan ordenadores cuánticos que hagan tareas de cierta importancia.

El progreso en el control y manipulación de los qubits intenta usar todo tipo de técnicas y sistemas: fotones en cavidades, espines controlados por RMN, electrones en puntos cuánticos, etc. Quizás una de las tecnologías más prometedoras consiste en confinar iones ultrafríos en trampas de Paul lineales. Las operaciones de control de los qubits se realiza dirigiendo haces láser a cada uno de los iones. Mediante un procedimiento similar se ha implementado la primera puerta CNOT cuántica.

CONCLUSIÓN

Estamos ante otra de esas revoluciones interdisciplinarias que producen gran cantidad de nuevas relaciones entre campos inicialmente sin conexión. La revolución cuántica está alcanzando también a la información, no sólo en sus métodos de procesado, sino en su propia concepción. La posibilidad de construir ordenadores cuánticos permitirá procesos relacionados con el tratamiento de la información, hasta ahora insospechados, además de poner de manifiesto ciertos comportamientos cuánticos fundamentales, hasta ahora sólo plasmados en los libros de texto. Quizás sea un buen momento de participar en este desarrollo.

- [1] Bennett, C. H. & Landauer, R; «The fundamental physical limits of computation», Scientific American 1985, July 38.
- [2] Feynman, R.; «Quantum mechanical computers», Found. Phys. 16 507 (1986).
- [3] Deutsch, D.; «Quantum theory, the Church-Turing principle and the universal quantum computer», Proc. R. Soc. London, A400 97 (1985).
- [4] Schumacher, B; «Quantum coding», Phys. Rev. A 51 2738, (1995)
- [5] D. Deutsch, «Quantum computational networks», Proc. R. Soc. London A 425, 73 (1989).
- [6] Deutsch,D. & Jozsa, R; «Rapid Solutions of problems by quantum computation», Proc. of the Roc. Soc., A439 553 (1992).
- [7] Shor, P.W.; Proceedings of the 35th Annual Symposium on Foundations of Computer Science (IEEE Computer Society, Los Alamitos CA 1994) p124. Shor, P.; «Polynomial-time algorithms for prime factorisation and discrete logarithms on a quantum computer», Proc. 35th Annual Symp. on Foundations of Computer Science, Santa Fe, IEEE Computer Society Press.
- [8] Beckman D.; Chari A.; Devabhaktuni S. & Preskill J.; «Efficient networks for quantum factoring», Phys. Rev. A 54 1034 (1996).
- [9] Grover, L.K.; «The advantages of superposition», Science 280 228, abril 1998. 10 Grover L.K., «A fast quantum mechanical algorithm for database search», Proceedings of the 28th Annual ACM Symposium on Theory of Computing , p212, Philadelphia 1996. 11 Steane, AM; «Multiple particle interference and quantum error correction», Proc. Roy. Soc. Lond. A 452 2551 (1996).





LA BÚSQUEDA DE INTELIGENCIA EXTRATERRESTRE: S.E.T.I.

Marc Caballero Gómez

*Estudiante de la ETSETB y Miembro de la Rama de Estudiantes del IEEE Barcelona
mave25@casal.upc.es*

INTRODUCCIÓN

Actualmente la humanidad se encuentra en un momento de su evolución que podríamos definir como la adolescencia tecnológica; en ella los avances tecnológicos se suceden con una celeridad asombrosa que nos sorprende con nuevos descubrimientos y aplicaciones a diario y nos conducen hacia una madurez todavía inimaginable y que, según parece, tardaremos aún bastante en alcanzar. Aunque estos avances permiten satisfacer muchos de nuestros deseos y necesidades, la humanidad aún posee muchas inquietudes y deficiencias por resolver; una de esas inquietudes, aunque algunos no quieran admitirla, es la de conocer si estamos solos en la inmensidad del universo. Parece lógico considerar que con toda la tecnología de que disponemos y los avances que la favorecen, nos dedicuemos a intentar conocer mejor el universo que nos rodea, un universo tan inmenso y antiguo que nuestra existencia individual ni siquiera sería tan relevante como la de una chispa, un universo que contiene aún tantos secretos y sorpresas que resulta el campo de trabajo perfecto para el enorme afán de investigación que caracteriza la comunidad científica actual. Una de las líneas de investigación se basa en considerar la posibilidad de que en algún confín adecuado del universo se haya podido desarrollar algún tipo de forma de vida; no creo que sea una idea muy descabellada, consideren sino una frase que escuché una vez y que afirmaba que «si estamos solos en el universo, vaya desperdicio de espacio».

Este tema aún parece un tabú en algunos círculos científicos pero, poco a poco, se está convirtiendo en un referente a considerar y en una base sólida para la evolución de ciertas ramas tecnológicas con aplicación en ciencias tan respetables como la astronomía. Esta especie de aversión que aún produce esta línea de investigación se debe básicamente al hecho que fácilmente se tiende a relacionarla con la búsqueda de hombrecillos verdes, platillos volantes y el platónico deseo de provocar un encuentro en la tercera fase; nada más lejos de la realidad. Lo cierto es que en los últimos años hemos alcanzado un nivel tecnológico suficientemente importante como para plantearnos la posibilidad de buscar señales de la posible existencia de civilizaciones más o menos avanzadas en los remotos confines del universo; simplemente se trata de rastrear el espacio exterior en busca de sistemas estelares cuya composición hubiese permitido la generación de los procesos físicos y químicos necesarios para la existencia

de vida, tal y como sucedió en nuestro sistema solar; una búsqueda que, además, contribuye plenamente en avances astronómicos sobre la caracterización del universo. También se procede, mediante el uso de radiotelescopios y otros sistemas de radiodetección extremadamente sensibles, intentando captar del espacio señales de radio cuya procedencia no pueda ser relacionada con fenómenos naturales o con señales de procedencia terrestre; la captación e identificación de estas señales indicaría la existencia de civilizaciones suficientemente avanzadas como para utilizar las comunicaciones vía radio, sin que esto signifique que sean más avanzadas que la nuestra pues debemos tener en cuenta que la banda de radio ocupa una gran parte del espectro electromagnético y podría haber sido descubierta por cualquier civilización más o menos avanzada tal y como hicimos nosotros. Creo, pues, que queda claro que la búsqueda de inteligencia extraterrestre se aleja claramente de la idea de una inminente invasión de terroríficos hombrecillos verdes procedentes de «Raticulín», es sólo un tópico que espero que el tiempo acabe de eliminar y así no resulte una mancha negra en el expediente de espléndidos científicos.

Siguiendo esta línea de investigación debemos, primero, ser conscientes de las pocas probabilidades de éxito que tenemos, para ello nada mejor que referirnos a la fórmula que propuso Frank Drake para estimar el número de civilizaciones extraterrestres que pueden existir en la galaxia:

$$N = R \cdot f \cdot n \cdot l \cdot i \cdot c \cdot L$$

N = Número de civilizaciones tecnológicamente evolucionadas en la galaxia.

R = Número medio de estrellas presentes en la galaxia.

f = Fracción de estas estrellas que pueden tener un sistema planetario.

n = Número de planetas en el interior de estos sistemas que podrían permitir la evolución de la vida.

l = Número de planetas donde actualmente se desarrolla vida.

i = Número de planetas donde la vida es inteligente.

c = Número de planetas en los que se han desarrollado tecnologías aptas para la comunicación.

L = Vida media de tales civilizaciones.

A tenor de esta expresión resulta realmente difícil considerar la posibilidad de hallar alguna, pero los últimos avances en astronomía han favorecido el descubrimiento de planetas orbitando alrededor de lejanas estrellas e incluso de un sistema planetario. Hasta hace poco la existencia de planetas extrasolares era sólo una creencia

pero en los últimos cinco años se han descubierto un total de veinte planetas fuera de nuestro sistema solar, sobre todo gracias a la labor del astrónomo estadounidense Geoffrey Marcy que halló catorce de ellos. Todos estos hallazgos hacen, pues, aumentar las posibilidades y, por tanto, suponen una gran fuente de motivación para los incansables investigadores del SETI. No es para lanzar las campanas al vuelo porque resulta que dieciocho de ellos son de tipo gaseoso y, por tanto, un poco inhóspitos como para albergar vida, al menos vida similar a la que conocemos, mientras que sobre los otros dos, aún siendo de tipo terrestre, no tenemos suficientes datos científicos como para extraer conclusiones; aún así estos descubrimientos resultan un gran avance y un preludio de lo que puede suceder en los próximos años.

Respecto a la identificación de señales de radio procedentes de otras civilizaciones los resultados no son nada gratificantes hasta ahora pero, en parte, es lógico debido a la gran dificultad que supone. Pensemos en lo amplio que resulta el estudio del espectro electromagnético y en que debemos analizarlo para cada una de las estrellas o sistemas que creamos que puedan ser propensos a albergar civilizaciones. Para hacernos una idea de lo que esto significa consideremos el caso particular de nuestra galaxia, la Vía Láctea, en ella debe haber, aproximadamente, unos 250.000 millones de estrellas y, según otro astrónomo, Carl Sagan, alrededor de un millón de ellas poseerían civilizaciones con cierto grado de tecnología; esto significa que menos de una estrella entre 250.000 tendría un planeta o sistema planetario con vida inteligente. Así pues, primero debemos identificar ese "reducido" número de estrellas candidatas y después, realizar el completo análisis de su espectro. Notemos, pues, el enorme esfuerzo que se requiere, pero es que aún así debemos considerar la posibilidad de que esas civilizaciones utilicen sistemas de comunicación totalmente desconocidos para nosotros y que, por tanto, aún buscando en la estrella adecuada, nunca reconocamos señal alguna que indique la presencia de una civilización, en este caso, más avanzada que la nuestra.

EVOLUCIÓN HISTÓRICA

El hecho de analizar el espectro de radio se debe a la suposición que si una civilización avanzada decidiese comunicarse con una menos avanzada la radio es el método más obvio para hacerlo, además, es la mejor técnica de larga distancia que posee nuestra tecnología.

El primer intento serio de escuchar señales de radio procedentes de otras civilizaciones se remonta al año 1959 en el que Frank Drake organizó el proyecto Ozma*; desde las instalaciones del "National Radio Astronomy Observatory" (NRAO) en Green Bank se observaron dos estrellas cercanas, Epsilon Eridani y Tau Ceti, durante algunas semanas; no se obtuvieron resultados favorables pero podemos considerarlo como el inicio del SETI.

Tras esto, a principios de 1970, la omnipresente "National Aeronautics & Space Administration" (NASA) se planteó la posibilidad de iniciar un ambicioso proyecto; desde el "NASA's Ames Research Center" (ARC) en Montain View se realizó un estudio, conocido como "Project Cyclops" y dirigido por Bernard Oliver, sobre la viabilidad del proyecto; lo realizaron un grupo de científicos del "Massachusetts Institute of Technology" (MIT), encabezados por Phillip Morrison, que aprobaron la idea y cuyo informe determinó un análisis sobre la ciencia SETI y sobre los requerimientos tecnológicos que constituyeron la base para trabajos posteriores.

A modo de estudio preliminar, a finales de los 70 los programas SETI se habían establecido en el "Jet Propulsion Laboratory" (JPL) en Pasadena, California, y en el NASA's ARC; ambos programas adoptarían una estrategia común para realizar un SETI a gran escala, Ames examinaría 1.000 estrellas similares al sol mediante una búsqueda selectiva y con capacidad para detectar señales débiles y esporádicas; por su parte, el JPL barrería sistemáticamente todas las direcciones. No fue hasta 1988 cuando después de más de una década de estudios y diseños preliminares, la dirección de la NASA adoptó formalmente esta estrategia y fundó el programa oficial que iniciaría las observaciones cuatro años después. Todo parecía propicio para el comienzo del ambicioso proyecto, incluso la fecha de inicio de las observaciones, pues coincidía con el quinto centenario del descubrimiento de América; pero ni la reputación de los miembros del MIT que realizaron el estudio inicial, ni el prefacio que contenía el informe Cyclops a cargo del Reverendo Theodore Hesburgh, presidente de la Universidad de Notre Dame, consiguieron que el programa tuviera el respaldo constante del congreso. Dicho programa estuvo caracterizado, ya desde su etapa preliminar, por constantes recortes presupuestarios y suspensiones periódicas, que finalmente desembocaron en su definitiva cancelación en septiembre de 1993, sólo un año después de su formalización, gracias, sobre todo, a la intervención del Senador por Nevada Richard Bryan.

Vale la pena remarcar que aparte de la NASA, desde el proyecto Ozma se habían realizado algunos modestos intentos tanto en los Estados Unidos como en Canadá, pero fue en la Unión Soviética donde, durante los 60, hubo cierta actividad destacable. Los soviéticos optaron por ciertas líneas arriesgadas de investigación; en vez de orientar la búsqueda hacia estrellas cercanas optaron por observar grandes porciones de cielo con antenas casi omnidiireccionales; no podemos, pues, darle otro calificativo que el de arriesgadas porque implícitamente estaban considerando la existencia de al menos unas pocas civilizaciones capaces de radiar con una gran potencia.

Paralelamente a la NASA, otras organizaciones y universidades decidieron iniciar proyectos similares y con el mismo objetivo; puede que no dispusieran de la misma

* El nombre de OZMA se debe al lugar exótico, distante y difícil de alcanzar que representa la tierra de Oz del libro de L. Frank Baum.



capacidad tecnológica y económica que la de la NASA pero con el tiempo algunos de ellos consiguieron el merecido reconocimiento de su labor y adquirieron especial relevancia al cancelarse definitivamente el programa oficial.

En 1981, el físico de Harvard y asesor de la NASA Paul Horowitz decidió seguir una de las propuestas sobre soluciones, de coste modesto, que proponía el informe de Morrison acerca del desarrollo de sistemas receptores de radio más sensibles y sistemas computerizados de procesado de señales. En esta línea propuso crear el «Suitcase SETI» que no era más que un receptor computerizado portátil para la búsqueda de señales en 131.000 radiocanales de banda muy estrecha. En esa época el proyecto de la NASA sufría uno de sus comunes recortes pero Horowitz se dirigió con su propuesta hacia la recién creada «Planetary Society» (PS) y decidieron instalar su sistema en las instalaciones donde la PS planeaba empezar su proyecto BETA, esto daba la posibilidad de iniciar un nuevo proyecto a más corto plazo, lo llamaron «Project Sentinel».

Aparte del Sentinel, la PS inició en otoño del 1985 el «8-million-channel project META» (Megachannel Extraterrestrial Assay) que se ocuparía de rastrear el cielo del hemisferio Norte terrestre; cinco años después, en 1990, se iniciaría el META II, un duplicado del proyecto que, desde Buenos Aires (Argentina), se encargaría de ampliar el proyecto con el rastreo del hemisferio Sur. El META escaneaba 8.388.608 canales con una resoluciónpectral de 0.05 Hz y 400 kHz de ancho de banda instantáneo y unos otros 1.048.576 canales redundantes para el control. El sistema, además, corregía la frecuencia observada ante los movimientos respecto a tres encuadres iniciales astronómicos. También ajustaba la frecuencia para compensar la rotación de la Tierra, que genera un cambio característico en la marca Doppler para señales de origen extraterrestre.

A finales de 1994, el META I había cubierto cinco veces el cielo utilizando una longitud de onda de 21 cm.; longitud que corresponde a la banda del Hidrógeno por lo que está más o menos libre de radiación cósmica de fondo y sería, pues, la frecuencia más propensa a ser utilizada como vía de comunicación para civilizaciones extraterrestres. El META I también consideró la idea de uno de los pioneros del SETI, Sebastian von Hoerner, de estudiar el segundo armónico de la frecuencia, a 10.5 cm. Por su parte, el META II había recorrido el cielo del hemisferio sur unas tres veces escaneando sólo en la banda del hidrógeno.

Sobre los resultados globales del META podemos decir que se detectaron algunas docenas de señales intrigantes pero ninguna de ellas se repitió y, por tanto, no se pudo eliminar la interferencia para identificarlas. En concreto podemos decir que a finales de 1994 el META I,

operado por Horowitz desde la universidad de Harvard, había examinado más de 60 trillones de canales diferentes de los que sólo 37 resultaron ser candidatos. En el hemisferio Sur, el META II, dirigido por Raúl Colomb en el Instituto de Radio Astronomía de Buenos Aires, encontró 19 posibles candidatos entre los 16 trillones de canales examinados. Todas estas señales candidatas tenían anchos de banda extremadamente pequeños que no parecía que siguieran la rotación terrestre; por desgracia todas estas señales no pudieron detectarse de nuevo en la reobservación y esto resulta imprescindible para su identificación como señal de procedencia extraterrestre.

La PS también participaría con su apoyo, a partir de 1996, en el proyecto SETI desarrollado en la «University of California» en Berkeley, el SERENDIP («Search for Extraterrestrial Radio Emissions from Nearby Developed Intelligent Populations»). Desde el 15 de abril de 1992, el SERENDIP III, tercera fase del proyecto, había estado operando desde el radiotelescopio de Arecibo, en Puerto Rico, que con sus 305 metros de antena se puede considerar como el mayor del mundo. El programa había analizado más de 200 trillones de canales de datos y había rastreado un 30 % del cielo, abarcando pues, más volumen que todos los anteriores proyectos SETI del mundo combinados. El sistema examinaba 4.2 millones de canales cada 1.7 segundos y se anotaron 200 millones de canales con picos espectrales por encima del ruido de fondo que generaron 400 señales candidatas para su posterior estudio detallado.

Las contribuciones de la PS permitieron a Dan Werthimer, miembro del programa, concluir el diseño e implantación de una nueva máquina, el SERENDIP IV, mucho más potente que su antecesora y que la sustituiría en Arecibo a partir de junio de 1997. Este nuevo sistema mejoraría las capacidades según un factor 40, por lo que permitiría analizar hasta 160 millones de canales cada 1.7 segundos. Desde Arecibo también se optó por analizar la banda del Hidrógeno por las ventajas antes comentadas y, también, por el hecho que la ley internacional prohíbe transmisiones de radio en esta parte del espectro frecuencial, por lo que la posibilidad de interferencia debida a la tecnología terrestre es mínima.

Antes de la intervención de la «Society», en Berkeley ya habían realizado dos proyectos SERENDIP, el I y el II. El primero de ellos se realizó entre 1980 y 1982 con un analizador de espectros con capacidad para el estudio de 100 canales de 1 kHz por canal montado en los radiotelescopios del «Hat Creek Observatory» al norte de California y el «Goldstone Observatory» del desierto del Mojave. El segundo de ellos se realizó entre 1986 y 1988 y era miles de veces más potente que su predecesor; podía analizar 65.000 canales por segundo desde el radiotelescopio del NRAO en West Virginia y, en menor medida, desde otros cuatro telescopios situados alrededor del mundo.

La maquinaria del SERENDIP se complementa con un ordenador capaz de almacenar las señales de interés y sincronizar el telescopio con estas señales para su ulterior y concienzudo estudio. También se utilizan una serie de algoritmos computacionales para eliminar la interferencia humana que claramente se impone sobre la aglomeración de señales captadas y que se conoce como «Radio Frequency Interference»(RFI).



Radiotelescopio de Arecibo en Puerto Rico, con un diámetro de 305 metros.

De las anteriores líneas se deduce la importante labor de colaboración de la Planetary Society con los diferentes proyectos SETI, pero existen también otras organizaciones que colaboran en la búsqueda, y entre ellas destaca el «SETI Institute», un ambicioso proyecto que pretende coordinar los diferentes centros de investigación y programas en un esfuerzo común de detección de señales extraterrestres. El Instituto colabora, por ejemplo, con el SERENDIP antes comentado y, además, inició su propio proyecto en febrero de 1995 denominando al programa «Project Phoenix».

Podemos considerar al Phoenix como el sucesor del proyecto SETI de la NASA; cuando éste fue cancelado los sistemas de procesado de señales TTS («Targeted Search System») que utilizaba el programa oficial estaban siendo mejorados, así que, lo que decidieron los miembros del Instituto fue terminar esas mejoras que habían quedado a medias y reestructurar el sistema para adecuarlo a las nuevas investigaciones que pretendían realizar e intentar tenerlo todo apunto para seguir los planes de observación

que tenía el programa de la NASA, esto suponía tenerlo todo listo en diciembre de 1994; se retrasaron un poco pero no demasiado, pues en febrero de 1995 se iniciaron las observaciones desde la antena de 64 metros del radiotelescopio de Parkes en New South Wales, Australia. Se escrutaron unas 200 estrellas de características solares en una ventana de microondas que iba de los 1.2 a los 3.0 GHz. La señal captada resultaba ser la primera de las dos necesarias para la aplicación de un sistema de confirmación en tiempo real mediante la utilización de otra antena con un sistema de procesado de señal independiente; la antena Mopra, de 22 metros, situada a unos 200 Km. al norte de Parkes, comprobaba cada una de las señales, propuestas como candidatas por la antena primaria, mediante el sistema FUDD («Follow-Up Detection Device») que realizaba un filtrado adaptado a la señal, de esta manera también se conseguía compensar la poca sensibilidad del telescopio utilizado.

Las observaciones en Australia concluyeron en junio de 1995, entonces el sistema de recepción se trasladó a California para realizar ciertas mejoras. A principios de septiembre de 1996 se instaló el sistema en el NRAO; utilizando la antena de 140 pies del complejo se realizaron observaciones hasta abril de 1998 utilizando el telescopio, más o menos, la mitad del tiempo disponible. Debemos tener en cuenta que la mayoría de programas deben compartir las antenas con otros investigadores por lo que sólo disponen de ellas a tiempo parcial; esto alarga claramente el tiempo real necesario para realizar las observaciones planeadas por el programa y, consecuentemente, el tiempo de obtención de resultados, pero mientras los recursos sean limitados esta es la mejor forma de que los diferentes campos de investigación implicados realicen sus estudios.

A mediados de 1998 el proyecto Phoenix subió un peldaño más trasladándose al radiotelescopio de 305 metros de Arecibo; allí, con una antena recién mejorada para dar mayor sensibilidad, dispusieron de un total de 2.600 horas de observación en sesiones de dos o tres semanas al año. En este enclave la antena de confirmación que se utilizó fue la del telescopio Lovell del «Jodrell Bank Observatory» de Inglaterra. La gran distancia entre ambas y la diferencia de latitudes la convertían en un inmejorable filtro frente a la RFI.

Como en la mayoría de los demás programas no se rastreaba todo el cielo sino que se focalizaba el estudio en los alrededores de estrellas cercanas y parecidas al sol; el hecho de centrarse en las de características solares se debe a que se cree que son las más propensas a tener planetas de suficiente antigüedad como para albergar vida. Se incluirían también las estrellas de las que se conocía la existencia de planetas orbitándolas; en total se seleccionaron unas 1.000 estrellas para su estudio, todas ellas comprendidas en un margen de 150 a 200 años luz de distancia. Se buscaban señales comprendidas entre los 1.000 y 3.000 MHz que supusieran una simple marca en el espectro, o

sea, señales de banda muy estrecha, característica que representaría la firma de una transmisión inteligente. Se analizan canales de 1Hz de ancho de banda, o sea, billones de canales por cada estrella seleccionada. Para ello deben monitorizarse simultáneamente millones de ellos por lo que la «escucha» debe realizarse por ordenador, los astrónomos simplemente realizan las decisiones críticas sobre las señales intrigantes que detecta el análisis informático. No debe pasarnos inadvertida la enorme capacidad computacional necesaria para procesar todos los datos, necesidad que aumenta con el tiempo y que se está convirtiendo en un problema.

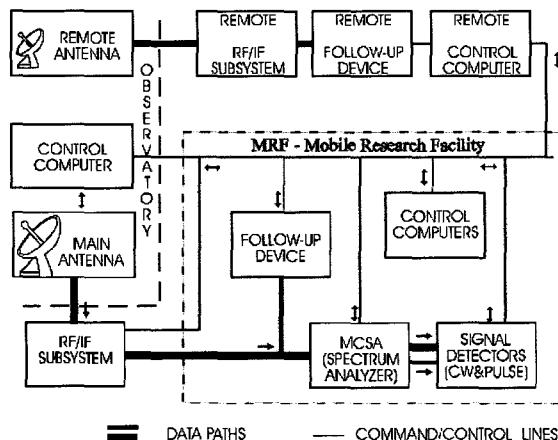
THE PHOENIX TARGET SEARCH SYSTEM

A finales de 1999 el Phoenix había examinado la mitad de las estrellas que configuran su lista de candidatas sin haber encontrado señales claramente extraterrestres.

Una vez analizada la evolución del SETI a lo largo de estos últimos años puede resultar interesante realizar un estudio detallado de alguno de los sistemas de recepción que se utilizan, en concreto nos referiremos al que utiliza el proyecto Phoenix debido a su actual relevancia entre los demás programas y a su relación con el sistema utilizado inicialmente por la NASA.

Como ya hemos mencionado anteriormente, el proyecto Phoenix utiliza un sistema de procesado de señales TSS, se trata de un sistema SETI portátil que se utiliza conjuntamente con los radiotelescopios existentes para obtener observaciones de alta sensibilidad. Sus características están totalmente determinadas por los requerimientos observacionales, en concreto se ocupa de rastrear señales artificiales con un ancho de banda realmente estrecho, menor de 300 Hz, en una ventana de microondas desde 1 GHz a 3 GHz, observando cada una de las bandas frecuenciales durante, como mínimo, 300 segundos; para ello utiliza los mayores telescopios disponibles para conseguir, así, mayor sensibilidad. Se observan unas 1.000 estrellas de características solares comprendidas en una

PROJECT PHOENIX -- TARGETED SEARCH SYSTEM



Flujo de datos y de control que definen el sistema TSS.

distancia de 150 años luz. También se requiere que el procesado de las señales se realice casi en tiempo real, permitiendo así que las señales consideradas de interés puedan ser testadas inmediatamente. El sistema resultante se ha automatizado al máximo para minimizar la interacción del operador y, por tanto, incrementar la calidad y uniformidad de la búsqueda.

El TSS se compone de una serie de subsistemas, cada uno responsable de uno de los aspectos del procesado de la señal captada, controlados por un sofisticado software que facilita la automatización del proceso.

Actualmente no hay observatorios que proporcionen una cobertura continua del total del ancho requerido por el sistema (de 1 GHz a 3 GHz), por lo que el TSS incluye su propio subsistema de recepción . El «Radio Frequency/Intermediate Frequency Subsystem» incorpora dos amplificadores HEMT criogénicamente refrigerados que cubren los rangos de 1.0 a 1.8 GHz y de 1.8 a 3.0 GHz, con bocinas de alimentación dieléctricamente separadas para cada banda. El sistema de recepción proporciona un ancho de banda instantáneo de 300 MHz para cada una de las dos polarizaciones circulares y una temperatura de sistema de 25 K o menor. En este subsistema también se realiza a una conversión a frecuencia intermedia y se selecciona un ancho de 20 MHz de la señal para pasarla a banda base, digitalizarla muestreando en cuadratura y enviarlo definitivamente al analizador de espectros multicanal.

El «MultiChannel Spectrum Analyzer» es el espectómetro del TSS y se compone dos unidades de polarización dual, cada una de las cuales divide 10 de los 20 MHz entregados en decenas de millones de simultáneos canales estrechos que pueden ser analizados en busca de señales por unos ordenadores diseñados a tal fin. Esta función básica podría conseguirse con una simple transformación de Fourier, pero el MCSA debe utilizar una aproximación de filtro polifásico para atender otros posibles requerimientos.

Para minimizar el efecto de la RFI, el MCSA dispone de dos capas de filtros digitales paso banda (BPF) seguidos de una transformación de Fourier. Cada uno de los filtros frontales de cada capa divide el ancho suministrado en aproximadamente 100 bandas menores, con las bandas adyacentes aisladas en más de 100 dB. Este alto grado de rechazo fuera de banda evita que señales potentes puedan contaminar completamente la banda observada. A las muestras de salida del segundo BPF de la capa se les aplica la transformada de Fourier para conseguir canales frecuenciales con resoluciones próximas a 1 Hz; una resolución realmente alta pero necesaria para la detección de señales continuas (CW) como, por ejemplo, portadoras.

Para poder ajustar señales pulsadas, el MCSA efectúa simultáneamente múltiples FFTs para dividir la banda

en canales de tres anchos de banda diferentes a elegir entre seis disponibles: 28.740.096 canales con ancho de 1 Hz, 14.370.048 de 2 Hz, 7.185.024 de 4 Hz, 4.105.728 de 7 Hz, 2.052.864 de 14 Hz o 1.026.432 de 28 Hz. Esto proporciona cierta sensibilidad a pulsos de una duración que oscila entre los 0.02 y los 1.5 segundos. Se debe tener en cuenta también que es bastante improbable que los pulsos estén sincronizados con los relojes terrestres que imponen el periodo de muestreo en el MCSA, por lo que se opta por la superposición en tiempo, y al 50 %, de sucesivos espectros.

En el MCSA los canales también se solapan levemente en frecuencia para conseguir una respuesta casi óptima tanto para CW como para señales pulsadas, siempre y cuando esas señales se mantengan estables en un mismo canal o se desplacen por el espectro, como mucho, hacia un canal adyacente.

El montaje del equipo para el MCSA requiere dos armarios de equipamiento estandarizados para situar 8 placas computacionales comerciales, 72 sistemas circuitales de 6 tipos diferentes y un total de 384 DSP's diseñados especialmente para el sistema. Cabe también resaltar que el sistema logra una tasa de operación de aproximadamente unos 75 GFLOPS (75 billones de operaciones de coma flotante por segundo).

Los dos tipos de señales, tanto las CW como las pulsadas, requieren unos sistemas de procesado totalmente diferentes, por ello el «Signal Detection Subsystem» (SDS) se divide en dos sistemas independientes de detección. Estos operan en paralelo sobre los dos flujos de salida del MCSA y remiten los resultados básicos de la detección a un subsistema encargado de su evaluación, el SCS que después analizaremos.

El detector de señal continua (CW) analiza la salida del MCSA de mayor resolución para cada una de las polarizaciones separadamente, en ella busca señales con componente continua con una relación señal a ruido (SNR) por canal mayor o igual a 0.25. Para frecuencias de señal superiores a 2 GHz se utiliza una resolución de 1 Hz, mientras que por debajo, la resolución a 2 Hz se procesa para compensar el amplio margen de traslación Doppler que presentan las señales. Trabajando con la resolución de 1Hz, el detector de CW recibe 80 millones de medidas de potenciapectral por segundo (28.74 millones de canales por polarización cada 0.714 segundos). Si consideramos que el producto temporal de las muestras es unitario, la tasa de datos resulta ser la misma tanto para la resolución de 2 Hz como para cualquier otra. El algoritmo utilizado para realizar una suma eficiente de las potencias detectadas en cada canal a lo largo de toda la deriva potencial que sufre el recorrido de las señales en el dominio de frecuencia-tiempo, requiere que todos los datos deban almacenarse durante la observación; esto requiere la utilización de un Hardware de grabación que se implementa de la forma más económica posible mediante hileras de discos de grabación (RAID). Una vez

completada la observación los datos almacenados se procesan mediante un sistema formado por cuatro placas circuitales que realizan 3.2 billones de sumas por segundo; de mientras, un segundo grupo de RAID's almacena los datos de una nueva observación. Tras completar el proceso, el detector CW indicará al sistema si en alguno de esos recorridos se ha sobrepasado un umbral preestablecido de potencia.

Considerando que las señales pulsadas con una potencia media similar a la de las CW van a ser relativamente fuertes sólo cuando estén a nivel alto, los datos entregados por el MCSA deben ser cribados previamente al procesado. Únicamente los canales con unos niveles de potencia que superen un cierto umbral de decisión se entregarán al detector de señales pulsadas del SDS. De esta forma se puede establecer un umbral tal que en situaciones de ausencia de señal alguna, o sea única presencia de ruido, sólo un canal entre 10^5 pasará al detector. Esto supone que se dispone de una reducida serie de datos para procesar, por lo que podemos optar por representarlos mediante una matriz de dispersión en el dominio frecuencia-tiempo. El conjunto de datos de las dos polarizaciones estudiadas se almacena en un disco de 1GB de capacidad y un procesador i860 se encarga de buscar, entre los datos, series de tres pulsos regularmente espaciados. Cualquier «tripleta» de pulsos con una potencia global que excede del umbral estático predefinido se remite, al igual que para las señales CW, al SCS.

Este «System Control Subsystem» (SCS) está compuesto por los ordenadores y el software necesarios para realizar la configuración, monitorización y control de todos los subsistemas y observaciones del TSS. Un par de estaciones HP (HP9000/755 y /735) con arquitectura cliente/servidor configuran el SCS para el TSS a 20 MHz. Una estación 9000/715 controla el equipamiento de la estación remota. El software está basado en concurrentes procesos comunicativos, cada uno de los cuales implementa una función específica del sistema; son muchas las capacidades y funciones disponibles del SCS.

El observador puede programar de forma interactiva una serie de observaciones mediante un interfaz gráfico. Puede, por tanto, configurar los subsistemas seleccionando las resoluciones para el MCSA, los umbrales de decisión, eligiendo subseries de datos para visualizar o almacenar, o, simplemente, aceptando la configuración por defecto del sistema. Cuando se está a punto de iniciar una de las observaciones programadas, el SCS establece conexión con el ordenador de control del observatorio para que éste proceda a orientar el telescopio según la posición de la estrella a observar, y se mantiene a la espera de la confirmación del rastreo de ésta. Cuando la recibe, indica a los diferentes subsistemas que inicien la observación.

Mientras se lleva a cabo, el observador puede cambiar interactivamente el margen de frecuencias y la resolución de los datos monitorizados. Cuando se reciben



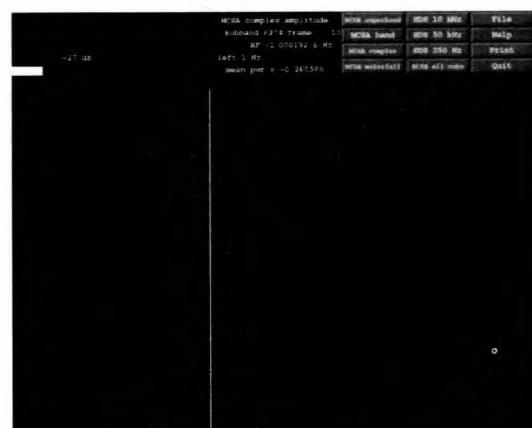
señales del SDS, el SCS compara las señales remitidas por los detectores con las de una base de datos que contiene señales interferentes conocidas o previamente observadas. Las señales que no pueden clasificarse como interferencia se remarcán para un posterior análisis con el «Follow-Up Detection Device» (FUDD).

Para poder tener un análisis inmediato e independiente de las señales candidatas, sin que esto signifique una pérdida del preciado tiempo de telescopio disponible, se utiliza un subsistema específico. El FUDD aplica un intensivo procesado de la señal en un ancho de banda relativamente estrecho, alrededor de la señal. Dadas las características básicas de la señal candidata (frecuencia, tasa de desplazamiento en frecuencia, potencia), el FUDD puede utilizar una mayor resolución para conseguir mayor sensibilidad y exactitud. Además, cuando esas características son conocidas con suficiente exactitud, el sistema puede optar por generar un filtro adaptado que aún mejora más la sensibilidad. Esta ganancia de sensibilidad que se obtiene con la utilización del filtro adaptado permite el uso de una antena relativamente pequeña para confirmar la detección preliminar de la gran antena principal. Esta confirmación de señal obtenida en un observatorio independiente se considera esencial para la clasificación de la señal como de origen extraterrestre.

En la práctica este proceso se lleva a cabo simultáneamente mediante dos FUDD situados en ambas antenas. Cuando el SCS determina que una señal remitida por el SDS no puede ser reglada como interferencia, las características de la señal se entregan a los dos FUDDs. Mientras los principales subsistemas del TSS proceden a estudiar una nueva frecuencia de observación sobre la estrella en cuestión, el FUDD sintoniza con la frecuencia de la señal candidata y la observa simultáneamente. Si la señal es persistente, el FUDD en la antena principal puede rápidamente detectar y mejorar las medidas de las características de la señal. Estos parámetros mejorados de la señal junto con los factores de transformación geométrica entre los dos lugares se entregan al FUDD de la antena remota, donde se diseña un filtro adaptado optimizado. La sincronización entre antenas hace que ese filtro actúe sobre los datos que se recogen simultáneamente en la antena principal, así, si la señal se detecta en los dos FUDDs, entonces podemos empezar a considerarla como una auténtica candidata a ser una señal extraterrestre.

El FUDD se implementa en un ordenador convencional con procesador Pentium. Una placa para la FFT implementada con chips Plessey FFT (PDSP16510) puede procesar unas 16 bandas sintonizables de 10 KHz de entre un ancho de banda de 10 MHz; por lo que, en realidad, se utilizan dos unidades de FUDD en cada localización para cubrir los 20 MHz de ancho del TSS. La alta resolución del espectro y la utilización de filtros adaptados para cada banda de señal candidata se generan con el procesador.

Otro aspecto a destacar del sistema es la forma como se monitorizan los datos, el «waterfall plot» o representación en frecuencia-tiempo permite visualizar la evolución de un margen seleccionado de canales. Estos canales, unos 1.000, se representan en forma de una serie de puntos sobre una línea horizontal en la pantalla del ordenador. El tamaño de cada punto es proporcional a la potencia del canal que representa. Cada vez que una nueva serie de medidas llega, cada 0.7 segundos, toda línea de puntos se desplaza totalmente hacia la posición inmediatamente inferior de la pantalla, para dejar espacio para los nuevos datos; así, en general se tiene una pantalla llena de líneas horizontales de puntos donde los nuevos datos aparecen en la línea superior de la pantalla mientras que los más antiguos desaparecen por la parte inferior de ésta.



Una señal constante en frecuencia y siempre activa produce una línea vertical en la pantalla.

Si no hay señales sintonizadas, los puntos en la pantalla adquieren un patrón aleatorio conocido vulgarmente como nieve, similar al aspecto que toma una pantalla de televisión cuando no se tiene sintonizado ningún canal; en cambio, si hay presencia de señal y es suficientemente potente, se observará un cierto patrón destacable entre la nube de puntos aleatorios. En cada canal, la potencia de las señales se suma a la del ruido de fondo presente en el ancho del canal y se produce, entonces, un punto mucho más destacable que los demás. Con el tiempo esos puntos destacados se diferenciarán fácilmente de los puntos de ruido y mostrarán como evoluciona la señal; así, por ejemplo, una señal constante en frecuencia y siempre activa produciría una línea vertical en la pantalla, mientras que si se desplazara en frecuencia generaría una línea inclinada.

Tras lo expuesto podemos afirmar que el sistema del proyecto Phoenix se ha diseñado para poder detectar débiles señales de comunicación procedentes de años luz de distancia y permitir el rechazo de la cacofonía provocada por las comunicaciones terrestres; pero como en todo sistema necesitamos probar su correcto funcionamiento. Idealmente nos gustaría disponer de una señal ET estándar para poder calibrar la sensibilidad y poder comprobar que toda la electrónica y el software de las dos localizaciones,

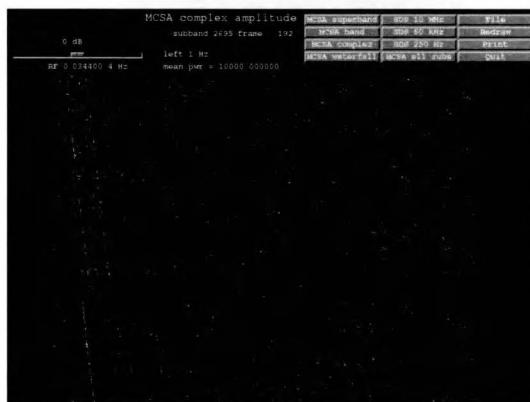
la antena principal y la de verificación, funcionan apropiadamente. Desgraciadamente, hasta que no se descubra la primera, no tendremos ese tipo de señal.



Una señal que se desplazara en frecuencia generaría una línea inclinada.

El reto está, pues, en conseguir algún tipo de señal que nos pueda servir de modelo y que resulte una buena aproximación de una señal ET.

En 1972 se lanzó la Pioneer 10; esta sonda, tras enviarnos las primeras imágenes cercanas de Júpiter y Saturno, ha continuado viajando a través del sistema Solar y más allá de él. El hecho que actualmente se halle a una distancia de más de 10.000 millones de kilómetros y que, además, continúa transmitiendo con unos pocos Wattios de potencia, convierten su emisión en una excelente señal para los test del sistema Phoenix.



Señal que produce el Pioneer 10 representada mediante el «Waterfall plot».

Al igual que muchos otros experimentos radioastronómicos, el Phoenix no se compensa frente al movimiento de la Tierra durante las observaciones. Esto significa que muchos de los transmisores terrestres aparecerán representados por una línea continua como la vista anteriormente, pero, por otro lado, las señales procedentes del espacio profundo mostrarán desplazamientos frecuenciales debidos al cambio de la velocidad relativa

entre el transmisor, ya sea el de una nave o el de otro planeta, y el radiotelescopio en la Tierra. A modo de ejemplo podemos observar la señal que produce el Pioneer 10 representada mediante el «Waterfall plot».

Otra representación de interés es la obtenida durante la observación al modificar la configuración de la pantalla para poder mostrar una de las bandas laterales en las que el Pioneer 10 transmite información. La imagen se puede dividir en dos zonas, en la parte inferior se muestra una parte del espectro (643 Hz) que contiene la portadora, mientras que los dos tercios superiores muestran otra parte diferente del espectro que también contiene señal de datos.

A lo largo del texto hemos visto como se realizan las observaciones y la cantidad de elementos que componen los sistemas de control y detección. También se ha podido observar la cantidad de información que se debe procesar y se ha hecho alusión a los problemas que comporta ese exceso de datos. Para intentar solucionar este serio problema que adquiere, además, mayor relevancia año tras año, se ha iniciado un programa a escala mundial que pretende utilizar la capacidad de procesado de un amplio conjunto de ordenadores esparcidos por todo el mundo para ayudar en la evaluación de los datos captados por los programas de Arecibo; se le conoce como el «seti@home».

Un grupo de científicos de la Universidad de Berkeley, bajo la dirección de David P. Anderson, decidieron implementar un software descargable de la red que, una vez instalado, actuase cuando el ordenador permaneciese inactivo. Funciona a modo de salvapantallas sólo que durante el periodo que actúa, aparte de mostrar en pantalla una imagen variante, se encarga de procesar las señales que se le han entregado desde Berkeley y que se recogen en Arecibo a través del proyecto SERENDIP. Una vez analizada la información, se reenvía automáticamente mediante la conexión a Internet necesaria para colaborar en el proyecto.

La idea es realmente revolucionaria y, a mi entender, una magnífica forma de implicar a todo aquel que disponga de unos mínimos medios informáticos, en una búsqueda que, de concretarse, afectaría a cada uno de los habitantes de nuestro planeta. Sería, además, muy gratificante poder decir que todo el mundo ha podido contribuir de forma relativamente fácil y totalmente desinteresada en el que sería uno de los mayores descubrimientos de la humanidad.

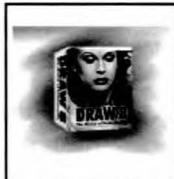
REFERENCIAS

- [1] <http://www.seti.org>
- [2] <http://www.planetary.org>
- [3] <http://setiathome.ssl.berkeley.edu>

Notas:

- Se ha mantenido el sistema de medida americano para mantener la concordancia con la nomenclatura de los proyectos. De manera que al referirnos a billones, debemos considerar su valor americano: 1 billón = 10^9 .





HEDY LAMARR, DE HOLLYWOOD A LA TELEFONÍA MÓVIL

Francesc Comellas, Javier Ozón

Departament de Matemàtica Aplicada i Telemàtica, UPC
comellas@mat.upc.es ozzy@mat.upc.es

El pasado mes de Enero murió a los 86 años la actriz Hedy Lamarr. Es posible que, fuera de los círculos cinematográficos, su nombre no sugiera apenas nada. Su rostro, en cambio, no sólo ha sido portada de innumerables revistas de actualidad y servido de reclamo publicitario de algunos productos emblemáticos, entre ellos RC-Cola y Corel-Draw, sino que pudo, en su día, alterar con su sola presencia el canon de belleza publicitado por los magnates de Hollywood. Hedy Lamarr fue, en efecto, una rutilante estrella del firmamento californiano de los años cuarenta, llegando a ser calificada como la mujer más bella del mundo y protagonizando una serie de productos típicamente americanos, desde *Algiers* en 1938 junto a Charles Boyer, hasta *The Female Animal*, su última aparición en pantalla, en el año 1958.

La misma vida de Hedy Lamarr podría inspirar un típico filme romántico o de aventuras hollywoodiense

Pero Hedy Lamarr era, además de todo eso, una mujer inteligente e imaginativa que detestaba su imagen encantadora, convencida de que «cualquier joven puede tener glamour, basta estarse quieta y parecer estúpida». Esa inquietud y talento, que ya se habían manifestado en su Viena natal (a los cuatro años, por ejemplo, se interesó por el funcionamiento del reloj de oro de su padre), le sirvieron a la postre para patentar numerosos ingenios, entre ellos un collar para perros con propiedades fluorescentes, una técnica de alisamiento del cutis y un sistema de control remoto de torpedos, invento que posteriormente ha sido aplicado tanto en la industria militar como en la telefonía móvil celular y que le reportó fama así como numerosos premios y reconocimiento.

La misma vida de Hedy Lamarr podría inspirar un típico filme romántico o de aventuras hollywoodiense. La huida de un marido traficante de armas, para lo cual hubo de drogar a su asistenta; su legendario desnudo en el filme checo *Ecstasy*; el proceso emprendido contra la editorial que difundió una, según ella, falsa autobiografía; sus seis matrimonios y correspondientes divorcios; las mencionadas patentes de ingenios militares; los encausamientos por

pequeños robos en Drugstores y otros establecimientos, o las actuaciones no anunciadas en un club de Greenwich Village donde cantaba sus propias composiciones en los últimos años de su vida, conforman una personalidad extraordinaria. Hija de un banquero y una pianista de origen judío, Lamarr nació bajo el nombre de Hedwig Eva Maria Kiesler el 9 de Noviembre de 1913, en Viena. En el año 1931, luego de haber abandonado sus estudios y de colaborar en el teatro berlínés con el legendario director Max Reinhardt, Hedy inicia su carrera cinematográfica, alcanzando celebridad dos años más tarde, merced a la secuencia, insertada en *Ecstasy*, en la que por espacio de diez minutos aparece completamente desnuda, primero inmersa en un lago y luego corriendo por la campiña checa.



Casada a los 19 años, en un matrimonio de conveniencia arreglado por sus padres, con el fabricante de armas Fritz Mandl, Hedy calificó posteriormente esa época como de auténtica esclavitud. Fritz era un filonazi despótico que había suministrado armas y municiones a las tropas de Mussolini durante la ocupación de Abisinia y que intentó infructuosamente hacerse con todas las copias existentes de la película en que su flamante esposa aparecía en cueros. Obligada a acompañar a su marido en innumerables cenas de negocios, Hedy tuvo que abandonar su incipiente carrera cinematográfica y cualquier otro tipo de actividad que no fuera el de simple comparsa de Fritz: así, por ejemplo, la actriz sólo podía bañarse cuando

su marido estaba a su lado, acechándola. A pesar de ello, Hedy aprovechó las interminables cenas y reuniones en que escoltaba a Fritz y sus clientes y proveedores para escuchar y aprender algunos pormenores de la tecnología armamentística de su época, conocimientos que más tarde iba a aprovechar para idear y patentar, en los años cuarenta, la técnica de conmutación de frecuencias que le devolvería notoriedad en los últimos años de su vida.

Su fuga rocambolesca de Italia a París primero (drogó, como se ha dicho, a su asistente y se deslizó furtivamente por una ventana) y más tarde a Londres, le permitió viajar finalmente a Hollywood, donde Louis B. Mayer, mandatario de la Metro Goldwyn Mayer, le había de proporcionar un nuevo nombre (inspirado en el de una antigua actriz de la época muda, Barbara La Marr, muerta por sobredosis en 1926) y catapultarla al estrellato. Hedy Lamarr sucedía así a la rubia Jean Harlow en el firmamento hollywoodiense, encarnando un nuevo canon de belleza: el de la morena enigmática y elegante, exótica y sofisticada. Allí, pese a algunos sonados patinazos, como su renuncia a encarnar los papeles protagonistas de Luz de Gas y Casablanca (personajes que posteriormente habían de dar fama internacional a otra actriz europea: Ingrid Bergman), Hedy Lamarr intervino en más de una veintena de películas al lado de actores de renombre, como Clark Gable, James Stewart, Robert Taylor, Ray Milland y Spencer Tracy, obteniendo su mayor éxito con un clásico producto hollywoodiense: el Sansón y Dalila de Cecil B. DeMille.



En el año 1941 Hedy conoció, en el transcurso de una fiesta en Hollywood, al compositor vanguardista George Antheil (1900-1959), un espíritu, como ella, inquieto y cultivado con el que en seguida trabó amistad. Lamarr, que desde los tiempos de su primer marido alimentaba un profundo rencor por el régimen nazi, había ofrecido por entonces sus servicios al recién creado National Inventors Council. La oferta de Hedy, como era de prever, fue declinada por las autoridades competentes, que muy amablemente le aconsejaron contribuir, con su

glamour y estatuto de estrella, a la venta de bonos de guerra y la emisión de posters propagandísticos. Hedy, que pensaba que con su ingenio y bagaje técnico podía contribuir a la victoria aliada, asistía con temor al avance en Europa de las tropas germanas. En aquel momento se veía como una posibilidad la derrota del ejército inglés y el subsiguiente triunfo del régimen nazi.

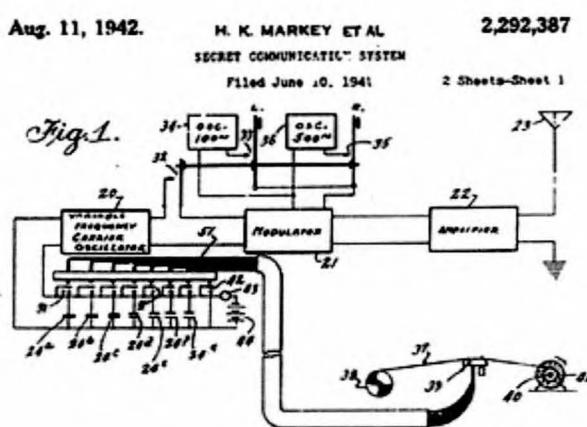
Hedy tuvo la idea de aplicar alguna de las técnicas musicales de George al control remoto de los misiles de guerra

A pesar de todo, Hedy no se desanimó y continuó dándole vueltas a alguna de las ideas que le rondaban por la cabeza. Una de las principales preocupaciones de la opinión pública respecto al conflicto, como manifestaría en una entrevista años más tarde, era el desequilibrio con que, más allá del Atlántico, combatían las aviaciones británica y germana. Así, mientras los aparatos ingleses entraban en territorio enemigo apenas habían abandonado la base y cruzado el canal, los aviones alemanes podían sobrevolar su propio territorio durante cientos de millas antes de llegar a la zona del conflicto. Hedy intuía que la fabricación de un misil teledirigido podía suponer una nivelación de la balanza, solución que el ejército americano no se atrevía a acometer, según algunos testimonios, por miedo a que las señales de control fueran fácilmente interceptadas o interferidas por los efectivos nazis.

Una tarde, mientras estaba sentada al piano con George Antheil, Hedy tuvo la idea de aplicar alguna de las técnicas musicales de George al control remoto de los misiles de guerra (las distintas versiones difieren sobre este punto, habiendo quien sitúa la anécdota del piano más adelante, cuando Hedy y George resuelven aplicar la técnica de los rodillos). Una radioseñal emitida a una determinada frecuencia por las tropas americanas para controlar un torpedo podía ser fácilmente interceptada y bloqueada por el ejército alemán. ¿Por qué no emitir entonces a distintas frecuencias, una en cada intervalo de tiempo, y según una secuencia que pudiera variar en cada ocasión? La idea, que era simple, requería sin embargo una solución práctica. Para ello Hedy y George, que pasaron largas veladas sentados en una alfombra del recibidor de la mansión de Hedy simulando los distintos ingenios con cerillas y una cajetilla de plata, diseñaron un dispositivo inspirado en los rollos perforados de las pianolas y en las cacofonías de algunos experimentos musicales de George (en su famoso Ballet Mécanique 16 pianolas sonaban simultáneamente en una misma sala, sincronizadas por este tipo de mecanismo.) En el diseño final sendos rollos perforados eran incorporados a las estaciones de emisión y recepción, que podían así sincronizar y comu-



tar sus frecuencias (en inglés, frequency hopping) de acuerdo con las instrucciones inscritas en los rollos. De este modo, cualquier intruso que intentara interceptar (o interferir) la señal no podría detectar más que un extraño ruido, perfectamente comprensible, sin embargo, para aquellos que tuvieran en su poder los rollos perforados con la precisa información de la secuencia acordada en cada caso.



El 11 de Agosto de 1942, fecha en la que los Estados Unidos habían ingresado definitivamente en el conflicto, la patente era registrada en Washington con el número de serie 2.292.387, y poco más tarde, cedida al ejército norteamericano. En las imágenes que la documentan puede leerse la inscripción H.K Markey et al. Las iniciales H.K. son las de Hedwig Kiesler (Hedy Lamarr), siendo Markey su apellido de casada de la época. Poco tiempo después, el 1 de Octubre de ese mismo año, aparecía en el New York Times la primera mención pública del invento, a pesar de lo cual, y aunque nadie puso en duda el interés y relevancia del ingenio, las autoridades de la época no consideraron la posibilidad de su realización práctica debido a impedimentos tecnológicos. El propio George Antheil atribuyó el rechazo de su patente a algunas de las indicaciones que habían adjuntado en la documentación y que, dada su fuente de inspiración, se apoyaban en símiles musicales. Imagínense, había escrito, a un hombrecillo de Washington leyendo esas explicaciones y preguntándose cómo diablos iban a introducir una pianola dentro de un torpedo.

De este modo, el ingenio fue olvidado hasta que, años más tarde, las nuevas tecnologías basadas en el transistor de silicio y la fabricación de los primeros microprocesadores permitieron la implantación de métodos eficaces, capaces de incluir la técnica de conmutación de frecuencias. En 1957, quince años después de que la patente de Hedy y George fuera registrada, la firma americana Sylvania Electronics desarrolló un dispositivo de control remoto basado en el frequency hopping y en el que, como es lógico, se habían sustituido los primitivos rollos perforados por circuitos electrónicos. A pesar de esta obligada y lógica innovación, el equipo de ingenieros reconoció en la patente de Lamarr y Antheil (que iba a quedar obsoleta en 1959, año de la muerte del músico) una

precursora de su invento. La primera aplicación conocida de dicho principio se produjo poco tiempo después, durante la crisis de Cuba de 1962, en que la flota naval enviada por los Estados Unidos empleó la conmutación de frecuencias para el control remoto de boyas rastreadoras. Después de Cuba la misma técnica fue incorporada en alguno de los ingenios utilizados en la guerra del Vietnam y, más adelante, en el sistema norteamericano de defensa por satélite (Milstar) hasta que en los años ochenta el hopping vio sus primeras aportaciones en ingeniería civil. Así, con la irrupción masiva de la tecnología digital a comienzos de los años ochenta, la conmutación de frecuencias pudo implantarse en la telefonía móvil celular (con el objeto tanto de proteger la señal de interferencias como de garantizar la intimidad de las llamadas), y más en general en la transmisión de datos sin cable, campo en el que, en palabras de David Hugues, todavía no se han explorado todas sus posibilidades.

Así, con la irrupción masiva de la tecnología digital a comienzos de los años ochenta, la conmutación de frecuencias pudo implantarse en la telefonía móvil celular

Como recompensa a la trascendencia de su proyecto inicial, y por iniciativa y empeño de David Hughes (investigador él mismo y animador de una serie de proyectos que en el seno de la Natural Science Foundation de EE.UU. han empleado técnicas de hopping), la Electronic Frontier Foundation otorgó el prestigioso EFF Pioneer Award a Hedy Lamarr y, a título póstumo, a George Antheil en una ceremonia celebrada en San Francisco el 12 de Marzo de 1997 a la que asistió, en representación de la actriz (que por entonces vivía recluida en Miami), uno de sus hijos, Anthony Loder (un comerciante orgulloso del ingenio de Hedy y dedicado precisamente al negocio de la telefonía.) No fue éste el único reconocimiento oficial. En 1997 Lamarr y Antheil recibieron también el Bulbie Gnass Spirit of Achievement Award, así como una distinción honorífica concedida por el proyecto Milstar. Un año más tarde, en Octubre de 1998, Hedy recibió en Viena (su ciudad natal) la medalla Viktor Kaplan otorgada por la Asociación Austriaca de Inventores y Titulares de Patentes. Finalmente, en el verano de 1999, el Kunsthalle de Viena organizó un proyecto multimediatílico, que incluía una retrospectiva de su carrera cinematográfica, en homenaje a una de las actrices e inventoras más singulares que ha conocido el siglo. Según se dice, cuando le comunicaron a Hedy la concesión del premio de la EFF, ésta se quedó impertérrita y exclamó, escuetamente: «it's about time». Ya era hora.