Appunti di Sistemi Distribuiti

Matteo Gianello

17 febbraio 2014

Indice

1 Introduzione

A partire dalla metà degli anni '80, grazie a due innovazioni tecnologiche si fecero diversi passi avanti nell'uso dei calcolatori. La prima di queste innovazioni fu lo sviluppo di microprocessori potenti; la seconda grande innovazione fu l'invenzione delle reti di computer con l'introduzione delle **LAN** (*Local Area Network*) che consentirono a centinaia di macchine di essere connesse le une alle altre e permisero lo scambio di piccole quantità di informazioni in pochi microsecondi. Il risultato di questa innovazione tecnologica è che oggi mettere insieme una grande quantità di computer tramite una rete ad alta velocità è diventato molto semplice. Questo tipo di sistemi sono solitamente chiamate *reti di computer* o **sistemi distribuiti**.

1.1 Definizione di sistema distribuito

Esistono diverse definizioni di *Sistema distribuito* ma tutte quante sono abbastanza insoddisfacenti. Daremo ora una prima definizione che è sufficiente per i nostri scopi:

Un sistema distribuito è una collezione di computer indipendenti che appare ai propri utenti come un singolo sistema coerente

Da questa definizione possiamo ricavare diversi caratteristiche di un sistema distribuito, la prima è che i sistemi distribuiti sono costituiti da componenti autonomi; la seconda è che gli utenti, siano essi persone o altri programmi, vedono il sistema come un'unica entità. Il che significa che i diversi componenti devono in qualche modo collaborare.
Quello che non viene specificato in questa definizione è il tipo di computer usati per i componenti ne come questi sono interconnessi.
Le caratteristiche più importanti dei sistemi distribuiti sono il fatto che le differenze tra i vari computer e le loro modalità di comunicazione risultano per lo più nascoste agli utenti finali. Inoltre gli utenti possono interagire con un sistema distribuito in modo *consistente* e *uniforme* ovvero indipendentemente da dove e quando avviene l'interazione.
Teoricamente i sistemi distribuiti dovrebbero essere facilmente espandibili e scalabili, inoltre, i sistemi distribuiti sono di norma sempre disponibili anche se alcune sui parti sono momentaneamente fuori uso.
Allo scopo di supportare reti eterogenee e sistemi operativi differenti alle volte si introduce uno strato software tra lo strato di applicazione e i diversi sistemi operativi, questo strato è chiamato **middleware** come mostrato in figura 1.1.

1.2 Obiettivi

La possibilità costruire sistemi distribuiti non implica che tutti i sistemi debbano essere costruiti come sistemi distribuiti. Per far si che sia utile progettare e costruire un sistema distribuito dobbiamo rispettare alcune caratteristiche. un sistema distribuito dovrebbe:

- rendere le risorse facilmente accessibili,
- nascondere il fatto che le risorse sono distribuite sulla rete,
- essere aperto,
- essere scalabile.

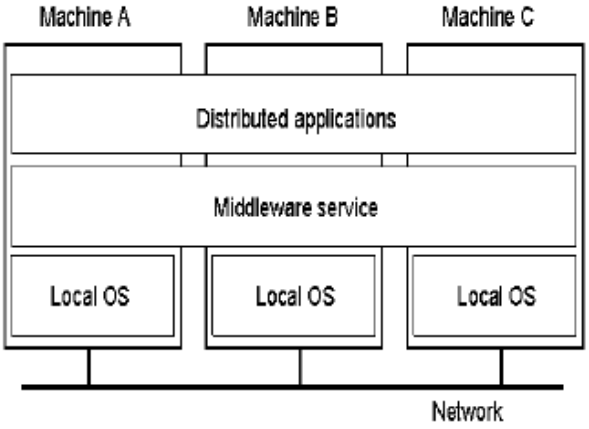


Figura 1: Schema di un middleware

1.2.1 Accessibilità delle risorse

L'obiettivo principale di un sistema distribuito è quello di rendere facile l'accesso alle risorse remote e condividerle in maniera efficiente e controllata.
Ma che cosa intendiamo per risorse? Con il termine *risorse* possiamo indicare qualsiasi cosa, alcuni esempi tipici sono stampanti, computer, dati, file, pagine web o intere reti. Le ragioni che portano a voler condividere le risorse sono molteplici, la prima è sicuramente quella economica, pensiamo ad esempio a ricercatori che condividono un super-computer o ad una stampante condivisa in un ufficio. Inoltre, la connessione di più utenti facilita la collaborazione come avviene nei **groupware** dove gruppi di persone lavorano insieme anche stando in diverse parti del mondo.
Tutto questo incremento di connessione e collaborazione dovrebbe portare però ad una necessaria crescita anche in termini di sicurezza, anche se nella pratica attuale tale incremento nei sistemi di sicurezza non è ancora avvenuto; non è raro trovare sistemi in cui password e altre informazioni sensibili sono inviate come testo in chiaro. Altri problemi legati alla sicurezza sono l'aumento delle *junk mail* o mail di *spam* e l'invio e la raccolta di informazioni riguardanti l'utente per creare un profilo mentre è connesso.

1.2.2 Trasparenza

Uno degli obiettivi principali in un sistema distribuito è quello di nascondere che i processi e le risorse sono distribuiti. Un sistema in grado di presentarsi come un singolo computer è detto **trasparente**.
Possiamo catalogare la trasparenza in diversi tipi in quanto questo concetto può riguardare molti aspetti di un sistema distribuito.
La **trasparenza all'accesso** riguarda le differenze nella rappresentazione dei dati e la modalità di accesso alle risorse da parte degli utenti. Ovvero si desidera nascondere le differenze nelle macchine e trovare un accordo nella rappresentazione dei dati. Un altro importante tipo di trasparenza è la **trasparenza di ubicazione** che si prefigge l'obiettivo di nascondere agli utenti la localizzazione di una risorsa. I *nomi* in questa tipo di trasparenza giocano un ruolo importante in quanto è possibile raggiungere tale trasparenza assegnando ad ogni risorsa un nome logico indipendente dalla sua locazione, un esempio di tale tecnica sono gli *URL*.
Alcuni sistemi distribuiti che consentono lo spostamento delle risorse senza compromettere la possibilità di accesso devono fornire la **trasparenza alla migrazione**. Nel caso in cui le risorse

possono essere spostate *durante* l'utilizzo senza che utenti o applicazioni notino tale spostamento si deve garantire anche la **trasparenza al riposizionamento**.
La **trasparenza alla replica** riguarda la possibilità di fornire una o più copie della stessa risorsa per aumentarne la disponibilità e migliorare le prestazioni, tutto questo nascondendo all'utente il fatto che la risorsa è replicata.
Come già detto l'obiettivo principale dei sistemi distribuiti è la condivisione di risorse, ma questa porta in alcuni casi ad avere una condivisione di tipo *competitivo* ovvero, più utenti vorrebbero accedere alle stesse risorse (es. una tabella di un database) tutto ciò deve essere evitato tramite la **trasparenza alla concorrenza** che deve lasciare la risorsa in uno stato consistente. Questa consistenza può essere ottenuta tramite diverse meccanismi tra cui ad esempio il *locking* nel quale gli utenti ottengono a turno l'accesso esclusivo ad una risorsa.
Per introdurre l'ultimo tipo di concorrenza partiamo da un'altra definizione di sistema distribuito

Ne apprendi l'esistenza quando il crash di un computer di cui non hai mai sentito parlare ti impedisce di portare a termine qualunque lavoro

Questa definizione pone un altro aspetto importante della progettazione di un sistema distribuito, quello della gestione dei guasti; rendere un sistema **trasparente ai guasti** significa far si che un utente non si renda conto che una risorsa smette di funzionare correttamente. La difficoltà più grande è distinguere quando una risorsa è morta da quando è semplicemente molto lenta.
Sebbene in generale si preferisce avere sistemi trasparenti ci sono situazioni in cui nascondere completamente agli utenti la distribuzione del sistema non è una buona idea. Come nel caso si voglia contattare un servizio che sta dall'altra parte del mondo e si voglia una risposta in un tempo inferiore ai 35ms; o quando si vuole che due repliche siano sempre consistenti, nel caso di server in due continenti diversi un aggiornamento potrebbe richiedere alcuni secondi.
Il problema principale che limita però la trasparenza è la trasparenza stessa, infatti, ammettendo che la completa trasparenza di un sistema è *impossibile* è *saggio* cerca di ottenerla a tutti i costi? Rendere la distribuzione esplicita può aiutare gli utenti a capire eventuali comportamenti *anomali* del sistema.

1.2.3 Apertura

un altro obiettivo dei sistemi distribuiti è l'apertura. un sistema distribuito **aperto** è un sistema che offre servizi rispettando delle regole standard che descrivono la sintassi e la semantica dei servizi stessi.
Nei sistemi distribuiti i servizi sono descritti tramite **interfacce** per lo più utilizzando un linguaggio denominato *IDL (interface description language)* che però descrive soltanto la sintassi delle interfacce, ovvero, il nome delle funzioni e i tipi di parametri, i valori di ritorno o le possibili eccezioni sollevate. Per la descrizione di che cosa fa il servizio, invece, si usa solitamente il linguaggio naturale.
Se l'interfaccia è ben specificata un processo che ha bisogno di una determinata interfaccia può comunicare con un altro processo che implementa tale interfaccia; inoltre, consente a due processi distinti di implementare tale interfaccia in due modi completamente diversi, il che porta a due sistemi distribuiti che però operano allo stesso modo.
Una specifica però deve essere *completa* e *neutrale*, completa significa che viene specificato tutto ciò che è necessario per realizzare un'implementazione, ma ottenere la completezza è molto difficile e perciò di solito un programmatore deve aggiungere dettagli specifici dell'implementazione. Per neutrale si intende, invece, che la specifica non deve imporre dettagli sull'implementazione. Completezza e neutralità portano ad altre due importanti caratteristiche che i sistemi distribuiti

dovrebbero soddisfare, **interoperabilità** che significa che due implementazioni di vendor diversi possono collaborare e coesistere basandosi unicamente su di uno standard comune. **Portabilità** indica la possibilità di eseguire un applicazione scritta per un sistema distribuito A su di un sistema distribuito B senza dover apportare modifiche all'applicazione.
Infine un altro obiettivo che i sistemi distribuiti dovrebbero prefissarsi è che l'aggiunta o la sostituzione di componenti dovrebbe risultare facile e non influire sui componenti già presenti; questa caratteristica sta ad indicare che il sistema distribuito è **ampliabile**. Per ottenere la flessibilità in un sistema aperto è fondamentale che esso sia organizzato come un insieme di componenti piccolo e e facilmente sostituibile e adattabile. Ma questo comporta fornire le interfacce non solo dei componenti che si interfacciano direttamente con gli utenti ma anche dei componenti interni.

1.2.4 Scalabilità

La scalabilità sta diventando uno degli aspetti più importanti dei sistemi distribuiti a causa della grande diffusione di internet. Esistono diversi tipi di scalabilità la prima si ha quando un sistema è scalabile rispetto alla sua dimensione il che significa che possiamo aggiungere utenti e risorse. Un sistema può essere scalabile geograficamente quando utenti e risorse sono situati in luoghi molto lontani. Ed, infine, un sistema può essere scalabile anche dal punto di vista dell'amministrazione ovvero quando comprende molte infrastrutture indipendenti rimane comunque facile da gestire.
La scalabilità richiede di affrontare molti problemi. Prendiamo ad esempio la scalabilità rispetto alla dimensione, alcuni servizi in un sistema distribuito possono essere forniti da un unico server questo fa si che aggiungendo utenti quel particolare server diventa un collo di bottiglia per l'intero sistema. A volte l'uso di un solo server è inevitabile come nel caso della gestione di dati sensibili. Per quanto riguarda la scalabilità a livello geografico anch'essa comporta innumerevoli problemi, infatti, la maggior parte dei sistemi distribuiti che lavorano su LAN si basano sulla comunicazione **sincrona** nella quale un *client* richiede una risorsa e resta in attesa che tale risorsa sia disponibile. Questo meccanismo non è applicabile per sistemi globali nei quali la comunicazione tra due computer può richiedere anche qualche millisecondo. In oltre si deve tener conto che la comunicazione su WAN(*wide area network*) è inaffidabile e praticamente sempre punto a punto (*point-to-point*). Al contrario le reti locali più affidabili permettono anche il *broadcasting* ovvero l'invio simultaneo a tutte le macchine delle rete dello stesso messaggio.
Dopo aver visto i problemi di scalabilità ci chiediamo come risolvere tali problemi nei sistemi distribuiti. I problemi di scaling nei sistemi distribuiti si presentano come problemi nelle prestazioni dovuti alle limitate capacità dei server. Ad oggi esistono soltanto tre tecniche di *scaling*: nascondere le latenze, la distribuzione e la replica.
Nascondere le latenze permette di ottenere la scalabilità geografica; l'idea di base è quella di limitare il più possibile i tempi di attesa delle risposte dai servizi remoti. Alcune possibili soluzioni sono anticipare il più possibili la richiesta al server remoto e durante l'attesa della risposta svolgere qualche altra operazione; questo tipo di comunicazione è detta **comunicazione asincrona**. Quando arriva una risposta l'applicazione si interrompe e viene richiamato un gestore (*handler*) speciale per completare la richiesta sollevata. Inoltre la comunicazione asincrona è spesso utilizzata nei sistemi *batch* e nelle applicazioni *parallele*.
Un'altra soluzione è quella di eseguire un *thread* per la richiesta il quale, anche se si blocca, permette agli altri thread di proseguire.
Esiste una classe di applicazioni, però, che non può utilizzare la comunicazione asincrona, questo tipo di applicazioni sono le applicazioni *interattive* nelle quali il client dopo aver effettuato una richiesta ad un server remoto non ha niente di meglio da fare che aspettare la risposta. In questo

caso l'unica cosa da fare è limitare il tempo di attesa e per fare ciò solitamente si sposta il carico di lavoro computazionale che solitamente è svolto dal server sul client. Come ad esempio nel caso di compilazione di *form* per un accesso alla base di dati. Si può inviare al server ogni campo della form per poi aspettare dal server la conferma della correttezza dei dati, oppure, più efficiente è far controllare direttamente al client la correttezza dei dati ed inviare al server l'intera form completa.
Un'altra soluzione ai problemi di scalabilità è la **distribuzione** che comporta prendere un componente spezzarlo in parti più piccole e distribuire tali parti nel sistema. Un esempio molto noto di questo tipo di tecnica è il DNS (*domain name system.* Lo spazio dei nomi è organizzato in un albero dei *domini* divisi in zone non sovrapposte. I nomi di ogni zona sono gestiti da un unico server.
L'ultima tecnica di scalabilità è la **replicazione** che consiste nel duplicare quelle risorse che sono maggiormente richieste in modo da evitare colli di bottiglia e bilanciare il carico sulle risorse.

1.2.5 Tranelli

I sistemi distribuiti si differenziano dal software tradizionale in quanto sono i componenti sono sparsi per la rete. Il fatto di non tenere conto di questa dispersione in fase progettuale rende i sistemi inutilmente complessi. Questi errori sono dovuti a delle ipotesi (false) che ognuno fa quando progetta un'applicazione distribuita:

1. La rete è affidabile
2. La rete è sicura
3. La rete è omogenea
4. La topologia non cambia
5. La latenza è zero
6. L'ampiezza di banda è infinita
7. Il costo di trasporto è zero
8. C'è un solo amministratore

1.3 Tipi di sistemi distribuiti

Esistono diverse categorie di sistemi distribuiti

1.3.1 Sistemi di calcolo distribuiti

Una classe di sistemi distribuiti è quella utilizzata per calcoli ad alte prestazioni; questa categoria può essere divisa in due sottogruppi. Nei **sistemi di calcolo a cluster** l'hardware è composto da una serie di workstation connessi ad una rete locale ad alta velocità e in cui ogni nodo ha lo stesso sistema operativo. Nei **sistemi con tecnologia grid** invece, si intendono sistemi distribuiti costruiti come un gruppo di computer risiedenti in domini di amministrazione diversi con hardware e software differenti.

Sistemi di calcolo a cluster I sistemi di calcolo a *cluster* divennero popolari quando il rapporto prezzo/prestazioni delle *workstation* divenne vantaggioso. In quasi tutti i casi i sistemi di calcolo a cluster sono utilizzati per per la programmazione parallela.
Un esempio di sistema a cluster molto diffuso è il **Beowolf** un sistema basato su linux in cui l'accesso al sistema avviene tramite un singolo nodo principale, che gestisce l'allocazione dei programmi sui nodi. In realtà il nodo master esegue il *middleware* necessario per l'esecuzione dei programmi e la gestione del cluster. Una parte fondamentale di questo middleware sono le librerie con il quale è sviluppato che forniscono solo funzionalità di comunicazione basate sui messaggi e non sono in grado di gestire errori, sicurezza ecc.
Un altro sistema cluster molto diffuso è il **MOSIX** che fa sembrare il cluster un singolo sistema offrendo così ai programmi la trasparenza di distribuzione.

Grid computing Al contrario dei sistemi di calcolo a cluster in cui i componenti hardware sono omogenei nei sistemi *Grid Computing* vi è un alto livello di eterogeneità sia hardware sia software.
Solitamente nella tecnologia grid le risorse di diverse aziende vengono unite per consentire la collaborazione che viene anche detta **organizzazione virtuale**.

1.3.2 Sistemi informativi basati sulle imprese

Un'altra classe di sistemi distribuiti si trova in strutture aziendali che si sono confrontate con una gran abbondanza di applicazione in rete ma per le quali l'interoperabilità non è stata facile. In alcuni casi l'integrazione riguardava solamente un server che veniva contattato da diversi client i quali confezionavano una richiesta e la inviavano a tale server. Un'integrazione più approfondita avrebbe permesso ai client di confezionare una richiesta diretta a molteplici server che avrebbero eseguito un **applicazione distribuita**.

Sistemi transazionali Sono sistemi incentrati su applicazioni relative a basi di dati. In pratica le operazioni sulle basi di dati sono portate a termine sotto forma di **transazioni**. In un sistema distribuito transazionale abbiamo la particolarità che all'interno di una transazione possiamo avere delle chiamate a procedura remote ottenendo così un sistema **RPC transazionale**.

1.3.3 Sistemi distribuiti pervasivi

I sistemi distribuiti visti fin qui hanno la particolarità di essere caratterizzati dalla stabilità: i nodi sono fissi ed hanno a disposizione una connessione di alta qualità. Con l'introduzione di dispositivi mobili ed embedded però abbiamo a che fare con sistemi distribuiti in cui la caratteristica principale è l'instabilità; tali sistemi sono detti **sistemi distribuiti pervasivi**.
I sistemi pervasivi hanno delle caratteristiche particolari, intanto non esiste un amministratore umano ma dopo una prima configurazione da parte del proprietario tali sistemi devono adattarsi al meglio all'ambiente circostante. Inoltre tali sistemi devono avere tre requisiti fondamentali:

- accettare cambi di contesto
- incoraggiare la composizione *ad-hoc*
- riconoscere la condivisione di default

Ovvero un dispositivo deve essere a conoscenza che il suo ambiente è in costante cambiamento e che i dispositivi che compongono il sistema saranno utilizzati in modo diverso da utenti diversi.

In questo tipo di sistemi non vi è alcun tipo di trasparenza, bensì la distribuzione dei dati dei processi e del controllo è intrinseca nei sistemi ed è quindi più efficiente esporla che nasconderla. Alcuni esempi di sistemi pervasivi sono:

- Sistemi domestici
- Sistemi elettronici per l'assistenza sanitaria
- Reti di sensori

2 Architetture

I sistemi distribuiti sono spesso definibili come pezzi di software sparsi su molte macchine. Al fine di dominare la loro complessità è necessario che questi sistemi siano organizzati. Un modo semplice per distinguere l'organizzazione di un sistema distribuito, è quello di distinguere l'organizzazione logica dei componenti software e la relativa realizzazione fisica.
Per *architetture software* intendiamo l'organizzazione e l'interazione dei vari componenti software; mentre l'effettiva realizzazione di un sistema distribuito richiede che i componenti software siano instanziati su macchine reali, l'architettura risultante viene detta *architettura di sistema.*Analizzeremo per prime le architetture centralizzate in cui un server implementa la maggior parte delle funzioni mentre client remoti accedono al server tramite semplici mezzi di comunicazione.

2.1 Stili architetturali

Iniziamo l'analisi dalle diverse tipologie di architetture software in quanto progettazione e adozione di una adeguata architettura sono fondamentali per la riuscita e la manutenibilità del sistema.
Introduciamo ora la nozione di *stile architetturale* che esprime in termini di componenti mezzi di comunicazione e messaggi scambiati. Un *componente* è un'unità modulare con interfacce ben definite e sostituibile nel suo ambiente.
La comunicazione tra i diversi componenti avviene tramite quello che è definito *connettore* ovvero un sistema che implementa le chiamate a procedure remote piuttosto che il passaggio di messaggi o lo streaming dei dati.
Usando componenti e connettori possiamo ottenere diverse configurazioni che a loro volta sono classificati in diversi *stili architetturali.* I più importanti stili architetturali ad oggi identificati sono:

- architettura a livelli (*layer*)
- architetture basate sugli oggetti
- architetture centrate sui dati
- architetture basate sugli eventi

Per quanto riguarda lo stile a livelli è quello più semplice nel quale i componenti sono organizzati a strati in cui un componente del livello L_i può chiamare un componente del livello L_{i-1} ma non può contattare i componenti dello stesso livello. Questo modello è uno dei più utilizzati nelle applicazioni di rete, le richieste scendono lungo la catena gerarchica mentre le risposte risalgono.
Le *architetture basate sugli oggetti* hanno un organizzazione meno rigida, ogni oggetto corrisponde ad un componente e tutti gli oggetti sono connessi tramite *chiamate a procedura remota*; questo tipo di architettura corrisponde esattamente al caso client-server ed insieme a quella a livelli costituiscono il 90% delle architetture dei sistemi distribuiti presenti oggi.
Le *architetture basate sui dati* si sviluppano attorno all'idea che i processi comunicano attraverso un *repository* comune.
Nelle *architetture basate sugli eventi* i processi comunicano essenzialmente attraverso la propagazione degli eventi, i più noti tipi di sistemi distribuiti che utilizzano la propagazione degli eventi sono i sistemi *publish/subscribe* nei quali alcuni processi pubblicano degli eventi ed è il middleware ad assicurarsi che questi eventi siano ricevuti soltanto da qui processi che si sono iscritti a quel

determinato evento.
Nel caso in cui si combinino le architetture basate sugli eventi con quelle basate sui dati si ottiene una architettura chiamata *spazio di dati condivisi* che permette ai processi di essere disaccoppiati anche nel tempo in quanto non è necessario che i processi siano attivi nell'istante in cui avviene la comunicazione.

2.2 Architetture di sistema

Abbiamo visto fino ad ora alcune scelte architetturali, vediamo ora come sono effettivamente organizzati la maggior parte dei sistemi distribuiti considerando dove sono posizionati i componenti software. La scelta di quali componenti usare e di come posizionarli porta alla realizzazione della cosiddetta *architettura di sistema.*

2.3 Architetture centralizzate

La prima architettura che analizzeremo è quella *centralizzata*, in quanto pensare ad un sistema in termini di *client* che richiedono dei servizi a dei *server* facilita la comprensione e la gestione dei sistemi distribuiti.
Nel modello client-server i processi sono suddivisi in due gruppi; un **server** è un processo che implementa uno specifico servizio. Un **client** è invece un processo che richiede un servizio ad un serve inviandogli una richiesta e quindi attendendo una risposta.
La comunicazione tra client e server può essere implementata per mezzo di un semplice protocollo senza connessione quando la rete sottostante è molto affidabile. In questo caso quando un client richiede un servizio invia un messaggio al server indicando il servizio richiesto e i dati di input. Quest'ultimo all'arrivo della richiesta la processa e confeziona i risultati in un altro messaggio.
L'utilizzo di un protocollo senza connessione ha il vantaggio di essere efficiente fino a quando non abbiamo perdita di pacchetti. Si potrebbe pensare di impostare il client affinché rinvii il messaggio quando non riceve alcuna risposta, ma non si può stabilire se è stato perso il messaggio o la risposta e quindi il server ha compiuto oppure no l'operazione richiesta. In alcuni casi le richieste sono *idem-potenti* ovvero possono essere ripetute senza danni.
Per risolvere il problema della perdita di messaggi, molte architetture client-server utilizzano dei protocolli affidabili orientati alla connessione.

Stratificazioni delle applicazioni Il problema principale dell'architettura client server è che non vi è una netta distinzione tra quali sono le funzionalità del client e quelle del server; è stata così introdotta una un nuovo tipo di suddivisione delle diverse funzionalità in base che esse siano più vicine all'utente o ai dati. I tre livelli identificati sono:

- il livello dell'interfaccia utente
- il livello applicativo
- il livello dei dati

Il livello dell'interfaccia utente contiene tutto ciò che è necessario per interfacciarsi con l'utente come la gestione della grafica. Il livello applicativo di solito contiene le applicazioni. Il livello dei dati gestisce tutto ciò che concerne i dati da gestire.
Esistono diversi modi in cui questi tre livelli possono essere istanziati sull'hardware come possiamo vedere in Figura 2. La configurazione più utilizzata è quella nel quale il client implementa solo il livello dell'interfaccia utente (*thin client*) ma esistono altri sistemi una parte dal livello

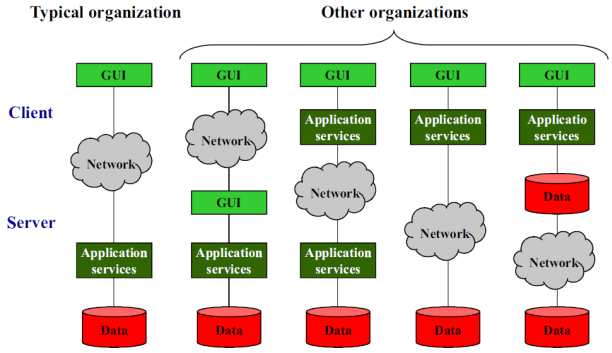


Figura 2: Esempio di suddivisione dei layer

applicativo si trova implementata nel client, in questo caso si parla di *fat client*. Per quanto riguarda il livello dei dati molte volte è gestito da meccanismi che rendono i dati **persistenti** anche quando non vi sono altre applicazioni in esecuzione. In alcuni casi molto semplici il livello dei dati consiste in un *filesystem* ma nella maggioranza dei casi si tratta di una *base di dati*

Architetture multi livello La suddivisione delle applicazioni in tre livelli logici suggerisce anche una distribuzione fisica delle applicazioni client server su molte macchine, la distribuzione più semplice è quella su due macchine:

1. una macchina client che contiene solo i programmi dell'interfaccia utente
2. una macchina server contenete il resto dei programmi e i dati

Si possono suddividere le applicazioni anche in altri modi come visto in Figura 2.
La tendenza degli ultimi anni è quella di suddividere i diversi livelli su più macchine in modo da creare una architetture multi livello. Un esempio pratico ad esempio è quello mostrato in Figura 3 nella quale si mostra un architettura a 3 livelli.

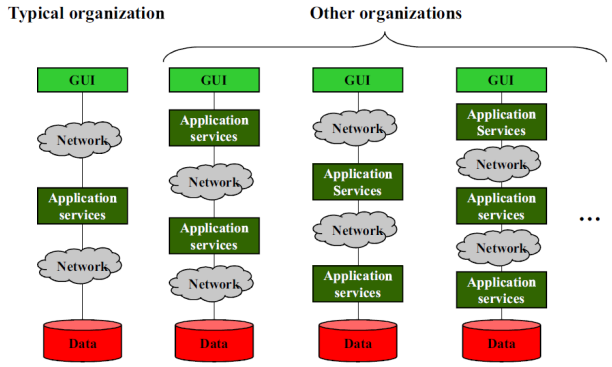


Figura 3: Esempio di suddivisione dei layer su 3 livelli

2.3.1 Architetture decentralizzate

La suddivisione delle architetture client server in livelli denota un tipo di distribuzione detta *verticale* in quanto i componenti sono divisi su più macchine seguendo un criterio *logico*.

Questo tipo di distribuzione è utile con le architetture client-server ma nel caso in cui la distribuzione che conta è quella dei client e dei server e non delle funzioni allora si parla di *frammentazione orizzontale*. Un esempio molto conosciuto di architettura con distribuzione orizzontale sono i **sistemi peer-to-peer**.
I processi che costituiscono un sistema peer-to-peer sono tutti uguali, di conseguenza l'interazione tra processi è quasi del tutto simmetrica e i processi agiscono sia da client che da server e per questo sono anche detti **servent**.
Dato questo tipo di comportamento il problema principale delle reti *p2p* è quello di organizzare i processi in una rete *overlay* ovvero una rete nella quale i processi costituiscono i nodi e i collegamenti rappresentano i canali di comunicazione. Con questa struttura un processo non può comunicare direttamente con un altro processo arbitrario ma deve seguire i canali di comunicazione disponibili.
Esistono due tipi principali di reti *overlay* quelle strutturate e quelle non strutturate.

Architetture peer-to-peer strutturate In una rete p2p strutturata la rete overlay è costruita usando una procedura deterministica, quella più comune è l'uso di una **hash tabel distribuita** (DHT). In questo tipo di struttura ai dati viene assegnato un identificatore univoco in uno spazio molto grande (129:160 bit). Anche ai nodi del sistema viene assegnato un identificatore nello stesso spazio dei dati. Il punto cruciale di questa architettura è quello di creare uno schema efficiente e deterministico che associ univocamente la chiave di un dato con l'identificativo di un nodo basandosi su un'opportuna metrica di distanza. Inoltre, è importante che quando si effettua la ricerca di un dato sia restituito l'indirizzo del nodo al quale questo dato è associato, e ciò si ottiene *instradando* la richiesta al nodo associato.
Ad esempio nel sistema *Chord* i nodi sono organizzati logicamente ad anello, ed i dati sono organizzati assegnando i dati con identificativo k al nodo con più piccolo identificativo $id > k$; questo nodo è chiamato *successore* della chiave k ed è identificato come $succ(k)$ come mostrato in Figura 4.
La parte più importante della gestione di una rete overlay strutturata è la **gestione dell'appartenenza** da parte dei nodi. Quando un nodo vuole unirsi alla rete genera un id casuale ed effettua una ricerca di id a questo punto il sistema restituirà $succ(id)$. Alla fine per inserirsi il nuovo nodo contatterà il successore individuato dalla ricerca e il suo predecessore e si inserirà nell'anello. Inoltre, l'inserimento causa la migrazione dei dati associati ad id da $succ(id)$. L'uscita è molto semplice il nodo informa della sua dipartita il $succ(id)$ e il suo predecessore e trasferisce i dati a $succ(id)$.

Architetture peer-to-peer non strutturate I sistemi peer-to-peer non strutturati si basano su algoritmi casuali per costruire la rete *overlay*. L'idea è quella che ogni nodo mantenga una lista dei suoi vicini. Inoltre, anche i dati sono posizionati sui nodi in modo casuale, di conseguenza l'unico modo di effettuare una ricerca è inoltrare la richiesta a tutta la rete.
L'obiettivo principale di un sistema p2p non strutturato è la creazione di un **grafo disordinato**. Per fare ciò ogni nodo mantiene una lista di c vicini scelti a caso dall'insieme dei nodi *vivi*; questa vista è detta **vista parziale**. Ora supponiamo che un nodo voglia aggiungersi alla rete, esso contatta in altro nodo arbitrario dalla lista dei punti di accesso; questo punto di accesso è un normale nodo della rete che però ha un alta probabilità di essere disponibile. I meccanismi che creano la rete overlay sono detti *push* e *pull* e permettono lo scambio delle informazioni per la costruzione della rete tra i nodi. Questi meccanismi usati singolarmente possono portare alla creazione di reti overlay disconnesse, per questo si usano entrambe le tecniche.

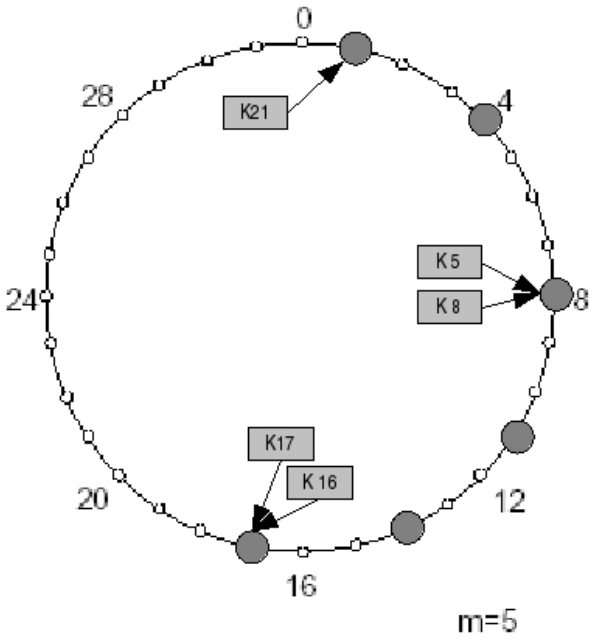


Figura 4: Esempio di sistema Chord

L'uscita dalla rete è molto semplice, infatti, visto che i nodi si scambiano periodicamente le viste parziali basta solo che il nodo lasci la rete, gli altri nodi con il passare del tempo si accorgono dell'assenza del nodo che ha lasciato la rete e lo elimina dalla sua vista parziale.

Gestione della topologia di una rete overlay Anche se i sistemi p2p strutturati e non strutturati sembrano costruire due classi completamente diverse, in realtà selezionando attentamente gli elementi scambiati nelle viste parziali è possibile costruire reti overlay con tipologie specifiche. Un esempio molto interessante sono quelle funzioni che cercano di cogliere la **vicinanza semantica** dei dati che creano reti **overlay semantiche** che permettono ricerche efficienti nei sistemi p2p non strutturati.

Superpeer Specialmente nelle reti non strutturate al crescere delle dimensioni potrebbe diventare problematico localizzare i dati. Questo è dovuto al fatto non esiste un modo deterministico per instradare una richiesta di ricerca.
Per evitare questo problema molte reti p2p hanno proposto l'utilizzo di nodi speciali che mantengono un indice dei dati. Questo tipo di nodi sono detti **superpeer**. I superpeer sono organizzati tra loro tramite una rete p2p; si forma così una struttura ad albero. L'accesso ad un normale peer avviene attraverso l'accesso al superpeer.

2.3.2 Architetture ibride

Fino ad ora abbiamo visto le architetture centralizzate client-server e alcune architetture peer-to-peer ora vediamo come queste due tipologie di architetture possono combinarsi per dare vita ad altre architetture dette *ibride*

Sistemi edge-server I sistemi *edge-server* sono una classe molto diffusa di architetture ibride. In questa classe i server sono posti ai bordi della rete Internet ovvero tra la rete Internet e quelle

aziendali come nel caso degli **Internet service provider**. I client si connettono alla rete internet tramite un edge-server il quale fornisce i contenuti dopo aver applicato dei filtri e funzioni di transcodifica.

Sistemi distribuiti collaborativi Le strutture ibride sono largamente utilizzate nei sistemi distribuiti collaborativi dove sono necessarie velocità nell'entrare nel sistema e per questo viene utilizzato uno schema di tipo client-server. Dopo di che si usa uno schema completamente decentralizzato.
Un sistema concreto che utilizza questo meccanismo è il sistema *BitTorrent*

2.4 Architetture e middleware a confronto

Considerando le questioni architetturali viste fino ad ora ci chiediamo quale ruolo giochi il middleware visto nei capitoli precedenti.
Il middleware in realtà costituisce uno strato tra le applicazioni e le piattaforme distribuite ed il suo obiettivo è quello di fornire un certo grado di trasparenza alla distribuzione. Ciò che accade in realtà è che i middleware seguono uno specifico stile architetturale (ad oggetti come *CORBA* o ad eventi come *TIB/Rendezvous*). Avere un middleware basato su di un certo stile architetturale rende più semplice la progettazione e la realizzazione delle applicazioni.
Lo svantaggio più grande è però il fatto che un middleware può non essere ottimale per una determinata applicazione ciò porta ad avere o middleware molto grandi o a diverse versioni per una specifica classe di applicazioni.

2.4.1 Interceptor

Concettualmente un *interceptor* non è altro che un costrutto software che interrompe il normale flusso di controllo e consente ad altro codice di essere eseguito. Rendere però un interceptor generico è molto arduo e a volte averne uno con funzionalità limitate migliorerà sia la gestione del software che che il sistema distribuito nel suo complesso.
Il concetto è che un oggetto A può richiamare un metodo appartenete all'oggetto B anche se quest'ultimo risiede su una macchina diversa da A. I passi per eseguire questa chiamata remota sono:

1. All'oggetto A viene fornita un interfaccia locale esattamente uguale a quella fornita dall'oggetto B; a questo punto A richiama il metodo disponibile nell'interfaccia

2. La chiamata di A è trasformata in un'invocazione a un oggetto generico disponibile tramite un'interfaccia generale messa a disposizione dal middleware.

3. L'invocazione a questo oggetto generico viene trasformata in un messaggio inviato attraverso l'interfaccia di rete.

3 Modelling

Nel precedente capitolo abbiamo visto il perché di alcune scelte architetturali nei sistemi distribuiti; in questo capitolo invece analizzeremo alcuni dei modelli esistenti e di come questi siano realmente utilizzati.

3.1 Architettura service oriented

Partendo dal concetto di architettura client-server si può pensare di costruire una architettura costruita interamente attorno al concetto di servizio (*service provider, service consumer, service brokers*). Un servizio rappresenta un insieme di funzionalità vagamente legate tra loro che sono messe a disposizione di una unita chiamata **fornitore di servizi** (*service provider*). Il **brokers** mantiene la descrizione dei servizi disponibili che possono essere cercati dai **consumer** che poi li richiamano quando ne hanno bisogno.
Con il termine *orchestration* si indicano l'insieme di invocazioni di determinati servizi in un flusso di lavoro per soddisfare un determinato obiettivo.

3.2 REST style

Lo stile REST (*REpresentational State Transfert*) è allo stesso tempo un buon modo di descrivere il web e un insieme di principi che definiscono come gli standard del Web dovrebbero essere utilizzati.
Gli obiettivi principali del REST includono :

- La scalabilità delle interazioni tra componenti
- Generalità delle interfacce
- Sviluppo indipendente dei componenti
- Componenti intermedi per ridurre le latenze aumentare la sicurezza ed incapsulare i componenti legacy.

Le principali caratteristiche del sistema REST sono che anch'esso è basato su un architettura client-server; le interazioni sono di tipo *stateless*, gli stati devono essere trasferiti di volta in volta dal client al server;. I dati che giungono come risposta ad una richiesta devono essere etichettati come cacheable oppure non-cacheable; ogni componente non può comunicare con se non con i layer più vicini. I client devono supportare il *code-on-demand*; ed infine, i componenti devono esporre un interfaccia uniforme.
Per quanto riguarda l'ultimo punto le interfacce dei componenti devono soddisfare quattro vincoli principali:

- Tutte le risorse devono essere identificate da un id (solitamente un *URI*) ed ogni risorsa con un id è una risorsa valida
- Manipolazione delle risorse tramite la loro rappresentazione, i diversi componenti comunicano tramite il trasferimento di rappresentazioni delle risorse in formati standard (XML) selezionati dinamicamente in base alle capacità o alle informazioni desiderate.
- Messaggi auto-descrittivi, i messaggi contengono al loro interno dei metadati che ne indicano il tipo di richiesta oppure il significato delle risposte. Questa tecnica è utilizzabile per parametrizzare le richieste

- Link ipermediali, i client cambiano il loro stato attraverso le richieste che avvengono tramite dei link ipermediali.

3.2.1 Peer-to-Peer

Come visto nel capitolo 2 nei sistemi peer-to-peer non esistono dei ruoli definiti ma tutti i componendi giocano lo stesso ruolo. Come già detto i sistemi p2p a differenza di quelli client-server permettono di scalare in modo migliore

3.3 Object oriented

Nel caso *Object Oriented* i componenti distribuiti incapsulano i dati e permettono l'accesso e la modifica solo tramite un interfaccia messa a disposizione da ogni componente. I diversi componenti interagiscono tramite RPC. Questo tipo di sistema si basa sul modello p2p ma il più delle volte è utilizzato per implementare meccanismi client-server

3.4 Data centered

Nel caso incentrato sui dai i componenti comunicano, solitamente in modo passivo, con un repository centrale nel quale i dati possono essere recuperati o aggiunti. La comunicazione avviene tramite chiamata a procedura remote e l'accesso ai repository è solitamente sincronizzato.

3.4.1 Il modello Linda

Il modello Linda è un modello introdotto negli anni '80 ed incentrato sulla condivisione dei dati, tale modello è usato principalmente nei sistemi di calcolo parallelo.
In questo modello la comunicazione è persistente e basata sul contesto si ottiene così un alto grado di disaccoppiamento. Le caratteristiche principali di Linda sono:

- I dati sono memorizzati in sequenza in base al tipo di campi (*tuple*)
- Le tuple sono memorizzate in uno spazio persistente e globale (*spazio delle tuple*)
- Operazioni standard come **out**(t) che scrive le tuple nello spazio delle tuple **rd**(p) che legge le tuple che coincidono con il pattern p

Il problema principale di questo sistema è che il modello a spazio di tuple non è facilmente scalabile soprattutto quando aumenta l'area del dominio. Il sistema è proattivo, ovvero esso risponde solo a delle richieste.

3.5 Event-Based

Nei sistemi basati sugli eventi i componenti collaborano per scambiarsi delle informazioni al verificarsi di alcuni *eventi*. In particolare esistono dei componenti che *pubblicano* le informazioni relative all'evento e altri componenti che si *sottoscrivono* a tale eventi.
Il sistema è di tipo asincrono basato su messaggi di tipo multicast ed anonimo in quanto non è importante sapere chi pubblica.

3.6 Mobile Code

Questo modello è diverso dai precedenti, è basato sull'abilità di reallocare i componenti dei sistemi distribuiti a run-time. Esistono diversi tipi di mobile code:

- *Strong mobility:* è la possibilità del sistema di migrare sia il codice sia lo stato di esecuzione.
- *Weak mobility:* è la possibilità di muovere il codice attraverso differenti ambienti di esecuzione

4 Processi

In questo capitolo vedremo come i processi giochino un ruolo fondamentale nei sistemi distribuiti. Il concetto di processo proviene dall'ambito dei sistemi operativi ed è definito come un programma in esecuzione.
Per organizzare efficacemente un sistema client-server è spesso necessario utilizzare tecniche di *multithreading* in quanto questa tecnica permette ai client e ai server di essere costruiti in modo tale che la comunicazione e l'elaborazione locale siano sovrapposti ottenendo un alto livello di prestazione.

4.1 Thread

Anche se i processi costituiscono la base di tutti i sistemi, la loro granularità non è sufficiente a soddisfare i bisogni dei sistemi distribuiti. una gestione più fine, sotto forma di **thread**, rende più facile la costruzione di applicazioni distribuite e ottenere prestazione migliori.

4.1.1 Introduzione ai threads

Prima di capire che cos'è un thread e che ruolo esso gioca nella costruzione di applicazioni distribuite è utile capire che cos'è in realtà un processo e che ruolo ha con i thread.
Per eseguire un programma, un sistema operativo crea un certo numero di processi virtuali. Per tener traccia di questi processi il sistema operativo mantiene aggiornata una **tabella dei processi** contenente elementi che vanno dalla memorizzazione dei registri della CPU, alla mappe della memoria, alla lista dei file aperti alle informazioni sugli *account* e così via.
Il sistema operativo fa si che processi indipendenti non possano in alcun modo influire sulla correttezza degli altri processi, ovvero, è reso trasparente il fatto che più processi possano condividere concorrentemente la stessa CPU e le altre risorse hardware. Questa concorrenza però è ottenuta ad un prezzo abbastanza alto; ogni volta che viene creato un processo il sistema operativo deve creare uno spazio degli indirizzi completamente indipendente. Allocare memoria può voler dire inizializzare segmenti di memoria azzerando segmenti dati, copiare il programma in un segmento di testo e preparando uno *stack* per i dati temporanei. Altrettanto costoso è il passaggio tra un processo ed un altro a livello di CPU, in quanto oltre a salvare il contesto è necessario cambiare i registri ed invalidare la cache.
Come un processo un *thread* esegue il suo pezzo di codice indipendentemente dagli altri threads. A differenza dei processi nei threads non si cerca di ottenere un alto grado di trasparenza, in quanto il fatto di cercare di mantenere la trasparenza fa degradare le prestazione; di conseguenza un sistema basato sui thread gestisce l'insieme minimo delle informazioni per gestire la CPU. Infatti, il **contesto di un thread** è spesso costituito solamente dal contesto della CPU e dalle informazioni per gestire il thread stesso come ad esempio lo stato dovuto al blocco di una variabile *mutex*. È quindi compito degli sviluppatori proteggere l'accesso ai dati tra i vari threads di un singolo processo.

Utilizzo dei thread nei sistemi non distribuiti Il vantaggio principale dell'utilizzo dei thread nei sistemi non distribuiti deriva dal fatto che in un processo a singolo thread quando viene effettuata una chiamata di sistema bloccante l'intero processo viene messo in pausa. Come nel caso di un foglio elettronico dove più celle sono collegate tra loro; in questo caso quando l'utente modifica il valore di una cella anche altre celle vengono rielaborate, ma tale rielaborazione è impensabile in un sistema a singolo thread in quanto il processo resterebbe bloccato in attesa di input e non calcolerebbe il valore delle altre celle.

Un altro vantaggio del multithreading è la possibilità di sfruttare il parallelismo quando si esegue il programma su sistemi multiprocessore. Il multithreading è usato anche nelle grandi applicazioni, le quali solitamente sono sviluppate come un insieme di processi cooperanti; tale cooperazione è realizzata tramite meccanismi di comunicazione tra processi (*IPC, interprocess comunication*), ma questi meccanismi solitamente richiedono molti cambi di contesto che ne rallentano notevolmente le prestazioni. Invece di usare i processi un'applicazione può essere costruita mediante l'utilizzo di threads e la comunicazione tra questi avviene mediante l'uso dei dati condivisi, ed il passaggio da un thread all'altro può essere eseguito a livello utente.

Implementazione dei thread I thread sono spesso forniti sotto forma di pacchetto contente le operazioni di creazione e distruzione dei threads sia le operazioni per la loro sincronizzazione come *mutex* e *condition*. Gli approcci per implementare un pacchetto di thread sono due. Il primo è costruire una libreria che viene eseguita completamente a livello utente, il secondo è lasciare che il kernel sia conscio dei threads e si occupi del loro scheduling.
Usare una libreria utente ha notevoli vantaggi, prima di tutto la creazione e la distruzione dei threads a livello utente è molto meno costosa in quanto il costo è dovuto solo all'allocazione della memoria per creare uno *stack*. Inoltre il cambio di contesto a livello utente può essere fatto con poche istruzioni. L'inconveniente principale dei thread a livello utente però è che una chiamata bloccante di sistema bloccherà l'intero processo e quindi bloccherà tutti i thread del processo.
Questo problema può essere raggirato implementando i threads a livello del kernel ma questo comporta che ogni operazione eseguita su un thread (creazione, distruzione, sincronizzazione e così via) dovrà essere eseguita a livello del kernel richiedendo quindi una chiamata a sistema che risulta essere molto più lenta e costosa come quella di un processo.
La soluzione ai problemi sta nell'uso di una forma ibrida chiamata **processi lightweight**. Un processo leggero viene eseguito nel contesto di un singolo processo (pesante) e per ogni processo ci possono essere più processi leggeri. Oltre a questi il sistema fornisce un pacchetto a livello utente per i threads mettendo a disposizione le solite operazioni. Il pacchetto dei thread è condivisibile da tutti i processi leggeri; questo significa che ogni processo leggero può eseguire il suo thread. Le applicazioni multithread vengono costruite creando dei thread e successivamente assegnando questi thread a un processo leggero.
Il pacchetto dei thread ha una singola routine per pianificare il thread successivo. Quando si crea un processo leggero gli si assegna uno *stack* e lo si mette alla ricerca di un thread da eseguire. I thread in esecuzione sono sono salvati in una tabella in una tabella alla quale i processi leggeri accedono in mutua esclusione tramite l'uso di *mutex* nello spazio utente. Questo significa che la sincronizzazione tra threads è interamente eseguita a livello utente senza la necessità di informare il kernel.
Nel caso in cui vi sia una chiamata di sistema bloccante il contesto di esecuzione passa dalla modalità utente a quella kernel ma continua comunque nel contesto del processo leggero attuale. Nel momento in cui il processo leggero non può più proseguire allora il sistema può decidere di proseguire con un altro processo leggero ritornando alla modalità utente.
I vantaggi di utilizzare un sistema ibrido sono molti. Innanzitutto la creazione, la distruzione e la sincronizzazione dei threads è relativamente poco costosa in quanto avviene a livello utente. Se un processo ha abbastanza processi leggeri allora una chiamata bloccante di sistema non bloccherà l'intero processo. A livello di architetture multiprocessore processi leggeri diversi possono essere eseguiti su CPU diverse. L'unico inconveniente che si presenta è che i processi leggeri devono essere creati e distrutti ma fortunatamente tali operazioni non sono comuni.

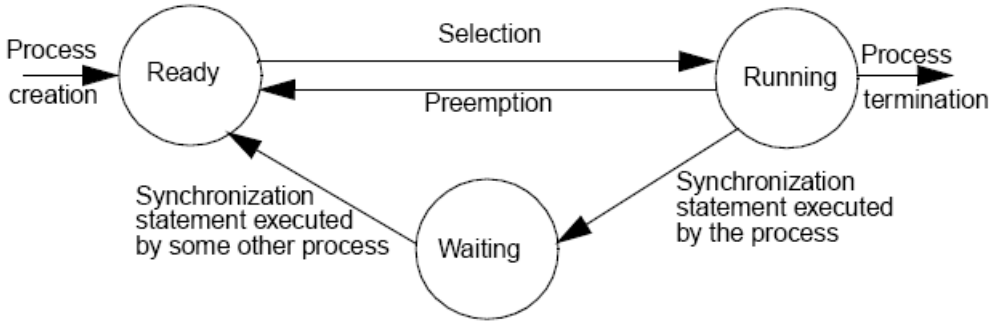


Figura 5: Modello preemtive

4.1.2 Thread nei sistemi distribuiti

Come abbiamo visto il vantaggio principale dell'uso dei threads è che una chiamata di sistema bloccante non blocca l'intero processo. Questa caratteristica è molto vantaggiosa nel caso di realizzazione di comunicazioni multiple come ad esempio la gestione di comunicazioni client-server.

Client multithread Per raggiungere un buon grado di trasparenza alla distribuzione i sistemi distribuiti che operano su reti globali hanno la necessità di nascondere lunghi tempi di propagazione dei messaggi. La tecnica più comune per nascondere la latenza dei messaggi è quella di avviare la comunicazione ed immediatamente iniziare a fare qualcos'altro. Un esempio molto diffuso sono i browser web che iniziano la comunicazione, ricevono una parte del codice HTML ed iniziano a visualizzare la pagina prima ancora di aver concluso la comunicazione.

Server multithread Anche se l'uso di client multithread offre notevoli vantaggi, il vero uso del multithreading è lato server. La pratica dimostra come l'uso del multithreading semplifica la codice e rende più facile lo sviluppo di applicazioni parallele per ottenere un alto livello di prestazioni.
Vediamo il caso di un *file server* dove un **dispacher** riceve in ingresso su di una porta le richieste provenienti da diversi client. Dopo averla esaminata il dispacher seleziona un **worker thread** inattivo a cui assegnare la richiesta. Il *worker* procede con la richiesta ed esegue una lettura bloccante sul file system locale; questo può far si che il thread vengo bloccato in attesa della lettura da disco, in tal caso viene selezionato un altro thread (*worker* o *dispacher*) che procede con la sua esecuzione.

4.1.3 Il modello preemtive

Nei sistemi moderni oltre ai thread viene utilizzato il modello *preemtive*, ovvero è possibile forzare un processo ad abbandonare il suo stato di esecuzione. Solitamente questo meccanismo è utilizzato per implementare un meccanismo di *time slicing* come mostrato in Figura 5.

4.2 I thread in C

Tutti i sistemi UNIX sono multitasking con il sistema preemptive; tradizionalmente tutti i processi sono creati allo stesso modo tramite l'uso della primitiva `fork()`. La *fork*produce una copia del processo chiamante; questa copia è esattamente identica all'origina tranne per il valore

restituito dalla `fork` che per il processo figlio vale *0* mentre nel padre il valore restituito è il *pid* del figlio. Un piccolo esempio: La fork restituisce due copie completamente indipendenti

```
1 /*do parent stuff*/
2 ppid = fork ();
3 if (ppid < 0) {
4         fork_error_function ();
5 } else if (ppid == 0) {
6         child_function ();
7 } else {
8         parent_function ();
9 }
```

Codice 1: Esempio di uso della fork

dello stesso processo, questa indipendenza permette la protezione della memoria e la stabilità ma causa dei problemi quando si vuole che diversi processi lavorino sullo stesso problema; infatti sarebbe necessario usare *pipes* oppure *SysV IPC*. Inoltre il costo di switching tra processi multipli è molto alto, la sincronizzazione è lenta ed esistono dei limiti sul numero di processi che possono essere schedulati efficacemente.
Per questo sono stati introdotti i threads che invece possono essere schedulati all'interno del processo e risolvono molti problemi del lavoro multi processo. L'API più popolare per creare una applicazione multithread in ambiente UNIX è la *pthread* (*POSIX thread*).
Le operazioni che si possono eseguire con quest'API sono la creazione, la distruzione, la sincronizzazione (*join*), lo scheduling, il controllo dei dati e l'interazione con il processo principale. I threads dello stesso processo condividono le istruzioni di processo, gran parte dei dati, i descrittori dei file aperti, i segnali e lo user e il group id. Mentre per ogni thread abbiamo un distinto *ThreadID*, un certo numero di registri, uno stack pointer ed una certa priorità come possiamo vedere Figura 6. Vediamo ora quali sono le funzioni della API *pthread*

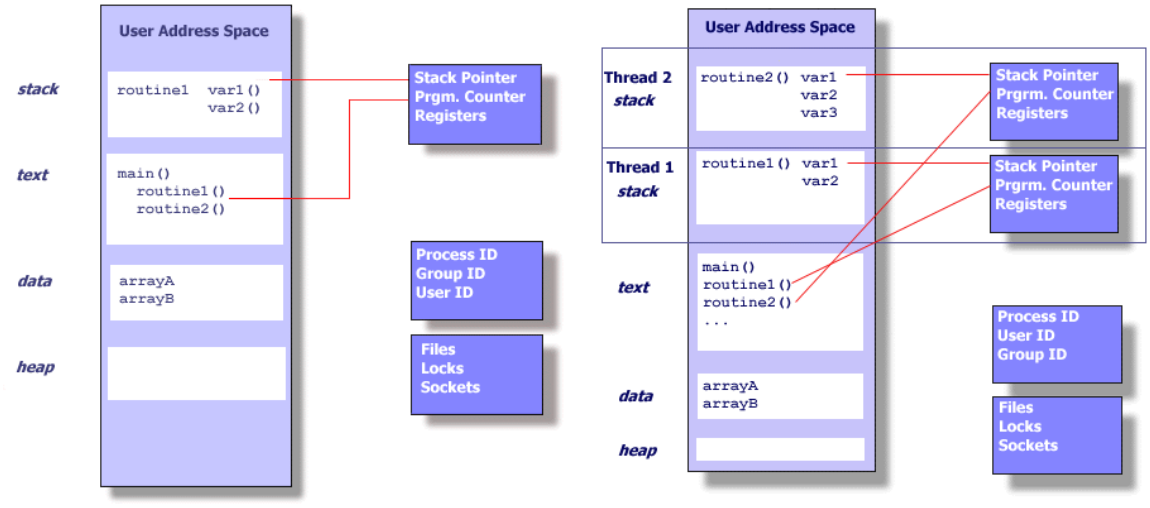


Figura 6: Memoria nel caso di processo (a) e di thread (b)

Thread creation La funzione per la creazione dei thread è: dove i valori sono:

```c
int pthread_create (pthread_t *id, const pthread_attr_t *attr, void *(*routine)(void *), void *arg)
```

Codice 2: Funzione di creazione dei thread

id: un valore che identifica il thread che viene restituito dalla funzione.

attr : un attributo che può essere utilizzato per impostare alcuni valori del thread. Se viene impostato a *NULL* vengono impostati i valori di default.

routine: indica la funzione C che il thread eseguirà una volta creato.

arg: un singolo argomento che può essere passato a`routine`, deve essere passato come riferimento ad un puntatore di tipo `void`; in caso non vi siano valori si imposta a *NULL*.

Thread termination Esistono diversi modi in cui un pthread può terminare:

- Il thread termina la sua routine.
- Nel thread viene richiamata la `pthread_exit`.
- Il thread è cancellato da un altro thread tramite la chiamata della funzione `pthread_cancel`.
- L'intero processo termina quando viene chiamata una delle funzioni `exec` o `exit`.

Tramite la `pthread_exit` è possibile specificare uno stato di terminazione che può essere restituito alla sincronizzazione del thread. Inoltre è molto importante ricordare che la `pthread_exit` non chiude i file ed ogni file aperto all'interno del thread rimane aperto anche alla sua terminazione. Se la funzione *main* termina con una `pthread_exit` prima che i threads siano conclusi i threads proseguono la loro esecuzione altrimenti terminano alla conclusione del *main*. Vediamo un esempio di creazione e terminazione dei thread in C nel Listato 3

```c
#include <stdio.h>
#include <stdlib.h>
#include <pthread.h>

#define NUM_THREADS 5

/* Esempio di programma che utilizza la libreria pthread */
void *PrintHello(void * threadid) {
       int *temp;
       temp = (int *) threadid;
       printf("\n%d: Hello World!\n", *temp);
       pthread_exit(NULL);
}

int main(int argc, char *argv[]) {
       pthread_t threads[NUM_THREADS];
       int rc,t;

       for (t = 0; t < NUM_THREADS; t++) {
              printf("Creazione del thread %d\n", t);
              rc = pthread_create(&threads[t], NULL, PrintHello, (void *) &t);
              if(rc) {
```

```
23                      printf("ERROR; return code from pthread_create() %d\n",rc);
24                      exit(-1);
25              }
26      }
27
28      for (t = 0; t < NUM_THREADS; t++) {
29              pthread_join(threads[t],NULL);
30      }
31      pthread_exit(NULL);
32  }
```

Codice 3: Esempio di uso della API pthread

Passaggio di argomenti Come abbiamo visto nella *pthread_create* è possibile impostare l'ultimo attributo con un attributo da passare alla routine che il thread eseguirà. Tale attributo deve essere convertito in un puntatore di tipo void. Tale passaggio presenta però alcuni tranelli, vediamo come nel Listato 4 come il passaggio per indirizzo crei un errore nell'esecuzione. Infatti, provando ad eseguire tale programma si rischia che più di un thread acceda contemporaneamente alla variabile t e si rischiano quindi di ottenere dei valori sbagliati.

```
1  int rc, t;
2  for(t=0; t<NUM_THREADS; t++) {
3          printf("Creating thread %d\n", t);
4          rc = pthread_create(&threads[t], NULL, PrintHello, (void *) &t);
5          ...
6  }
```

Codice 4: Errore nel passaggio di argomenti ad un thread

Un possibile risultato di questo codice è quello seguente dove si può vedere che il thread numero 3 stampa il valore 4 anche se il thread numero 4 non è ancora stato creato (In realtà tutti i thread accedono alla variabile t con un ritardo in quanto manca la stampa del thread numero 0).

```
Creazione del thread 0
Creazione del thread 1
1: Hello World!
Creazione del thread 2
2: Hello World!
Creazione del thread 3
3: Hello World!
4: Hello World!
Creazione del thread 4
```

Per passare un argomento ad una routine è necessario controllare l'accesso ai dati da parte dei threads in modo che non vi siano possibili conflitti come nel caso del Listato 5. Nel quale viene passato ad ogni routine un puntatore ad un dato diverso.

```
1  int *taskids[NUM_THREADS];
2  for(t=0; t<NUM_THREADS; t++) {
3          taskids[t] = (int *) malloc(sizeof(int));
4          *taskids[t] = t;
5          printf("Creating thread %d\n", t);
```

```
6         rc = pthread_create(&threads[t], NULL, PrintHello, (void *) taskids[t]);
7         ...
8 }
```

Codice 5: Metodo corretto nel passaggio di argomenti ad un thread

Joining threads L'operazione di *join* è uno dei modi nel quale si può implementare la sincronizzazione tra thread.

```
1 int pthread_join(pthread_t thid,void **thread_return)
```

dove i diversi campi sono:

thid è l'identificativo del thread su cui fare la join

thread_return è il possibile valore di ritorno che si ottiene dall'invocazione della `pthread_exit`

Il funzionamento della funzione di *join* è specificato in Figura 7

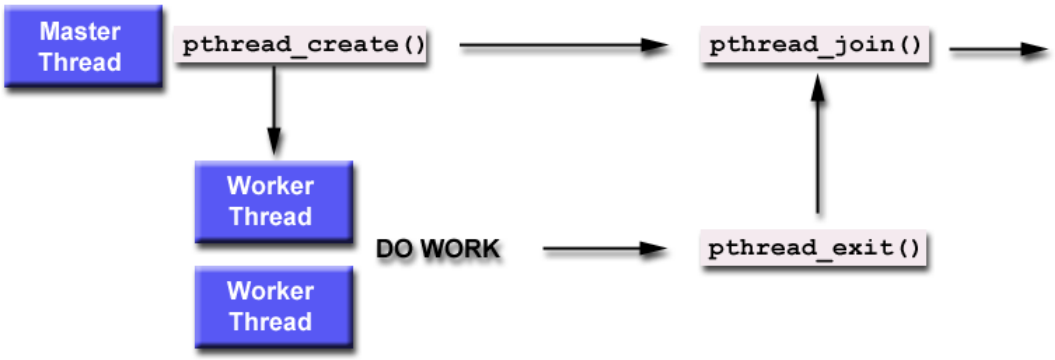


Figura 7: Funzionamento dell'operazione di join

I mutex Un *mutex* funziona come un *lock* proteggendo l'accesso a dei dati condivisi. Il concetto principale è che un solo thread alla volta può bloccare una variabile di tipo mutex, se più thread tenta di bloccare un mutex soltanto uno di questi effettuerà l'operazione con successo, inoltre i thread non possono bloccare un determinato mutex finché il thread che lo detiene non lo libera.
È compito del programmatore assicurarsi che ogni thread che utilizza dei dati condivisi usi i mutex.
Una variabile di tipo mutex può essere dichiarata sfruttando la parola chiave `pthread_mutex_t`, per inizializzare tale variabile, invece, è possibile sfruttare due metodi:

- Il primo è l'inizializzazione statica come ad esempio

  ```
  pthread_mutex_t mymutx = PTHREAD_MUTEX_INITIALIZER;
  ```

- Il secondo metodo è l'inizializzazione dinamica richiamando la routine

  ```
  pthread_mutex_init
  ```

In questo caso è possibile impostare alcuni parametri dell'oggetto mutex tramite le primitive `pthread_mutexattr_init` e `pthread_mutexattr_destroy` che rispettivamente creano e distruggono gli attributi di un mutex

In entrambi i casi di inizializzazione l'oggetto mutex è inizializzato *unlocked*. Infine, la routine `pthread_mutex_destroy` permette di rilasciare un mutex di cui non si ha più bisogno.
Esistono tre primitive per la gestione dei mutex, queste sono:

- `pthread_mutex_lock:` che si usa per acquisire un lock su di una variabile, nel caso in cui tale lock sia detenuto da un altro thread il thread che ha richiesto il lock si blocca finché il blocco non viene rilasciato.
- `pthread_mutex_trylock` molto simile alla routine precedente solo che nel caso in cui il blocco sia detenuto da un altro thread allora la routine restituisce un codice di errore che indica *busy*; è molto utile nel caso si vogliano prevenire condizioni di deadlock.
- `pthread_mutex_unlock:` questa routine permette di rilasciare il lock in possesso del thread ma restituisce un errore nel caso in cui si voglia rilasciare un lock su di una variabile già sbloccata o bloccata da un altro thread.

```
1  #include <pthread.h>
2  #include <stdio.h>
3  #include <stdlib.h>
4
5  typedef struct{
6         double *a;
7         double *b;
8         double sum;
9         int veclen;
10 } DOTDATA;
11
12 /*Define global variables and mutex*/
13 #define NUMTHREADS 4
14 #define VECLEN 100
15 DOTDATA dotstr;
16 pthread_t callThd[NUMTHREADS];
17 pthread_mutex_t mutexsum;
18
19 void *dotprod(void *arg) {
20        int i, start, end, offset;
21        double mysum;
22
23        offset = (int) arg;
24        start = offset * dotstr.veclen;
25        end = start + dotstr.veclen;
26        mysum = 0;
27        for (i = start; i < end; i++) {
28               mysum += (dotstr.a[i] * dotstr.b[i]);
29        }
30
31        pthread_mutex_lock (&mutexsum);
32        dotstr.sum += mysum;
33        pthread_mutex_unlock (&mutexsum);
34        pthread_exit((void *)0);
```

```
35 }
36
37 int main (int argc, char *argv[]) {
38        int i;
39        int status;
40
41        /* Assign storage and initialize values */
42        dotstr.a = (double*) malloc (NUMTHREADS*VECLEN*sizeof(double));
43        dotstr.b = (double*) malloc (NUMTHREADS*VECLEN*sizeof(double));
44        for (i=0; i<VECLEN*NUMTHREADS; i++) {
45               dotstr.a[i]=1;
46               dotstr.b[i]=1;
47        }
48        dotstr.veclen = VECLEN;
49        dotstr.sum=0;
50        /* initialize the mutex */
51        pthread_mutex_init(&mutexsum, NULL);
52        /* Create threads to perform the dotproduct */
53        for(i=0;i<NUMTHREADS;i++) {
54               pthread_create(&callThd[i], NULL, dotprod, (void *)i);
55        }
56        /* Wait on the other threads */
57        for(i=0;i<NUMTHREADS;i++) {
58               pthread_join( callThd[i], (void **)&status);
59        }
60        /* After joining, print out the results */
61        printf ("Sum = %f \n", dotstr.sum);
62        free (dotstr.a);
63        free (dotstr.b);
64        pthread_mutex_destroy(&mutexsum);
65        pthread_exit(NULL);
66 }
```

Codice 6: Esempio di uso delle variabili mutex

Condition variables Mentre i mutex implementano la sincronizzazione tramite il controllo degli accessi sui dati le *condition variables* permettono ai thread di sincronizzarsi in base ad un determinato valore di un dato. Senza quest'aspetto dei thread bisognerebbe implementare un polling per verificare quando una particolare condizione viene riscontrata. Le *condition variables* sono un modo per ottenere lo stesso risultato senza il polling, e possono essere utilizzate anche insieme ai mutex. L'utilizzo principale sono tutti quei problemi della categoria**produttore-consumatore**.
Come per i mutex le condition variabile sono dichiarate utilizzando la parola chiave `pthread_cond_t`; per l'inizializzazione esistono due metodi:

- Statico

  ```
  pthread_cond_t myconvar = PTHREAD_COND_INITIALIZER
  ```

- Dinamico tramite la funzione `pthread_cond_init` che permette di settare anche gli attributi della variabile tramite le due primitive

  ```
  pthread_condattr_init
  pthread_condattr_destroy
  ```

Che permettono rispettivamente di inizializzare e distruggere gli attributi della variabile

Infine tramite la primitiva `pthread_cond_destroy` è possibile liberare una variabile condizionale che non è più necessaria.
Per la gestione di questo tipo di variabile esistono diverse primitive:

- `pthread_cond_wait` è una routine che il thread finché una determinata condizione non si verifica, se chiamata quando vi è un lock attivo la routine sblocca il mutex e lo blocca nuovamente quando il thread si sveglia.
- `pthread_cond_signal` questa routine risveglia gli altri thread in attesa di una *condition variables*
- `pthread_cond_brodcast` può essere utilizzata al posto della routine precedente se ci sono più thread bloccati in uno stato di *wait*.

4.3 Concorrenza in Java

Come per il C anche il Java fornisce il supporto alla concorrenza a livello di linguaggio, esso mette a disposizione delle classi per istanziare ed eseguire nuovi thread più i metodi di sincronizzazione e le variabili di condizione. Il modo più semplice per creare un thread è quello di utilizzare la classe `java.lang.Thread` in questo caso è sempre necessario implementare un metodo `run()`.

```
1  public class MyThread extends Thread {
2         private String message;
3         public MyThread(String m) {message = m;}
4         public void run() {
5                 for(int r=0; r<20; r++)
6                         System.out.println(message);
7         }
8  }
9
10 public class ProvaThread {
11        public static void main(String[] args) {
12                MyThread t1,t2;
13                t1=new MyThread("primo thread");
14                t2=new MyThread("secondo thread");
15                t1.start();
16                t2.start();
17        }
18 }
```

Codice 7: Uso della classe Thread in Java

Come vediamo la nostra classe che implementa un thread estende l'oggetto *Thread*, in questa classe viene fatto l'override del metodo `run()` il quale non è altro che la routine che viene eseguita dal thread. Per far partire il thread è necessario richiamare il metodo `start()` dopo aver creato un nuovo oggetto.
Un'altra possibile soluzione è l'utilizzo dell'interfaccia `Runnable` la quale specifica soltanto che deve esistere un metodo `run()` che deve essere implementato. La classe `Thread` implementa anch'essa l'interfaccia `Runnable`. Come vediamo nel Listato 8 a differenza del caso precedente oltre all'oggetto *MyThread* deve anche essere creato un oggetto *Thread* corrispondente, al quale viene poi passato l'oggetto MyThread, ed infine il metodo `start()`viene invocato sull'oggetto Thread.

```
1  public class MyThread implements Runnable{
2          private String message;
3
4          public MyThread (String m) { message = m; }
5          public void run() {
6                  for (int r = 0; r < 20; r++)
7                          System.out.println(message);
8          }
9  }
10
11 class ProvaThread {
12         public static void main (String[] args) {
13                 Thread t1, t2;
14                 MyThread r1, r2;
15                 r1 = new MyThread("PrimoThread");
16                 r2 = new MyThread("Secondo␣Thread");
17                 t1 = new Thread(r1);
18                 t2 = new Thread(r2);
19                 t1.start();
20                 t2.start();
21         }
22 }
```

Codice 8: Utilizzo dell'interfaccia Runnable

L'esecuzione dei thread non segue un ordine predefinito ma lo stesso codice può produrre risultati diversi su diversi computer o addirittura sullo stesso. Questo caratteristica è chiamata *non-determinismo* ed è un punto focale nella concorrenza.
Java di per se implementa il modello *preemtive* e nel caso sia disponibile un meccanismo di *time-slicing* allora java esegue i thread con la stessa priorità tramite una meccanismo di *round-robin*.
Per definire quando un sistema multithread è corretto si devono rispettare due proprietà:

- *Sicurezza:* Un sistema si dice sicuro quando gli eventi malevoli non accadono.
- *Longevità:* Un sistema è longevo quando le cose buone possono accadere.

I possibili guasti che rientrano nella categoria Sicurezza sono quei guasti che avvengono a livello di esecuzione come i conflitti *read/write* e *write/write*. I meccanismi che invece riguardano la Longevità sono quei meccanismi che bloccano l'esecuzione del programma come:

- Lock
- Waiting
- CPU contention

Solitamente, purtroppo, le cose più semplici che si possono fare per aumentare la longevità ne riducono però la sicurezza e vice versa.
Vediamo ora quali sono i meccanismi che Java mette a disposizione per supportare la concorrenza.

Exclusion In un sistema sicuro ogni oggetto protegge se stesso da possibili violazioni della sua integrità. le tecniche di esclusione preservano l'invariante di un oggetto. Tre sono le tecniche principali per permettere l'*esclusione*:

- Immutabilità
- Esclusione dinamica (Locking)
- Esclusione strutturale

Per quanto riguarda l'**immutabilità** si ottiene creando le classi in modo che gli oggetti proteggano se stessi come nel Listato 9

```
1 class ImmutableAdder {
2         private final int offset;
3         public Immutableadder (int a) {
4                 offset = a;
5         }
6         public int addOffset (int b) {
7                 return offset + b;
8         }
9 }
```

Codice 9: Esempio di oggetto immutabile

I vantaggi di questa tecnica sono il fatto che non richiede sincronizzazione ed è molto utile per condividere degli oggetti tra i threads, ma sfortunatamente ha dei limiti di applicabilità.
Per parlare di sincronizzazione introduciamo prima l'esempio del Listato 10

```
1  public class RGBColor {
2         private int r;
3         private int g;
4         private int b;
5
6         public void setColor (int r, int g, int b)
7                 checkRGBVals(r, g, b);
8                 this.r = r;
9                 this.g = g;
10                this.b = b;
11        }
12 }
```

Codice 10: Esempio sincronizzazione

Ora immaginiamo che due thread chiamati *red* e *blue* vogliano impostare contemporaneamente il loro colore sullo stesso oggetto di tipo `RGBColor` a questo punto potrebbero verificarsi dei problemi in quanto i due thread tentano di scrivere lo stesso dato violando così la sua integrità.
Per risolvere questo problema java tramite il locking serializza l'esecuzione del codice dichiarato *synchronized*. Ogni istanza di un oggetto possiede tali meccanismi di lock in quanto derivati dalla classe `Object`, l'unica eccezione si ha con l'utilizzo di array, infatti, bloccare un array non blocca gli elementi di tale array.
Esistono due modi per bloccare una parte di codice, si può dichiarare *synchronized* un intero metodo o un singolo blocco di codice, in caso di singolo blocco la funzione `synchronized` richiede l'oggetto sul quale effettuare il lock (Listato 11 e Listato 12).

```
synchronized (object) {
        //Lock is held
        ...
}
//Lock is released
```

Codice 11: Sincronizzazione di una parte di codice

```
synchronized void f() {
        //Lock is held
        /* Body */
}
//Lock is released
```

Codice 12: Sincronizzazione di un metodo

I lock vengono automaticamente acquisiti all'ingresso del blocco o del metodo dichiarato `synchronized` e rilasciato all'uscita da esso.
Alcune regole chiave per l'uso della sincronizzazione sono:

- Sempre quando si effettua un aggiornamento a dei campi di un oggetto.

```
synchronized (point) {
        point.x = 5; point.y = 7;
}
```

- Tutte le volte che si accede a dei dati che potrebbero essere aggiornati.

```
synchronized (point) {
        if (point.x > 0) {...}
}
```

- Si può fare a meno di sincronizzare parti di metodo stateless

```
public void f() {
        synchronized (this) {
                state = ...;
        }
        operations();
}
```

- **Mai** sincronizzare parti di codice che contengono invocazioni ad altri oggetti

```
public void f() {
        synchronized (this) {
                ...
        }
        h.foo();
}
```

La strategia più sicura (ma non la più efficace) per realizzare un'applicazione OO concorrente è quella di utilizzare oggetti completamente sincronizzati, anche detti oggetti *atomici*, nei quali tutti i metodi sono sincronizzati, non esistono campi pubblici o altri tipi di violazione nell'incapsulamento, tutti i metodi sono finiti e hanno modo di rilasciare il lock, tutti i campi sono

inizializzati ad un valore consistente nel costruttore, ed infine lo stato dell'oggetto è consistente sia all'inizio che alla fine di ogni metodo anche in presenza di eccezioni.
Uno dei problemi principali della programmazione concorrente è il *deadlock*, tale problema si verifica quando due o più oggetti sono acceduti da due o più threads e tali thread detengono un lock mentre tentano di acquisire un lock detenuto da un altro thread.
L'assegnamento di un valore ad una variabile è un'operazione atomica (a parte per i *long* e i *double*), questo significa che generalmente non è necessario sincronizzare l'accesso ad una variabile. Tuttavia i thread solitamente memorizzano il valori delle variabile in memoria locale, questo significa che se un thread cambia il valore di una variabile un altro thread non vede il cambiamento. Per evitare questo meccanismo bisogna sincronizzare la variabile oppure dichiararla di tipo `volatile` che significa che ogni volta che una variabile è usata deve prima essere letta dalla memoria principale.
Il confinamento implementa l'incapsulamento garantendo che al massimo un'attività alla volta acceda agli oggetti. Questo meccanismo permette l'accesso ad un solo thread alla volta senza utilizzare i locking dinamici. Il punto principale è quello avere un punto di uscita dal thread. Esistono quattro categorie per verificare che un riferimento r ad un oggetto x può uscire da un metodo m:

- m passa r come argomento di un'invocazione ad un metodo o ad un costruttore di un oggetto
- m passa r come valore di ritorno di un metodo.
- m registra r in un campo accessibile da altre attività
- m rilascia un riferimento che però permette l'accesso ad r

Per quanto riguarda le collezioni, il framework `java.util.Collection` basata su uno schema *Adapter* permette la sincronizzazioni delle classi collection, infatti, ad eccezione di `Vector` e `Hashtable` le classi base per le collezioni (come `java.utili.ArrayList`) sono non sincronizzate. Sono state così costruite una serie di classi sincronizzate attorno alle classi base come la `Collection.synchronizedList`.
Come abbiamo detto prima però la sincronizzazione non è molto efficiente infatti richiamare un metodo sincronizzato richiede un tempo quattro volte maggiore rispetto a metodi non sincronizzati; inoltre, esso riduce la concorrenza e diminuisce le performance, infine non vi è alcun modo di controllare il meccanismo dei lock.
Con java versione 5 sono stati introdotti nuovi meccanismi di sincronizzazione come la *sincronizzazione condizionata*. Prendiamo come esempio un parcheggio con una certa capacità e dei metodi che permettono l'arrivo e la partenza di automobili come esemplificato nel Listato 13.

```
1  public class CarParkControl {
2         protected int space;
3         protected int capacity;
4
5         public CarParckControl (int n) {
6                capacity = space = n;
7         }
8
9         syncronized public void arrive() {
10               ...; --space; ...;
11        }
12        syncronized public void depart() {
```

```
13              ...; ++space; ...;
14      }
15  }
```

Codice 13: Esempio di controllore di un parcheggio

Come per il C esistono però dei metodi che permettono una gestione più efficiente del controllore rispetto all'uso della syncronized; questi metodi sono:

- `public final void notify()`: che risveglia un singolo thread in attesa.
- `public final void notifyAll()`: risveglia tutti i thread in attesa.
- `public final void wait() throws InterruptedException`: pone il thread in attesa di un *notify*. Quando un thread viene posto in uno stato di wait esso rilascia il lock acquisito e lo riacquista al suo risveglio.

```
1   public class CarParkControl {
2           private int space;
3           private int capacity;
4
5           public CarParkControl (int n) {
6                   capacity = space = n;
7           }
8           synchronized public void arrive() throws InterruptedException {
9                   while (space == 0) wait();
10                  --space;
11                  notifyAll();
12          }
13          synchronized public void depart() throws InterruptedException {
14                  while (space == capacity) wait();
15                  ++space;
16                  notifyAll();
17          }
18  }
```

Codice 14: Esempio di variabili condizionali in Java

Si può ridurre l'overhead dovuto al contex-switching sostituendo la `notifyAll` con la `notify`. Tale meccanismo può essere usato per migliorare le performance quando si è certi che almeno un thread è in atteso per eseguire un lavoro.
Alla condizione di *wait* è possibile associare un timer molto utile per migliorare la longevità del sistema in quanto tende a risolvere in modo automatico i deadlock.

5 Comunicazione

La comunicazione tra processi è il cuore di tutti i sistemi distribuiti, infatti, non ha senso studiare i sistemi distribuiti senza esaminare come i processi su posti macchine diverse si scambiano le informazioni. La comunicazione nei sistemi distribuiti si basa sempre sullo scambio dei messaggi a basso livello come fornito dalla rete sottostante, anche se ciò rende la realizzazione del sistema distribuito molto complicata.
In questo capitolo analizzeremo prima di tutto le regole che i diversi processi devono rispettare per comunicare tra loro, queste regole sono conosciute anche come protocolli e solitamente vengono strutturati a livelli.
Analizzeremo in seguito tre modelli di comunicazione molto diffusi, le chiamate a procedure remote (RPC, *remote procedure call*), i middleware orientati agli oggetti (MOM, *message-oriented middleware*) e gli *streaming* di dati. Ed infine analizzeremo il problema dell'invio di dati a destinatari multipli, ovvero, il *multicast*.

5.1 Il modello OSI

A causa della mancanza di una memoria condivisa tutta la comunicazione nei sistemi distribuiti avviene mediante tramite l'invio e la ricezione di messaggi a basso livello. Quando un processo A vuole comunicare con un processo B prima di tutto costruisce un messaggio nel proprio spazio degli indirizzi e poi effettua una *chiamata di sistema* che fa in modo che il SO si occupi dell'invio del messaggio attraverso la rete fino a raggiungere B. Anche se il principio è semplice esistono diversi ostacoli al completamente di questa operazione, prima di tutto A e B devono concordare sul significato dei bit inviati, esistono molti altri aspetti sui quali bisogna accordarsi, come il valore in volt usati per indicare un bit a 1, individuare l'ultimo bit del messaggio, bisogna inoltre capire se un messaggio è stato danneggiato o perso ecc.
Per poter trattare facilmente i numerosi aspetti di una comunicazione la *international standard organization* (ISO) ha sviluppato un modello di riferimento che identifica i vari livelli di comunicazione coinvolti, gli assegna dei nomi standard e identifica le diverse funzionalità per ogni livello. Questo modello è chiamato **open system interconnection reference model** o più comunemente modello **ISO OSI** ed è illustrato in Figura 8 Tuttavia è bene far presente che i protocolli sviluppati nel modello OSI non sono stati mai ampiamente utilizzati, tuttavia il modello sottostante si è rivelato particolarmente utile per comprendere le reti di computer.
Il modello OSI è progettato per consentire la comunicazione tra sistema aperti ovvero tra sistemi preparati per comunicare tramite regole standard che ne regolano il formato, i contenuti ed il significato di messaggi ricevuti ed inviati. Queste regole sono dette **protocolli** e devono essere concordate a priori per permettere la comunicazione tra gruppi di computer.
Esistono due grandi tipologie di protocolli, quelli **orientati alla connessione** nei quali mittente e destinatario stabiliscono esplicitamente una connessione prima di scambiarsi dei dati ed alla fine devono rilasciare tale connessione. Con i protocolli **senza connessione** non è necessaria alcuna premessa, quando il messaggio è pronto il mittente invia il messaggio come nel caso di un invio di una lettera.
Nel modello OSI la comunicazione è suddivisa in sette livelli o *layer*, ogni livello tratta uno specifico aspetto della comunicazione, in modo da suddividere il problema in parti gestibili ciascuna delle quali può essere trattata indipendentemente. Per realizzare questo meccanismo ogni livello fornisce un'interfaccia al livello superiore, la quale specifica un insieme di operazioni che il livello è pronto a fornire.
Quando un processo A sulla macchina 1 vuole comunicare con un processo B sulla macchina 2

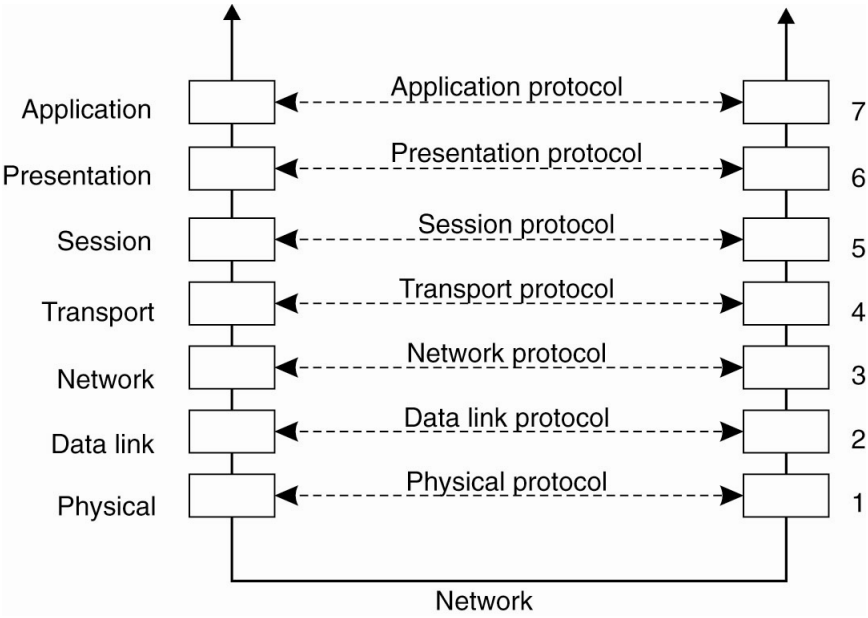


Figura 8: Modello ISO OSI

costruisce un messaggio e lo passa al livello applicativo sulla sua macchina; tale livello potrebbe essere una procedura ad una libreria oppure implementato in qualche altro modo come ad esempio all'interno del sistema operativo. Il software a livello applicativo aggiunge un intestazione (*header*) all'inizio del messaggio e passa tutto al livello di presentazione tramite l'interfaccia tra i livelli 6 e 7. A sua volta il livello di presentazione passa il messaggio al livello di sessione non prima di aver aggiunto il suo *header* e così via. Alcuni livelli aggiungono oltre all'header anche un *trailer* in chiusura al pacchetto come mostrato in Figura 9. Quando il messaggio raggiunge

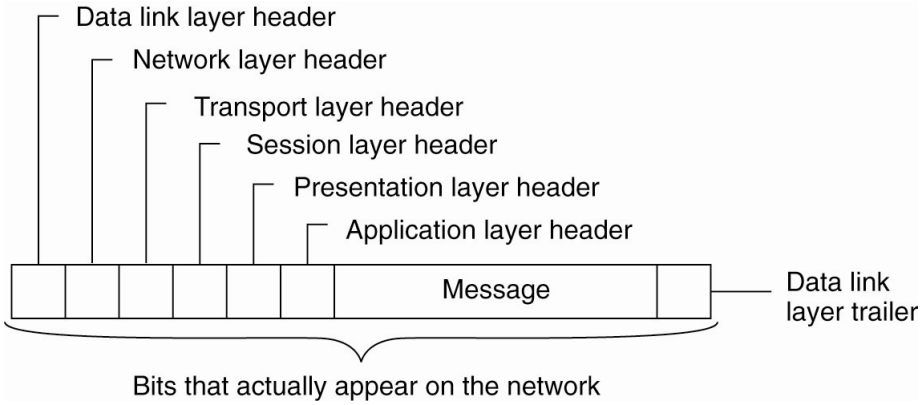


Figura 9: Elenco header e trailer di un messaggio che attraversa i vari layer

il fondo esso viene trasmesso dal livello fisico sul mezzo di trasmissione. Quando il messaggio raggiunge la macchina 2 viene passato verso l'alto e ogni livello stacca la sua intestazione e la esamina, infine il messaggio raggiunge il suo destinatario , ovvero il processo B il quale può rispondere utilizzando il percorso inverso.
Ogni livello ha il suo protocollo che può essere cambiato indipendentemente dagli altri ed è proprio questa indipendenza a rendere i protocolli a livelli interessanti.

L'insieme di protocolli usati in un particolare sistema è detto **suite di protocolli** o **stack di protocolli**.

5.1.1 I layer

Analizziamo ora i diversi layer che compongono il modello OSI, vedremo quali sono le loro funzionalità e dove possibile indicheremo quali protocolli sono attualmente utilizzati nell'ambito di internet. Partiamo dai tre livelli più bassi della suite di protocolli, insieme questi tre livelli implementano le funzioni base di una rete di computer.
Il *livello fisico* trasmette gli 0 e gli 1 sul canale fisico di comunicazione, elementi importanti per questo livello sono la quantità di volt che contraddistingue gli 0 e gli 1, il numero di bit al secondo, la possibilità di trasmettere in entrambe le direzioni, infine rivestono una notevole importanza la forma dei connettori (*plug*) ed il numero di piedini (*pin*). Il protocollo che identifica questo livello ha a che fare con la standardizzazione delle interfacce elettriche e meccaniche e di segnale; un esempio di questi standard è l'interfaccia RS-232-C per la comunicazione seriale.
Il livello *data link* si occupa della rilevazione e della correzione degli errori di trasmissione. Per fare ciò il data link layer raggruppa i bit in unità chiamate **frame** e controlla che ogni frame sia ricevuto correttamente. Per eseguire questo compito il livello di collegamento dei dati applica un *pattern* di bit all'inizio ed alla fine di ogni frame per delimitarli ed eseguire una **somma di controllo** (*checksum*) sommando tramite algoritmi specifici i byte del frame ed inserisce tale somma nei campi all'inizio o alla fine del frame. All'arrivo di un nuovo pacchetto il livello data link esegue la somma sui dati arrivati e la confronta con quella inviata insieme al pacchetto nel caso le due checksum coincidano il pacchetto è considerato corretto, in caso contrario il destinatario ne richiede la ritrasmissione grazie al numero di sequenza inserito nell'header del data link.
Il *livello di rete* si occupa del **routing** dei pacchetti, ovvero, della scelta del percorso ottimale che permetta ad un pacchetto di andare da un mittente ad un destinatario. Il problema si complica in quanto il percorso più breve non è sempre quello ottimo. Al momento il protocollo di rete più utilizzato è l'IP (**Internet Protocol**) senza connessione che è parte dei protocolli Internet. Un **pacchetto** IP può essere inviato senza alcun preparativo. Ogni pacchetto è instradato verso la sua destinazione indipendentemente da tutti gli altri.
Il *livello di trasporto* costituisce l'ultima parte di quelli che possiamo definire *stack del protocollo di rete di base* nel senso che implementa tutti i quei servizi non forniti dall'interfaccia del livello di rete ma necessari per l'implementazione di applicazioni di rete. In altre parole il livello di trasporto in un qualcosa di utilizzabile.
Uno dei compiti del livello di trasporto è quello di fornire una connessione affidabile anche se molte applicazioni gestiscono autonomamente la perdita di pacchetti. Quando arriva un messaggio dal livello applicativo il livello di trasporto lo spezza in parti sufficientemente piccole per essere trasmesse ed assegna un numero di sequenza. Le informazioni nell'header del livello di trasporto riguardano il numero di pacchetti inviati, il numero di pacchetti ricevuti, quali devono essere ritrasmessi e così via.
Connessioni di trasporto affidabili possono essere costruite sopra servizi di rete orientati alla connessione o senza connessione. Nel primo caso i pacchetti arriveranno nella sequenza corretta, nel secondo caso non vi è alcun metodo per stabilire a priori l'ordine di arrivo dei pacchetti, ed è compito del software del livello di trasporto di riordinare i dati. Un aspetto importante del livello di trasporto è quello di fornire questo comportamento *end-to-end*.
Il protocollo di trasporto di Internet è chiamato **TCP (transmission control protocol)**. La combinazione TCP/IP è oggi lo standard de facto per la comunicazione in rete. Oltre al TCP esiste anche un protocollo di trasporto senza connessione chiamato UDP (*universal datagram*

protocol) che è molto simile all'IP con alcune aggiunte minori.
Ulteriori protocolli sono proposti regolarmente, un esempio è il **real-time transport protocol (RTP)** il quale specifica il formato dei pacchetti per il trasferimento dei dati in tempo reale ma non fornisce alcun meccanismo per garantire la loro consegna.
Sopra il livello di trasporto sono il modello OSI identifica altri tre livelli in realtà solo il livello applicativo è sempre usato. Per quanto riguarda i sistemi middleware nel il modello OSI nel l'approccio Internet sono soddisfacenti. Ad esempio il livello di sessione mette a disposizione funzionalità per il controllo del dialogo fornendo la possibilità di impostare *checkpoint* che in caso di *crash* permettano la ripresa della trasmissione senza ricominciare da capo. Tale meccanismo non è mai implementato nella suite di protocolli Internet. Tuttavia nel contesto di soluzioni middleware il concetto di sessione sono piuttosto cruciali. Il compito del livello di prestazione è invece quello di dare un significato ai bit trasmessi. La maggior parte dei messaggi è composta da informazioni strutturate come nomi, indirizzi, quantità di denaro e così via. Nel *livello di presentazione* è possibile definire dei *record* contenenti campi come quelli precedentemente elencati in modo che il mittente possa comunicare al destinatario che il pacchetti contengono dei dati in un certo formato.
Il *livello applicativo* era originariamente inteso per contenere semplici applicazioni di rete come l'e-mail o il trasferimento di file, ora è divenuto il contenitore di tutte le applicazioni. Ciò che manca in questo livello è una netta distinzione tra protocolli di una specifica applicazione e protocolli più generali.

Protocolli middleware Il middleware pur essendo posizionato in maniera logica nel livello applicativo contiene molti protocolli specifici che ne giustificano l'esistenza in un livello proprio. i protocolli per supportare una gran varietà di servizi middleware sono diversi, molti sono pensati per stabilire un autenticazione, non essendo legati ad una specifica applicazione questo tipo di protocolli può essere accorpato in un sistema middleware come servizio generale. Un altro esempio di protocollo che rientra a far parte dei protocolli middleware è quello del *commit distribuito* dove un operazione è portata a termine solo se è portata a termine in tutte le sue parti. Questa proprietà è detta **atomicità**ed è ampiamente utilizzata in tutti i tipi di transizioni.
Come abbiamo visto dai due esempi appena fatti i protocolli middleware supportano servizi di comunicazione di alto livello, ma oltre a questi esistono protocolli per supportare lo *stream* di dati in tempo reale, oppure, protocolli più specifici del livello di trasporto ma che, dovendo tener conto dei requisiti delle applicazioni devono essere situati ad un livello più alto di quello di trasporto come ad esempio il caso di *multicast* che devono garantire la scalabilità.
Il modello che si viene così a creare è quello di Figura 10 nel quale il livello di sessione e quello di presentazione vengono sostituiti da un unico livello middleware contenete quei protocolli indipendenti dalle applicazioni.

5.1.2 Tipi di comunicazione

Esistono diverse alternative nella comunicazione che il middleware mette a disposizione delle applicazioni; partiamo dall'esempio mostrato in Figura 11. In questo caso possiamo pensare che ogni *host* esegua uno *user agent* ovvero un processo che esegue le operazioni di comunicazione tra i vari sistemi.

Prendiamo ora come esempio il caso di un invio di e-mail in questo caso i due *user agent* saranno rispettivamente il sistema che invia e quello che riceve le e-mail. L'agente dal lato del mittente passa la mail al sistema per la consegna delle mail nella convinzione che tale sistema consegni la mail, l'agente dal lato del destinatario a sua volta si connette al sistema per sapere se è giunta qualche nuova mail in caso affermativo la mail viene trasferita dal sistema allo user

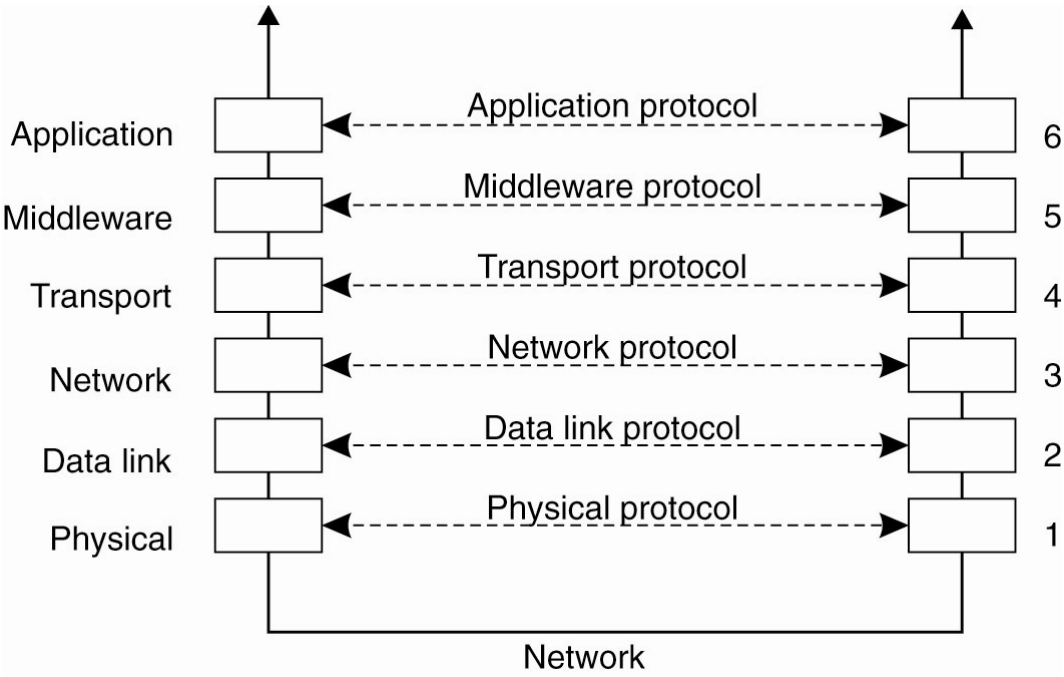


Figura 10: Modello OSI-Middleware

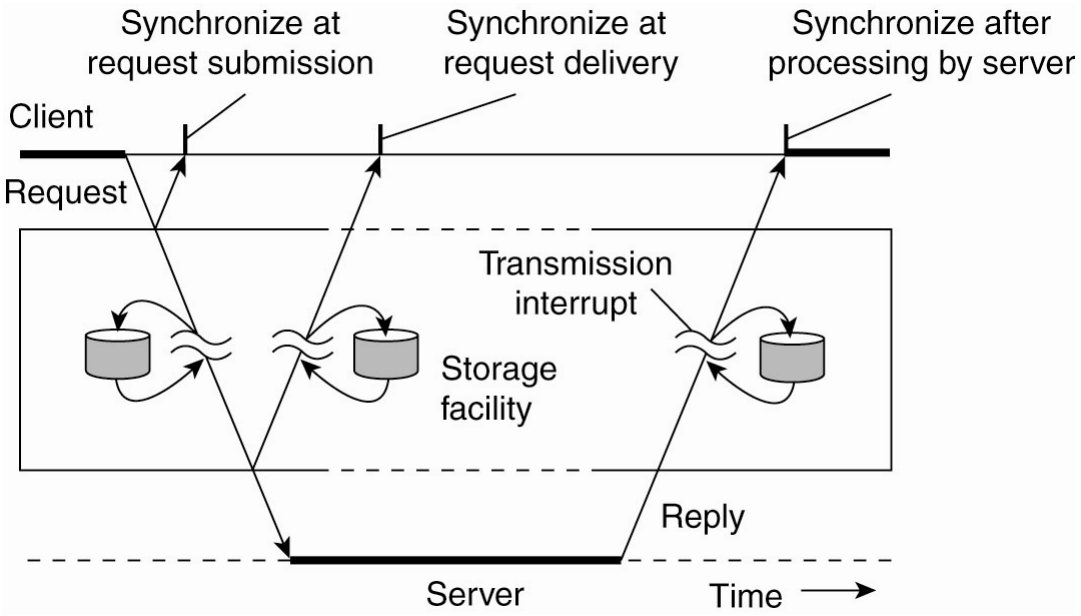


Figura 11: Esempio di tipi di comunicazione

agent del destinatario.
Questo tipo di meccanismo è detto **comunicazione persistente** in quanto il messaggio viene memorizzato dal middleware per tutto il tempo necessario affinché la consegna non vada a buon fine durante questo periodo non è necessario che i due elementi siano in esecuzione contemporaneamente. Nel caso di **comunicazione transiente** invece, il sistema memorizza i messaggi scambiati solo finché sia il mittente che il destinatario sono in esecuzione.
Oltre che persistente o transiente una comunicazione può essere asincrona o sincrona. Si parla di comunicazione **asincrona** quando il mittente continua la sua elaborazione subito dopo l'invio del messaggio. Nel caso di comunicazione **sincrona**, invece, il mittente è bloccato fino a quando la sua richiesta non viene accettata. Ciò può avvenire fino a quando il middleware non comunica la presa in consegna della richiesta, oppure quando la richiesta non viene consegnata al destinatario l'ultima possibilità è che il mittente resta bloccato fino alla fine dell'elaborazione da parte del destinatario e invio di una risposta all'elaborazione.
Esistono molti tipi in cui combinazione persistente, transiente, sincrona e asincrona possono essere combinati ma le più diffuse sono la persistenza e la sincronizzazione, la transiente con la sincronizzazione alla fine dell'elaborazione.
Dobbiamo distinguere, infine tra comunicazione discreta e a *stream*.

5.2 Chiamate a procedure remote

Molti sistemi si basano sullo scambio di messaggi esplicito tra processi, questo scambio avviene tramite l'utilizzo delle procedure `send` e `recive` che però non nascondono del tutto la comunicazione. Una nuova tecnica fu introdotta nel 1984 quando si pensò di permettere ai programmi di richiamare procedure situate su altre macchine. Quando un processo sulla macchina A richiama una procedura sulla macchina B il processo chiamante viene sospeso e ha luogo l'elaborazione sulla macchina B. Le informazioni sono passate dal chiamante al chiamato tramite i parametri e ritornano indietro come risultato della procedura. Questo tipo di meccanismo è conosciuto come **chiamata a procedura remota** o **RPC** (*remote procedure call*).
Sebbene l'idea di base risulti molto semplice i problemi relativi sono molti, primo fra tutti visto che chiamante e chiamata risiedono su due macchine diverse lo spazio degli indirizzi è completamente diverso, inoltre, il passaggio di parametri e dei risultati non è così semplice in quanto le macchine potrebbero avere rappresentazione dei dati differenti.

5.2.1 Operazioni di base sulle RPC

Iniziamo a vedere come funzionano realmente le RPC partendo dall'analisi di una chiamata a procedura locale per poi analizzare le chiamate remote.

Chiamate a procedura locali Per capire come lavorano le RPC è necessario capire comprendere come lavorano le chiamate a procedura convenzionali, ovvero su una singola macchina. consideriamo una chiamate in C tipo:

```
count = read(fd, buf, nbytes)
```

dove *fd* è un intero che indica un file, *buf* è un array di caratteri e *nbytes* è il numero intero di byte da leggere ed immagazzinare in *buf*.Se la chiamata è stata eseguita dal *main* allora lo *stack* della chiamata sarà quello mostrato in Figura 12. Come vediamo nella figura (b) il chiamante inserisce (*push*) i parametri nello stack in ordine inverso. Alla fine dell'esecuzione della procedura il valore di ritorno è posizionato in un registro vengono rimossi i parametri e viene restituito il controllo al chiamante.

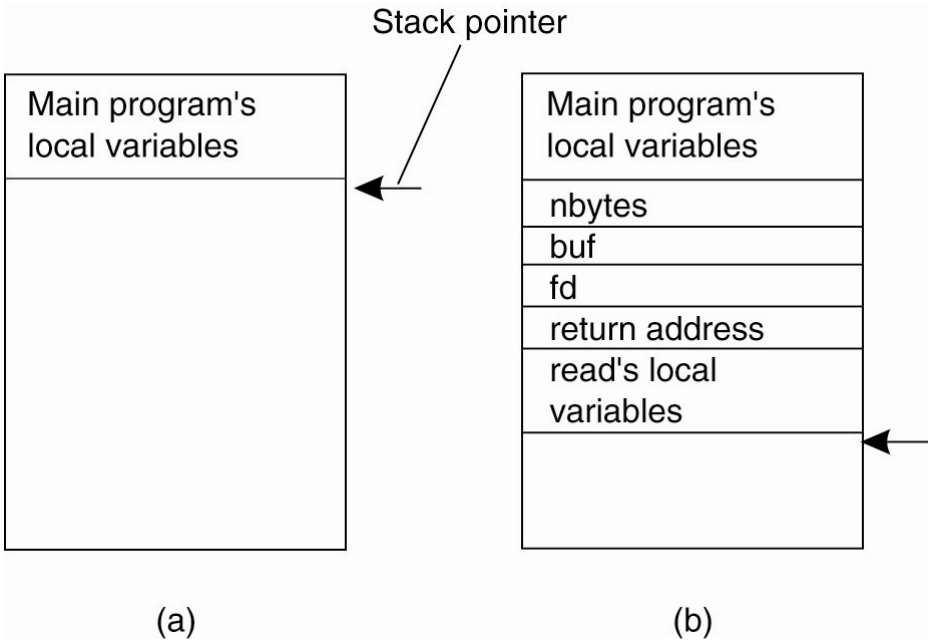


Figura 12: Stack prima e dopo la chiamata di una procedura

Notiamo ora che in C i parametri possono essere passati per **valore** come *fd* e *nbytes* per i quali il valore viene copiato nello stack o per **referenza** come nel caso di *buf* il quale è un puntatore ad un array di char; in questo caso nello stack della chiamata non vi è il valore dell'array ma semplicemente un indirizzo che indica dove l'array è situato. Nel caso in cui la procedura modifichi i valori contenuti nell'array tali cambiamenti saranno effettivi anche all'uscita dalla procedura.
Esiste un terzo meccanismo di passaggio dei parametri anche se non è usato in C ed è quello per **copia/ripristino**, questo passaggio consiste nel copiare il valore della variabile nello stack come nel caso di passaggio per valore, e quindi ricopiarla al termine della chiamata sovrascrivendo il valore originale.

Client e server stub L'idea di base delle RPC è quella di rendere le chiamate a procedura remote il più simile possibile ad una chiamata locale. Vogliamo che le RPC siano trasparenti alla distribuzione. Per ottenere tale trasparenza quando il linker assembla il codice al posto di mettere la versione di sistema della procedura, nel nostro caso la `read`, esso la sostituisce con una versione chiamata **client stub**. Come quella originale anche questa versione viene richiamata usando la sequenza di Figura 12, anche questa esegue una chiamata di sistema, ma a differenza di quella tradizionale questa chiamata impacchetta i dati in un messaggio e ne richiede l'invio al server tramite una `send`. Dopo l'invio il *client stub* richiama la procedura `receive` e si blocca in attesa della risposta come mostrato in Figura 13. Quando il messaggio raggiunge il server esso lo passa al **server sub**, l'equivalente lato server del *client stub*, questo pezzo di codice trasforma la RPC in una chiamata a procedura locale. Il server esegue il proprio lavoro e restituisce il risultato al chiamante. Quando il *server stub* riprende il controllo impacchetta il risultato in un messaggio e richiama la `send` per inviare la risposta al client e si rimette in attesa dell'arrivo di una nuova richiesta con la `receive`. Quando il messaggio di risposta arriva alla macchina client il sistema operativo lo indirizza al *client stub* che lo spacchetta, copia i dati nel buffer e restituisce il controllo al processo client.

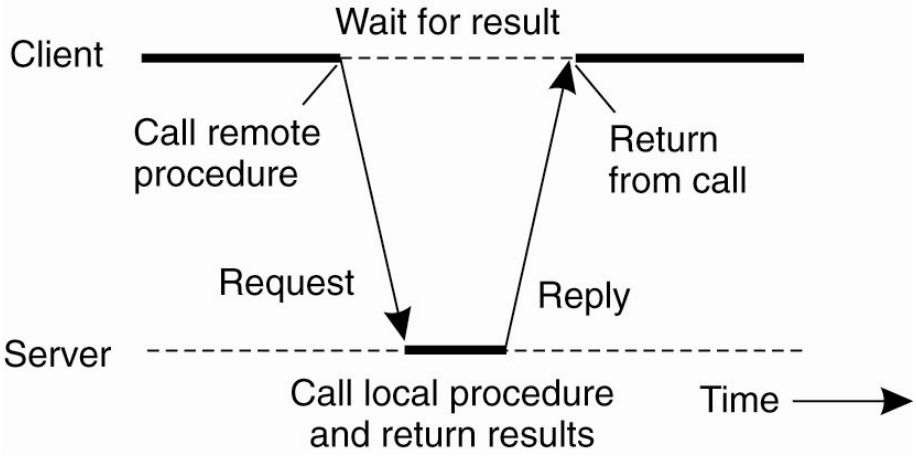


Figura 13: Esempio di chiamata a procedura remota

Quando il client riprende il controllo non ha idea di che cosa sia avvenuto e non ha idea se la chiamata è stata eseguita remotamente o in locale. Ricapitolando una chiamata a procedura remota segue i seguenti passi:

1. la procedura client richiama il *client stub* nel modo normale;
2. il *client stub* costruisce un messaggio e richiama il sistema operativo locale;
3. il SO del client invia il messaggio al SO remoto;
4. il SO remoto passa il messaggio al *server stub*;
5. il *server stub* spacchetta i parametri e richiama il server;
6. il server esegue il lavoro e restituisce il risultato allo *stub*;
7. il *server stub* lo impacchetta in un messaggio e richiama il su SO;
8. il SO del server invia il messaggio al SO del client;
9. il SO del client passa il messaggio al *client stub*;
10. lo *stub* spacchetta il risultato e lo restituisce al client.

5.2.2 Passaggio di parametri

La funzione del client stub è quella di prendere i propri parametri di impacchettarli e di inviarli al server stub. Il problema è che questa operazione pur sembrando molto semplice in realtà non lo è.

Passaggio di parametri per valore L'operazione di impacchettare i parametri in un messaggio è detta **marshaling dei parametri**. Come esempio consideriamo una procedura remota molto semplice come `add(i,j)` che prende in ingresso due valori interi i, j e restituisce la loro somma aritmetica. L'esecuzione di questa procedura è mostrata in Figura 15, nella parte sinistra vediamo la chiamata a tale procedura. Il *client stub* prende i due parametri e li mette in un messaggio semplicemente copiando i valori; insieme ai parametri viene anche inserito il

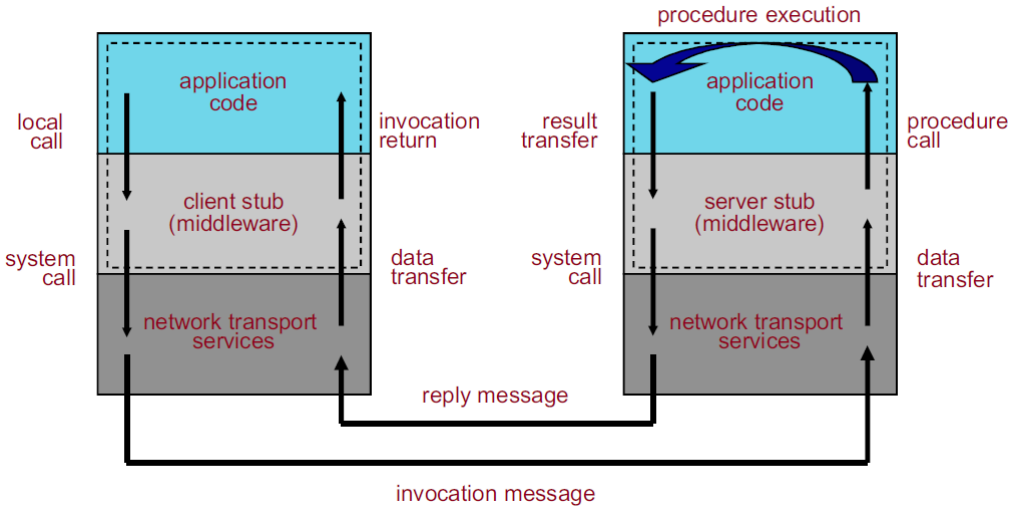


Figura 14: Esecuzione di una chiamata a procedura a procedura remota

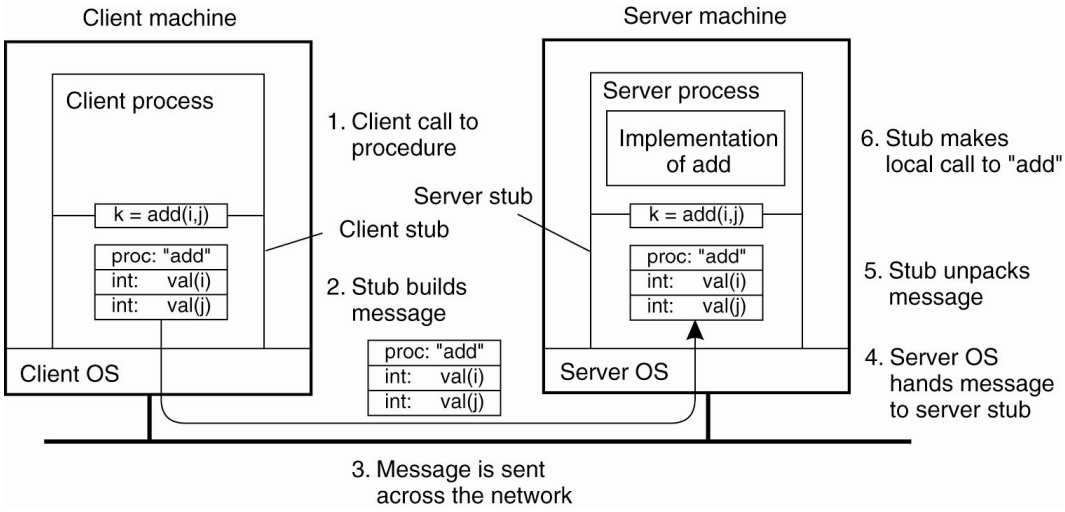


Figura 15: Passaggio di parametri ad una procedura remota

nome (o il numero) della procedura da chiamare in quanto il server potrebbe supportare diverse chiamate. Quando il messaggio raggiunge il server lo *stub* lo esamina per capire quale procedura è richiesta e quindi esegue la chiamata a tale procedura. La chiamata alla procedura è una normale chiamata locale l'unica differenza è che i parametri che vengono passati sono stati estratti dal messaggio e inizializzati nello spazio degli indirizzi dello *stub*.
Quando il server ha completato l'esecuzione il controllo ritorna al *server stub* il quale preleva il risultato e lo inserisce nel messaggio di ritorno al *client stub* il quale effettua *unmarshal* e restituisce il risultato al client.

I problemi iniziano a sorgere quando due macchine non sono dello stesso tipo, ad esempio i *mainframe IBM* usano una codifica *EBCDIC* per i caratteri mentre i personal computer usano l'*ASCII*; problemi simili possono verificarsi con la rappresentazione degli interi o dei numeri in virgola mobile, in quanto alcune macchine utilizzano la numerazione **little endian** (numerazione dei bit da destra a a sinistra) altre invece utilizzano la **big endian** (numerazione da sinistra a destra).

Passaggio di parametri per referenza Abbiamo visto il passaggio di parametri per valore, che a parte i problemi di rappresentazione non solleva grosse problematiche. Vediamo ora invece il passaggio di parametri tramite puntatori o riferimenti in generale.
Un puntatore ha senso solo nello spazio degli indirizzi del processo in cui è usato. Ad esempio nel caso della `read` visto in precedenza il secondo parametro è un puntatore ad un array di caratteri e contiene quindi l'indirizzo al primo elemento dell'array (ad esempio 1000). Se noi passassimo al server tale indirizzo non avrebbe alcun senso in quanto lato server l'indirizzo 1000 potrebbe essere occupato da una istruzione del programma. La strategia più semplice in questo caso è che il client stub, visto che conosce sia la tipologia che la dimensione dell'array, faccia una copia dell'array nel messaggio e lo invii al server stub il quale a questo punto può richiamare la procedura utilizzando come indirizzo quello dell'aria di memoria nella quale il server stub ha posizionato l'array appena giunto. Quando il server termina il server stub copia nuovamente l'array nel messaggio e lo restituisce al client. Quando il messaggio giunge al client stub esso applica le modifiche all'array originale. Il meccanismo di passaggio per riferimento è stato così sostituito da un meccanismo di **copia/ripristino**.
Esistono delle ottimizzazioni per rendere più efficace tale meccanismo, se gli *stub* conoscono se il parametro è un input oppure un output per il server una delle due copie può essere eliminata.
Abbiamo visto come passare un dato di cui conosciamo la dimensione, nel caso più generale di dati come strutture o oggetti i dati possono essere passati come uno stream di byte, questo meccanismo è chiamato *serializzazione*.

Specifica dei parametri e generazione degli stub Come abbiamo visto fino ad ora per nascondere una chiamata remota bisogna che il chiamante e il chiamato concordino sul formato dei messaggi e che eseguano gli stessi passi, in altre parole entrambi i lati di una RPC devono seguire lo stesso *protocollo*. Definire il formato dei messaggi non è però sufficiente è necessario anche che il client e il server concordino sulla rappresentazione delle strutture dati e su come i messaggi debbano essere scambiati, come ad esempio utilizzare un protocollo orientato alla connessione come il TCP/IP.
Una volta che il protocollo è stato definito bisogna implementare gli stub, fortunatamente questi differiscono soltanto per le loro interfacce verso le applicazioni . Per semplificare le cose le interfacce sono spesso definite tramite un **linguaggio per la definizione di interfacce** (IDL *interface definition language*). Un'interfaccia specificata tramite IDL viene successivamente compilata in un *client stub* e in un *server stub* unitamente alle interfacce *compile-time* o *run-time*.
Utilizzare un linguaggio per la definizione delle interfacce semplifica considerevolmente le applicazione client-server basate su RPC.

5.2.3 RPC in pratica

Abbiamo visto fino ad ora la teoria che sta dietro ad una chiamata a procedura remota, vediamo ora invece come viene implementato nella realtà un sistema basato sulle RPC.
Esistono due vie per produrre un sistema che sfrutta le chiamate a procedura remote, il primo è lo standard *de facto* ed è stato introdotto dalla Sun Microsystem. Questo sistema è parte del Network File System dei sistemi UNIX e specifica il formato dei dati tramite un XDR (*eXternal Data Representation*), il protocollo di trasporto utilizzato è indifferentemente il TCP o UDP, il passaggio di parametri è consentito solo tramite copia e con un massimo di un input ed un output per funzione, infine tale meccanismo fornisce dei meccanismi per la criptazione dei dati.
Il secondo meccanismo è quello proposto dall'Open Group ed è chiamato *Distributed Computing*

Environment (DCE) ed è un insieme di specifiche e riferimenti. Esistono diverse specifiche per la chiamata di una procedura:

- **At-Most-Once:** in cui una chiamata non viene portata avanti più di una volta
- **Idempotent:** in questo caso la procedura può essere ripetuta più di una volta senza problemi
- **Broadcast:** marcando una procedura in questo modo la richiesta sarà inoltrata a tutte le macchine della rete

Come primo passo per creare un'applicazione distribuita dopo aver scelto i meccanismi è quello di chiamare i programmi *uuidgen* i quali generano un prototipo di file IDL contente un identificatore d'interfaccia. A questo punto si completano i file IDL compilando i nomi delle procedure remote ed i loro parametri. Una volta completato il file IDL viene richiamato il compilatore IDL che genera un file *header* da includere nelle applicazione, il *client stub* e il *server stub*. A questo punto il programmatore deve sviluppare le due applicazioni client e server le quali verranno poi compilate e linkate insieme alle librerie e ai rispettivi client e server stub come mostrato in Figura 16 L'ultimo problema che dobbiamo affrontare è come *legare* un client al server che implementa la procedura richiesta dal client. Anche in questo caso Sun e DCE affrontano il problema in due modi differenti. Sun introduce un demone chiamato `portmap` che associa un server ad una porta, il server occupa una porta disponibile e comunica la sua scelta al *portmap* il client contatta il portmap e richiede la porta necessaria. Il problema principale è che il portmap risolve solo il problema di come stabilire la connessione al server ma deve già conoscere l'ubicazione del server.
DCE, invece utilizza due demoni, uno situato sulla stessa macchina del server che svolge la stessa funzione del portmap, ed un secondo demone situato su un server differente (*directory server*) il quale implementa la trasparenza alla distribuzione. Il meccanismo con il quale tale sistema opera è illustrato nella Figura 17.

5.2.4 Tipi di chiamate a procedure remote

Fino ad ora abbiamo dato per scontato che le chiamate a procedura remote fossero solamente di tipo sincrono con sincronizzazione alla fine dell'esecuzione della procedura remota. In alcuni casi, però, tale meccanismo non è efficiente in quanto spesso si sprecano risorse lato client. Una possibile variante sono le RPC asincrone le quali possono essere utilizzate quando non sono attesi valori di ritorno e lo sblocco del client avviene all'arrivo di un acknowledgment quando la richiesta viene presa in carico oppure all'arrivo di una *promessa* di risposta che verrà inviata tramite un'altra RPC asincrona dal server.
La Sun inoltre ha implementato delle RPC *batched* che sono particolari chiamate a procedura che non si aspettano un risultato, queste sono immagazzinate (*buffered*) dal client ed inviate in un unica comunicazione quando si presenta una chiamata non-batched.
Un esempio simile avviene nei dispositivi mobili; quando l'host è disconnesso immagazzina le richieste e periodicamente tenta l'invio di richieste, lo stesso viene fatto dal server per l'invio delle risposte. Le richieste e le risposte sono inviate/ricevute su due canali differenti.

5.2.5 Remote method invocation

Il *Remote Method Invocation* (RMI) è simile alle RPC ma è applicato alla programmazioni ad oggetti, un'importante differenza è che i riferimenti oggetti remoti possono essere passati. L'IDL

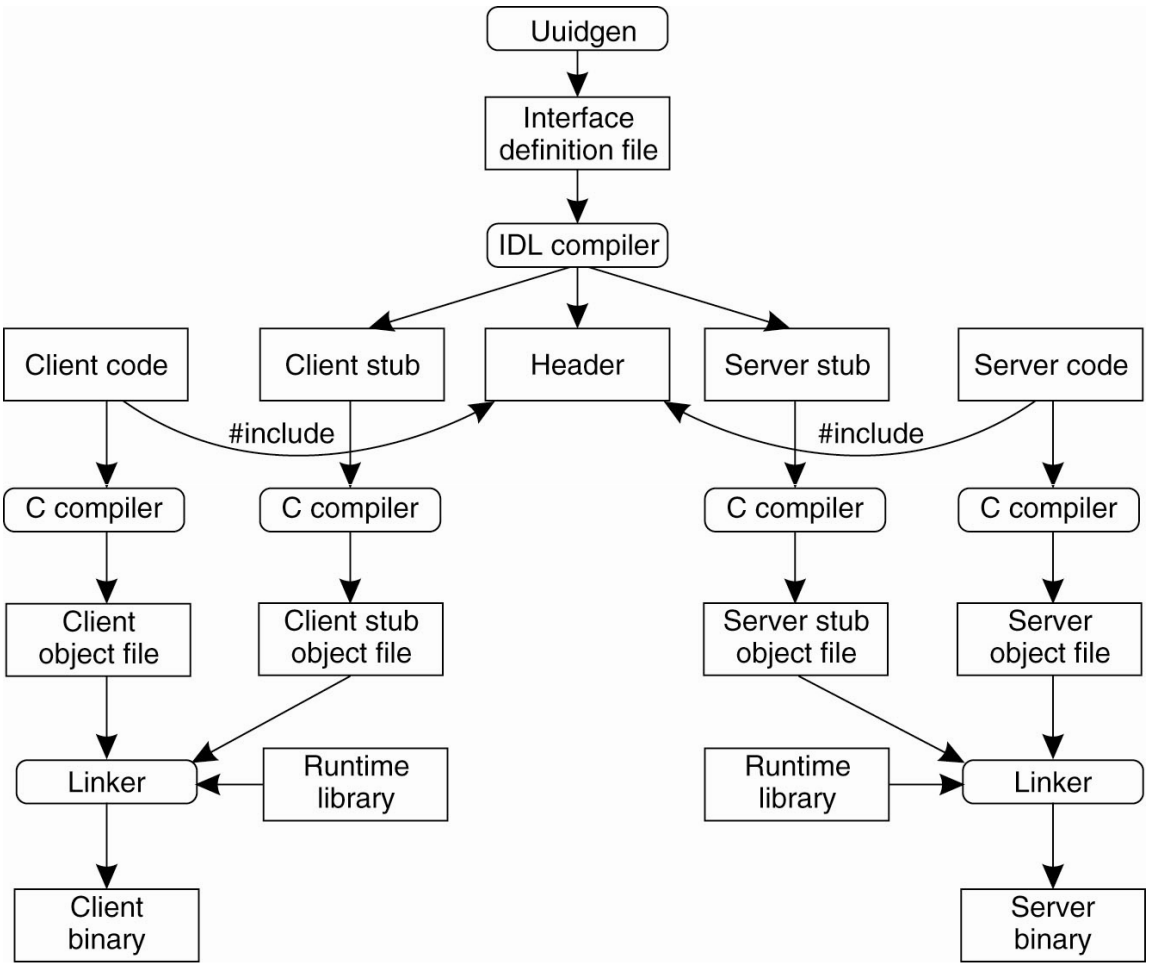


Figura 16: Flusso di sviluppo di un'applicazione distribuita

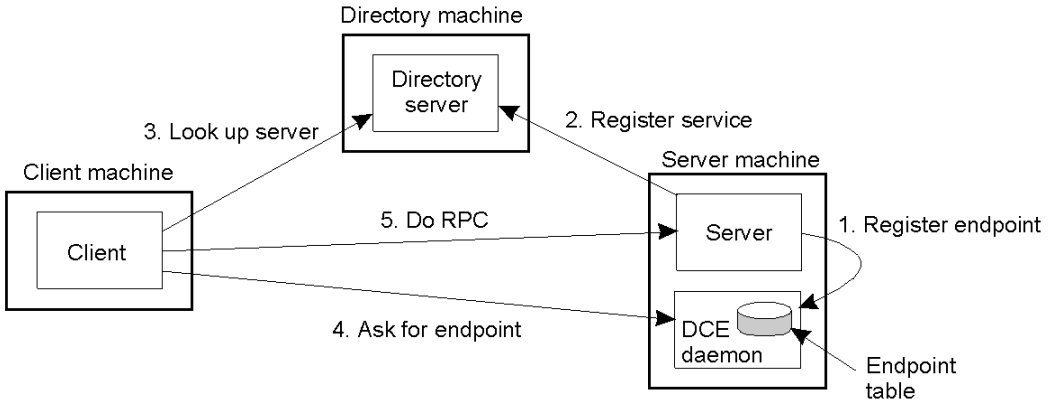


Figura 17: Esecuzioni di un binding tra client e server in DCE

nel caso di RMI è molto più completo includendo eccezioni ed altro in quanto la separazione tra interfaccia ed implementazione è alla base della programmazione ad oggetti.
Le più importanti tecnologie che implementano l'RMI sono *Java RMI* che può essere utilizzata solo implementando in Java ed utilizzando una Java Virtual Machine, è molto semplice e permette il passaggio di parametri sia per riferimento che per valore. *OMG CORBA* è multilinguaggio e multipiattaforma, anche *CORBA* permette il passaggio di parametri sia per riferimento che per valore ma nel secondo caso è compito del programmatore garantire la stessa semantica sia lato server che lato client.

5.3 Comunicazione orientata agli oggetti

Le chiamate a procedure remote o le RMI aiutano a nascondere la comunicazione nei sistemi distribuiti, ovvero forniscono un certo grado di trasparenza all'accesso. Tuttavia esistono dei casi in cui questi meccanismi non sono appropriati, come ad esempio nel caso in cui non si è certi che il destinatario sia attivo nell'istante di invio di un messaggio.La soluzione a queste problematiche sono i *messaggi*.

5.4 Comunicazione orientata ai messaggi

Le RPC come abbiamo visto hanno di per se una natura sincrona ed è necessario che sia mittente che destinatario siano in esecuzione all'invocazione della procedura. In molti casi però questo non è possibile o non è conveniente perciò le applicazioni vengono sviluppate utilizzando l'invio di messaggi.

5.4.1 Comunicazione transiente orientata ai messaggi

Molti sistemi e applicazioni distribuite sono costruite direttamente in base al semplice modello orientato ai messaggi fornito dal livello di trasporto. Tali applicazioni sono sviluppate attraverso l'utilizzo delle *socket*.

Berkeley socket Un'attenzione particolare è stata messa nella standardizzazione dell'interfaccia del livello di trasporto in modo da consentire ai programmatori di far uso dell'intera suit

Primitiva	Significato
`Socket`	Crea una nuova porta di comunicazione
`Bind`	Collega un indirizzo locale ad una socket
`Listen`	Annuncia la disponibilità ad accettare connessioni
`Accept`	Blocca il chiamante finché la richiesta di connessione non arriva
`Connect`	Cerca attivamente di stabilire una connessione
`Send`	Invia dati sulla connessione
`Receive`	Riceve dati sulla connessione
`Close`	Rilascia la connessione

Tabella 1: Primitive delle socket

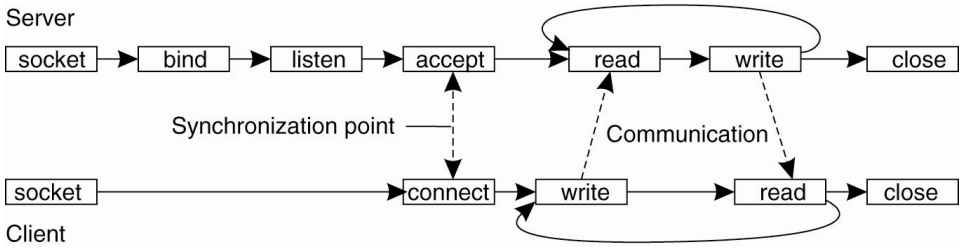


Figura 18: Schema generale di una connessione socket

di protocolli attraverso un semplice insieme di primitive. Una di queste interfacce è **l'interfaccia socket**introdotta in UNIX Berkeley, un'altra interfaccia importante è la **XTI (X/open transport interface)**. Le *socket* e XTI sono molto simili nel modello di programmazione ma differiscono nelle primitive.
Una socket è una porta di comunicazione su cui un'applicazione può scrivere dati da inviare e da cui poter leggere le risposte. Le primitive delle socket sono mostrate in Tabella 1 In generale i server eseguono le prime quattro primitive nell'ordine in cui sono presentate. Quando viene richiamata la primitiva `socket` il chiamante crea una nuova porta di comunicazione ed il sistema operativo locale riserva tale risorsa per la ricezione e l'invio di dati. La primitiva `bind` associa un indirizzo locale con la socket appena creata, il collegamento (*binding*) indica al sistema operativo che il programma vuole ricevere messaggi solo all'indirizzo e sulla porta specificati. La primitiva `listen` viene richiamata solo nel caso di comunicazione orientata alla connessione, serve per comunicare al SO quante risorse riservare in base al numero di connessioni massime che il server intende accettare. Tramite la primitiva `accept` il server si blocca fino all'arrivo di una richiesta di connessione, al suo arrivo il sistema operativo locale crea una nuova *socket* con le stesse proprietà dell'originale e la restituisce al chiamate. Questo meccanismo permette al server di effettuare una *fork* e gestire una comunicazione mentre attende una nuova connessione. Lato client il procedimento è leggermente differente, anche qui è necessario creare la risorsa tramite la primitiva `socket` ma il *binding* esplicito di tale risorsa non è necessario. La primitiva `connect` richiede che il chiamante specifichi l'indirizzo a cui va inviata la richiesta di connessione; il client è bloccato fino a quando la connessione non è stata stabilita dopo di che entrambi i lati possono iniziare a scambiarsi messaggi tramite le primitive `send` e `receive`. Nelle socket la connessione è simmetrica perciò la chiusura della connessione si ha quando sia il server che il client richiamano la primitiva `close`. Il modello generale di una connessione socket è mostrato in Figura 18.

Primitiva	Significato
MPI_bsend	Aggiunge un messaggio in uscita a un buffer per l'invio locale
MPI_send	Invia un messaggio e aspetta finché non viene copiato in un buffer
MPI_ssend	Invia un messaggio e aspetta finché non inizia la ricezione
MPI_sendrecv	Invia un messaggio e aspetta la risposta
MPI_isend	Invia il riferimento a un messaggio in uscita e continua
MPI_issend	Invia il riferimento a un messaggio e aspetta finché non inizia la ricezione
MPI_recv	Riceve un messaggio; si blocca se non ce ne sono
MPI_irecv	Controlla se ci sono messaggi in ingresso, ma non si blocca

Tabella 2: Elenco delle primitive MPI

Interfaccia per lo scambio di messaggi Con l'arrivo dei computer ad alte prestazioni gli sviluppatori hanno avuto bisogno di primitive orientate ai messaggi che consentissero la scrittura di applicazioni in modo facile ed efficiente. Queste caratteristiche fanno si che le primitive siano ad un livello di astrazione sufficientemente alto per sviluppare le applicazioni in modo veloce ma che richiedessero un incremento minimo rispetto alle *socket*, le quali non potevano essere sfruttate in quanto avevano un livello di astrazione troppo basso, ed erano state pensate solo per la comunicazione in rete tramite TCP/IP.
Serviva qualcosa che di adatto a reti ad alta velocità come quelle usate nei *cluster* ed inoltre serviva un'interfaccia più avanzata che potesse gestire funzionalità più avanzate come diverse forme di *buffering* e di sincronizzazione.
Si è arrivati così alla definizione di uno standard chiamato semplicemente **interfaccia per lo scambio di messaggi** o **MPI** (*message-passing interface*). Nato come progetto per le applicazioni parallele e come tale adatto alla comunicazione transiente, utilizza la rete sottostante e presuppone che eventi come la caduta di un processo siano fatali.
L'MPI suppone che la comunicazione avvenga tra un gruppo di processi conosciuti, a cui viene assegnato un identificatore del tipo (*groupID,processID*) che identifica univocamente un processo e che viene utilizzato in sostituzione dell'indirizzo del livello di trasporto.
Il cuore di MPI è costituito da primitive per lo scambio di messaggi per supportare la comunicazione transiente, esse sono elencate nella Tabella 2 LA comunicazione transiente asincrona è supportata dalla primitiva `MPI_bsend`, il mittente invia un messaggio che viene copiato localmente dal sistema runtime di MPI, quando il messaggio è stato copiato il mittente continua il proprio lavoro. Il sistema runtime locale cancellerà il messaggio dal suo buffer e si occuperà della trasmissione non appena un destinatario chiamerà una `receive`.
Esiste anche un'operazione di invio bloccante, la `MPI_send` la quale blocca il chiamante fino a quando il messaggio non è stato copiato nel buffer runtime del destinatario oppure fino a quando il destinatario non ha iniziato la ricezione. La comunicazione sincrona nella quale il mittente si blocca fino a quando la sua richiesta non è accettata avviene tramite la `MPI_ssend`, infine, è disponibile anche una comunicazione fortemente sincrona nella quale il sistema attende fino all'esecuzione della richieste e ricezione della relativa risposta che avviene tramite la primitiva `MPI_sendrecv`; quest'ultima corrisponde ad una normale RPC.
Le due primitive `MPI_send` e `MPI_ssend` hanno delle varianti che evitano di copiare il messaggio nel buffer locale, queste varianti corrispondono ad un tipo di comunicazione asincrona e si ottengono tramite la `MPI_isend` tramite il quale il mittente invia un puntatore ad un messaggio dopodiché il sistema runtime di MPI si occupa della comunicazione mentre il mittente prosegue con la sua esecuzione. La primitiva `MPI_issend` fornisce la sicurezza che il destinatario abbia accettato il messaggio e stia lavorando sulla richiesta.

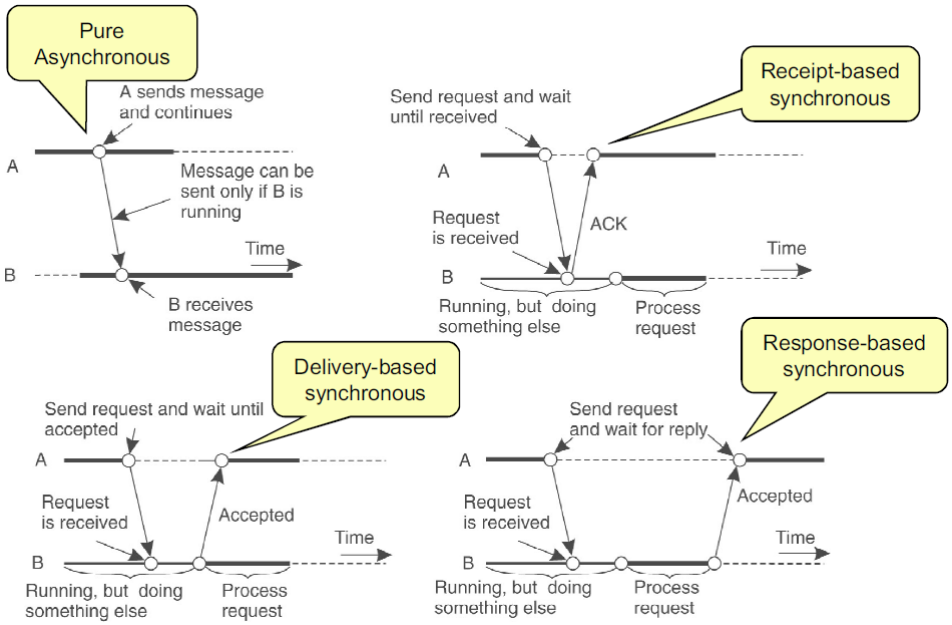


Figura 19: Esempi di comunicazione di tipo transiente

Le operazioni `MPI_recv` e `MPI_irecv` sono rispettivamente le primitive bloccanti e non bloccanti usate per la ricezione dei messaggi.
I diversi tipi di comunicazione sono mostrati in Figura 19.

5.4.2 Comunicazione persistente orientata ai messaggi

La classe più importante di servizi middleware orientati ai messaggi è quella conosciuta come **sistemi a code di messaggi** o semplicemente **middleware orientati ai messaggi** (MOM, *message oriented middleware*). Questo tipo di sistemi forniscono un ampio supporto alla comunicazione asincrona persistente. La caratteristica fondamentale di questi sistemi è che permettono di memorizzare i messaggi senza che il mittente e il destinatario siano attivi durante la trasmissione del messaggio.

Modello a code per lo scambio di messaggi L'idea di base è che le applicazioni comunicano inserendo i messaggi in code specifiche, questi messaggi vengono poi inoltrati tramite una serie di serve e alla fine consegnati al destinatario.
La particolarità è che un mittente non ha la garanzia che un messaggio venga effettivamente letto ma ha solo la garanzia che tale messaggio sarà inserito prima o poi nella coda dei messaggi del destinatario.
Questo aspetto fa si che vi sia un forte disaccoppiamento nelle comunicazioni e non vi è quindi la necessità che il destinatario sia in esecuzione quando viene inviato il messaggio, analogamente non vi è la necessità che il mittente sia in esecuzione quando il messaggio viene prelevato dal destinatario come mostrato in Figura 20. Questo significa che un messaggio rimarrà in coda fino a quando non verrà rimosso senza tenere conto se il mittente o il destinatario sono in esecuzione.
I tipi di messaggi che possono essere inviati sono infiniti, l'unica restrizione è che essi siano opportunamente indirizzati; per l'indirizzamento viene fatto fornendo un nome univoco a livello del sistema della coda di destinazione. La semplicità di questa architettura si rispecchia anche

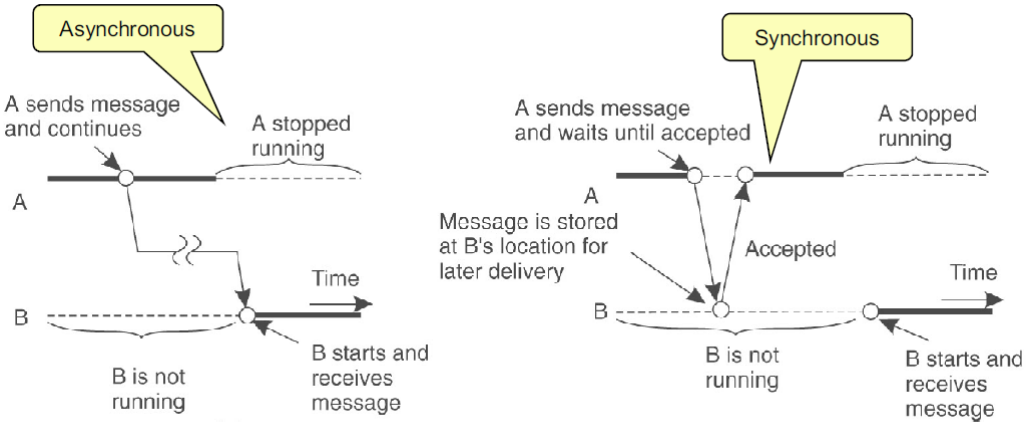


Figura 20: Esempio di comunicazione persistente orientata ai messaggi

Primitiva	Significato
`Put`	Aggiunge un messaggio ad una coda specifica
`Get`	Preleva il primo messaggio in coda nel caso sia vuota si blocca
`Poll`	Preleva il primo messaggio in coda ma non è bloccante
`Notify`	Installa un handler da chiamare quando arriva un messaggio nella coda

Tabella 3: Interfaccia di base per una coda in un sistema a code di messaggi

nella sua interfaccia che è mostrata in Tabella 3. La primitiva `put` viene richiamata da un mittente per passare un messaggio al sistema sottostante e posizionarlo in una coda specifica. La `get` è una chiamata bloccante attraverso la quale un processo autorizzato può rimuovere dalla coda specificata il messaggio pendente da più tempo; nel caso la coda sia vuota il processo viene bloccato. La corrispettiva chiamata non bloccante è la `pull`. Molti sistemi inoltre permettono l'installazione di un *handler* sotto forma di *funzione callback* che viene richiamata all'arrivo di ogni nuovo messaggio; molto utili nel caso si volesse avviare un processo per la ricezione dei messaggi.

Architettura generale di un sistema a code In realtà la vera architettura di un sistema a code di messaggi ha alcune restrizioni; prima fra tutti è che i messaggi possono essere inseriti solo in code *locali* al mittente ovvero code sulla stessa macchina o comunque sulla stessa LAN raggiungibile in modo *efficiente* tramite una RPC. Una coda di questo tipo è chiamata **coda sorgente**. Analogamente i messaggi possono essere letti soltanto da code locali dette **code di destinazione**. È responsabilità del sistema fare in modo che coda sorgente e coda destinazione siano disponibili e che i messaggi vengano recapitati nel modo corretto.
L'insieme delle code è distribuito su molte macchine perciò il sistema deve mantenere una corrispondenza tra le code e la loro posizione, questo significa mantenere una base di dati di **nomi delle code** e delle rispettive posizioni sulla rete come mostrato in Figura 21 Questo meccanismo è molto simile all'uso del *domain name system* (DNS). Le code sono gestite dai **gestori delle code** i quali interagiscono direttamente con le applicazioni che inviano e ricevono messaggi; esistono però gestori speciali che agiscono da *router* o **relay** che inoltrano messaggi in ingresso ad altri gestori. in questo modo un sistema a code può crescere e diventare una rete ***overlay*** completa basata un una rete di computer esistente.

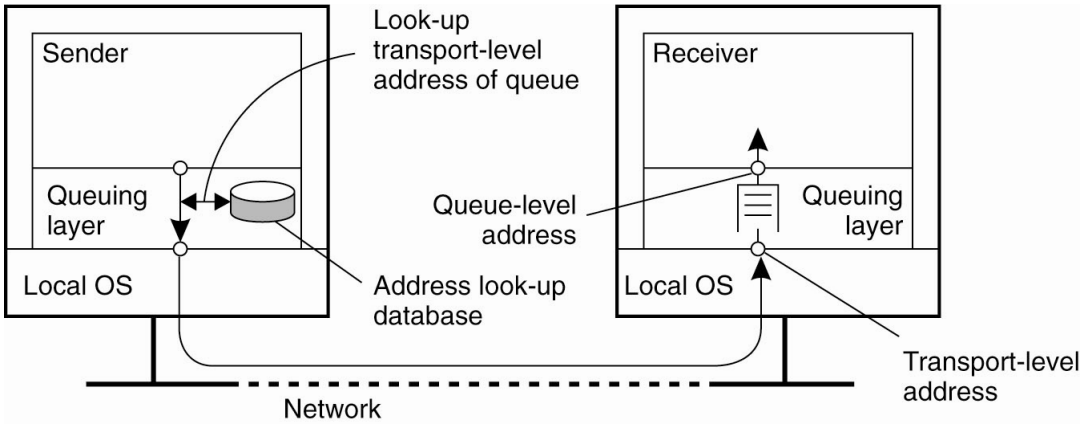


Figura 21: Esempio di uso di basi di dati per i nomi delle code

I *relay* possono essere utili per molti motivi, nei sistemi nei quali non è possibile mantenere un *naming server* generale il quale possa mantenere dinamicamente la corrispondenza coda-posizione e la topologia della rete è statica ed ogni gestore delle code deve mantenere una copia della corrispondenze coda-posizione allora in questo caso possono verificarsi problemi di gestione della rete. Una soluzione possibile è quella di utilizzare dei *router* i quali conoscono la topologia della rete. Quando un mittente A inserisce un messaggio per un destinatario B allora il messaggio è trasferito al più vicino *router* chiamato $R1$ come si vede nella Figura 22. A questo punto il router $R1$ potrebbe dedurre da alcune informazioni contenute nel messaggio che la direzione verso la quale inoltrare il messaggio è quella di $R2$. In questo modo è necessario solo aggiornare i router su quali code vengono aggiunte o eliminate mentre i gestori devono solo preoccuparsi di individuare il router più vicino.
I *relay* agevolano la costruzione di sistemi a code di messaggi scalabili, tuttavia al crescere della rete diventa impossibile la gestione manuale della rete, la soluzione è quella di adottare sistemi di *routing* dinamici come avviene nelle reti di computer.
Un altro vantaggio dei *relay* è quello di consentire un'elaborazione secondaria dei messaggi per ragioni di sicurezza e tolleranza ai guasti; oppure per trasformare i messaggi in una forma comprensibile al destinatario, in questo caso parliamo di *gateway*; infine, i relay possono essere usati per effettuare il *multicasting*.

Broker di messaggi Una caratteristica importante dei sistemi a code di messaggi è la possibilità di integrare applicazioni nuove con alcune già esistenti in un unico sistema distribuito. L'integrazione richiede che tutte le applicazioni del sistema possano interpretare i messaggi che ricevono, questo implica che i messaggi in uscita da un'applicazione siano nello stesso formato di quelli del destinatario.
Questa caratteristica non è sempre garantita da tutte le applicazioni. Una soluzione è quella di concordare un *protocollo* comune per tutte le applicazioni ma sfortunatamente tale meccanismo non funziona con i sistemi a code di messaggi in quanto il livello di astrazione è troppo alto e un meccanismo di questo genere ha senso solo se le informazioni da scambiare sono molte.
L'altro approccio è quello di provare a convivere con tutti questi formati e fornire un meccanismo di conversione il più semplice possibile. Nei sistemi a code di messaggi tale conversione è gestita da alcuni nodi speciali chiamati **broker di messaggi**. Un *broker* agisce da *gateway* a livello applicativo, il suo obiettivo è quello di convertire i messaggi in ingresso in modo che siano comprensibili dall'applicazione destinataria.

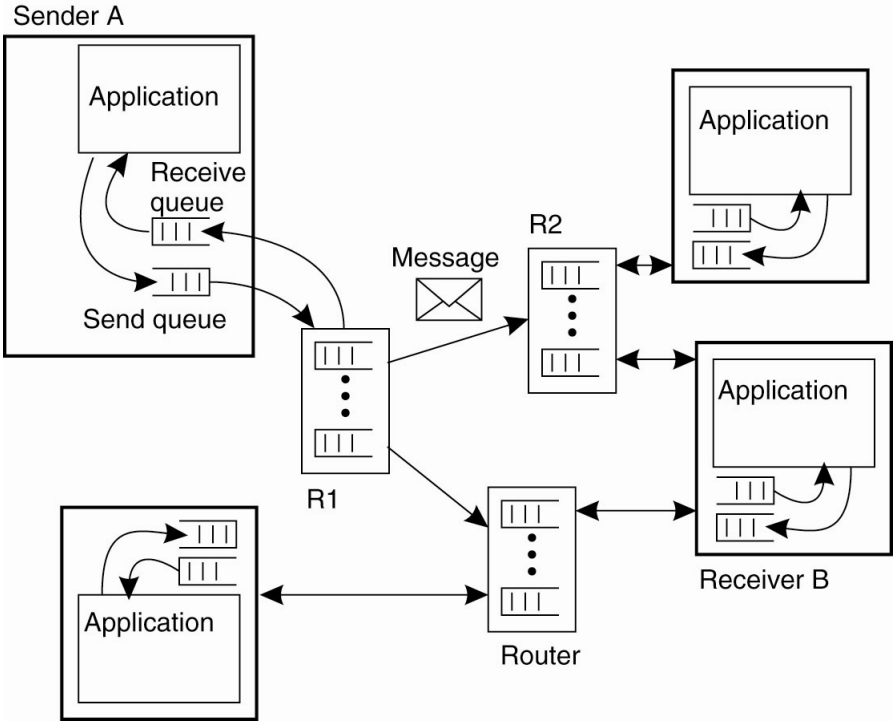


Figura 22: Organizzazione generale dei sistemi a code di messaggi con *router*

In un sistema a code di messaggi un *broker* non è altro che un'applicazione come si può vedere anche da Figura 23.

La funzione più comune di un *broker* è quella di effettuare l'**integrazione di applicazioni aziendali** (EAI,*enterprise application integration*), in questo caso oltre a convertire i messaggi il broker deve trovare una corrispondenza tra le applicazioni basandosi sui messaggi che si sono scambiati. In tale modello, chiamato **publish/subscribe**, le applicazioni inviano messaggi sotto forma di *publishing* riguardo ad un determinato argomento X al broker, le applicazioni che hanno dichiarato il loro interesse all'argomento X tramite una *subscription* riceveranno questi messaggi dal broker. Esistono due tipologie di sottoscrizioni, la *subject-based* nella quale l'argomento è determinato a priori, oppure, la *content-based* nella quale la sottoscrizione contiene un'espressione (filtro) che permette il filtraggio degli eventi in base al loro contenuto.
Il cuore di un broker è il *repository* delle regole e dei programmi che consentono di trasformare un messaggio del tipo $T1$ in un messaggio del tipo $T2$.

5.4.3 Event dispatcher

L'*event dispacher* è quel componente che, in un sistema basato sugli eventi, raccoglie le sottoscrizioni e distribuisce gli eventi ai vari client. Questo componente può essere centralizzato, oppure, per ragioni di scalabilità, distribuito; nel caso distribuito un insieme di broker sono organizzati in una rete *overlay* cooperante per raccogliere le sottoscrizioni, la tipologia della rete può essere ciclica o aciclica.
Nel caso di rete aciclica ogni broker immagazzina le sottoscrizioni dei client a lui collegati, i messaggi sono scambiati tra i diversi broker e inoltrati ai client solo se si sono sottoscritti. Per quanto riguarda le sottoscrizioni ogni broker inoltra le sottoscrizioni agli altri ma non viene mai mandata la stessa sottoscrizione due volte sullo stesso collegamento. I messaggi seguono le

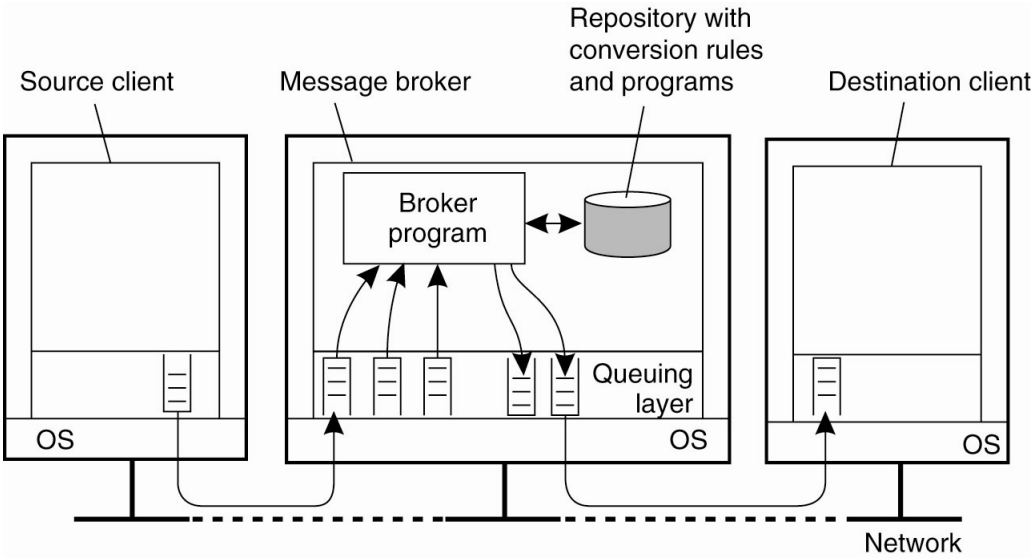


Figura 23: Esempio generale di un broker

stesse vie delle sottoscrizioni. In particolare ogni volta che un broker riceve un messaggio esso effettua un matching con un lista di filtri per determinare una lista di inoltro. L'efficienza di questo meccanismo dipende dalla complessità del linguaggio di sottoscrizione e dall'algoritmo di forwarding utilizzato. Il più comune nel caso aciclico è il forwarding di tipo gerarchico; messaggi e sottoscrizioni sono inoltrati verso la radice dell'albero, i messaggi ridiscendono solo se esiste un matching nelle sottoscrizioni lungo il percorso intrapreso.
Nel caso di grafo ciclico un sistema di tipo DHT (*distributed hash table*) organizza i nodi in una struttura nella quale il routing è efficiente e avviene nei nodi nei quali si ha un ID più piccolo o al più uguale ad un altro ID. Le sottoscrizioni a dei messaggi aventi un certo soggetto S vengono così processate:

- Si calcola l'hash (Hs) dell'argomento S
- Si utilizza la DHT per inoltrare la sottoscrizione a *succ(Hs)*
- Durante l'inoltro della sottoscrizione verso $succ(Hs)$ si comunicano delle informazioni ai nodi attraversati per il successivo inoltro dei messaggi.

La pubblicazione dei messaggi aventi un certo argomento S vengono così inoltrati:

- Si calcola l'hash dell'argomento Hs
- Si utilizza la DHT per inoltrare il messaggio al nodo $succ(Hs)$
- Durante l'inoltro verso $succ(Hs)$ utilizza le informazioni lasciate dalla sottoscrizione per inoltrare il messaggio ai nodi interessati.

Per quanto riguarda sistemi *content based* esistono diversi meccanismi sia per il *forwarding* sia per *routing*, per quanto riguarda il forwarding le tecniche utilizzate sono:

- Per Source Forwarding (PSF)
- Improved per source forwarding (iPSF)

- Per receiver forwarding (PRF)

Mentre per quanto riguarda il routing solitamente si utilizzano:

- Distance Vector (DV)
- Link-State (LS)

PSF: Per-Source Forwarding Nel caso del PSF si parte da un nodo e si calcola l'albero a cammino minimo; ottenuto quest'albero si riempe una tabella associata ad ogni nodo in questo modo con un campo *sorgente* un campo *successivo* ed un campo *predicato* come mostrato in Figura 24. Come *sorgente* si utilizza il nodo radice utilizzato per calcolare lo SPT, come *successivo* si utilizza uno dei nodi figli a quello in cui si sta riempiendo la tabella e nel campo *predicato* si esegue l'operazione di *or logico* sui predicati dei nodi figli. Nell'esempio di Figura 24 abbiamo che il nodo numero 1 è utilizzato per calcolare l'albero a costo minimo, la tabella che si sta calcolando è quella del nodo 2 i valori successivi sono $\{4, 5, 8\}$ i predicati che si ottengono sono *Pf* per quanto riguarda 4, *Pg+Pd* per quanto riguarda il nodo 5. Ricapitolando definendo il grafo $G = (V, E)$ dove $v \in V$ e definendo i predicati di un nodo come $pred(u)$ allora possiamo definire una serie di passaggi per il calcolo del PSF:

1. Per ogni nodo v si calcola il corrispettivo albero di costo minimo.
2. Per ogni albero partendo dalla radice v si calcolano tutti i figli che chiamiamo u.
3. Per ogni u si compila una tabella con tre campi *sorgente*, *next-hop* e *predicato*.
4. In *sorgente* si mette il valore corrente di v
5. In *next-hop* si mette uno dei vicini di u
6. In *predicato* si effettua la somma logica dei predicati raggiungibili da *next-hop*

iPSF: improved PSF Il pre-source forwarding implica un grosso dispendio di risorse in quanto le tabelle nei nodi hanno dimensioni molto grandi. Per ottimizzare tale meccanismo si è sviluppato *iPSF* il quale, nel caso in cui due SPT calcolati su due radici diverse hanno un sotto-albero comune allora le tabelle dei nodi del sotto-albero comune avranno come valore del campo *sorgente*, come possiamo vedere in Figura 25 dove i nodi 1 e 3 hanno lo stesso sotto albero comune in formato dai nodi $\{2, 4, 5, 8\}$

PRF: Per-Reciver Forwarding In questo caso il calcolo di chi è interessato al messaggio viene fatto direttamente all'origine e nell'header del messaggio vengono inseriti i nodi interessati. Per ogni nodo esistono due tabelle una che contiene i forwarding con i rispettivi predicati ed un'altra che contiene gli instradamenti ai nodi successi come mostrato in Figura 26

Algoritmi di routing Per quanto riguarda gli algoritmi di routing possiamo averne di due tipi, il primo sono gli algoritmi di tipo *Distance Vector* nei quali i nodi hanno soltanto una visione parziale della rete tramite la richiesta ai nodi vicini del loro stato e la risposta da parte di quest'ultimi, si ha così un algoritmo distribuito nel quale solo col passare del tempo un nodo ha la piena visione della rete. Nel caso degli algoritmi *Link-State* tutti i nodi utilizzano un pacchetto denominato *link-state packet* che viene inoltrato ad un nodo centrale il quale, dopo aver calcolato la rete, restituisce a tutti i nodi le informazioni su tale rete.

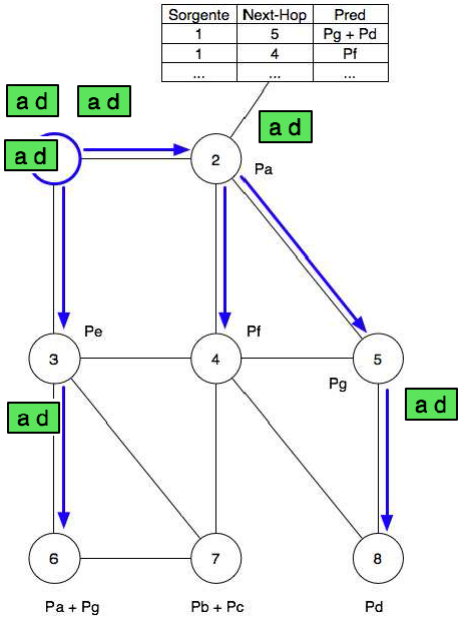


Figura 24: Esempio di PSF

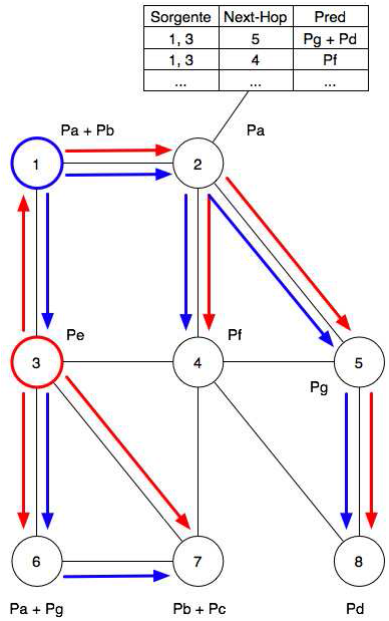


Figura 25: Esempio di iPSF

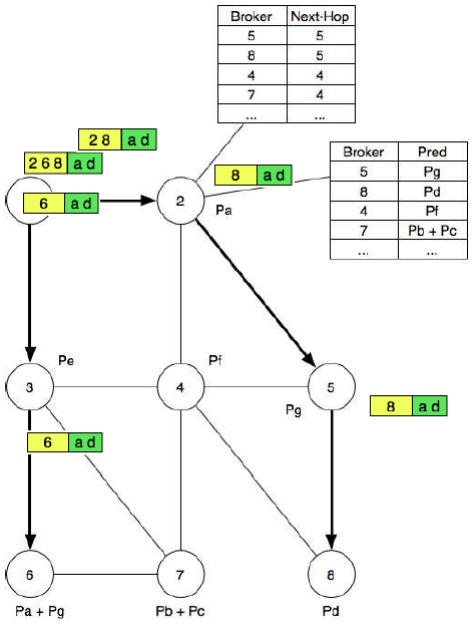


Figura 26: Esempio di PRF

5.5 Comunicazione orientata agli stream

Fino ad ora abbiamo visto dei meccanismi per lo scambio di informazioni nei quali non ha importanza l'istante in cui avviene tale scambio. Esistono però forme di comunicazione in cui il tempo ha un ruolo fondamentale, come nel caso degli *stream* audio nei quali per riprodurre un suono è necessario che i dati, nel caso stiamo parlando di un CD, siano riprodotti nell'ordine corretto e con una cadenza di 1/44100 sec. Eseguirli ad una velocità diversa riprodurrebbe un suono completamente diverso.

5.5.1 Supporto ai media continui

Il supporto allo scambio di informazioni dipendenti dal tempo spesso si traduce nel supporto ai media continui ovvero quei mezzi nei quali viene convogliata l'informazione come media media per la memorizzazione e la trasmissione o media per la presentazione. La caratteristica più importante dei media è come essi *rappresentano* l'informazione. Gli *stream* audio ad esempio possono essere codificati tramite campioni a 16 bit usando la PCM (*pulse code modulation*). Nei **media continui** le relazioni temporali tra i diversi dati sono fondamentali per interpretare correttamente l'informazione in essa contenuta. Ad esempio il movimento può essere rappresentato da una serie di immagini successive visualizzate ad un intervallo di tempo T regolare pari a 30-40 msec.

Stream di dati Per gestire lo scambio di informazioni continue i sistemi distribuiti forniscono supporto agli **stream di dati**. Uno stream non è altro che una sequenza continua di dati. Gli stream possono essere usati sia per i media continui che per quelli discreti, un esempio di stream di dati discreto sono le connessioni TCP/IP (stream di byte).
Il tempo è fondamentale negli stream di dati continui, nella modalità di trasmissione **asincrona** i dati di uno stream vengono trasmessi in sequenza uno dopo l'altro ma è l'unico vincolo imposto, questo è solo il caso degli stream di dati discreti.

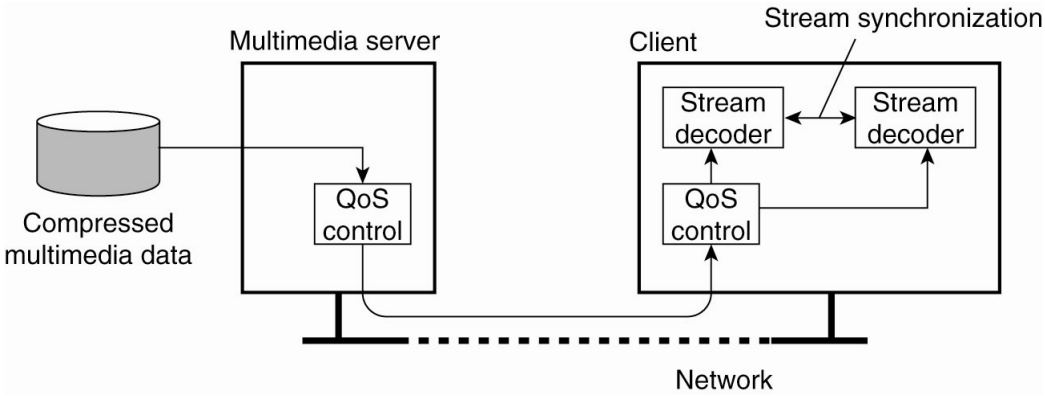


Figura 27: Esempio di sistema multimediale

Nella modalità di trasmissione **sincrona** viene definito un tempo massimo di trasmissione per ogni unità di dati in uno stream. Non è importante se un'unità è trasferita più velocemente del tempo massimo. Esiste infine una modalità di trasmissione **isocrona** nella quale è necessario che i dati siano trasferiti in un certo tempo. Ciò significa che i dati sono soggetti a vincoli di tempo sia minimi che massimi, questi limiti sono chiamati *bounded (delay) jitter*, questa modalità è abbastanza importante per i sistemi multimediali in quanto ha un ruolo importante nella rappresentazione audio e video.
Possiamo fare un'altra distinzione negli stream, essi possono essere **semplici** quando la sequenza di dati è unica, oppure **complessi** quando più stream semplici sono in relazione, tali stream sono chiamati **substream**. La relazione tra *substream* di uno stream complesso dipende spesso dal tempo. L'importante è che i *substream* siano continuamente sincronizzati. In altre parole le unità di dati provenienti da due substream devono essere trasmessi in coppia. Un possibile esempio di architettura di un sistema distribuito multimediale è quello mostrato in Figura 27.

5.5.2 Stream e qualità del servizio

I requisiti di tempo sono spesso espressi come requisiti della **qualità del servizio** (**QoS**,*quality of service*). Questi requisiti descrivono che cosa necessiti il sistema per assicurare che i vincoli siano rispettati. La QoS per gli stream di dati continui è essenzialmente relativa al tempo, al volume e all'affidabilità questo significa specificare:

- il *bit rate* a cui devono essere trasportati i dati;
- il tempo massimo di *set up* di una sessione (quando un applicazione può iniziare ad inviare dati);
- il tempo di trasporto massimo;
- la varianza massima del tempo di trasmissione o *jitter*;
- il tempo massimo di risposta.

Tuttavia la maggior parte dei sistemi distribuiti orientati agli stream è costruita sullo stack IP il quale è costituito da un servizio *datagram* di tipo *best-effort* il quale permette di cancellare pacchetti ogni qual volta la comunicazione diviene troppo pesante.

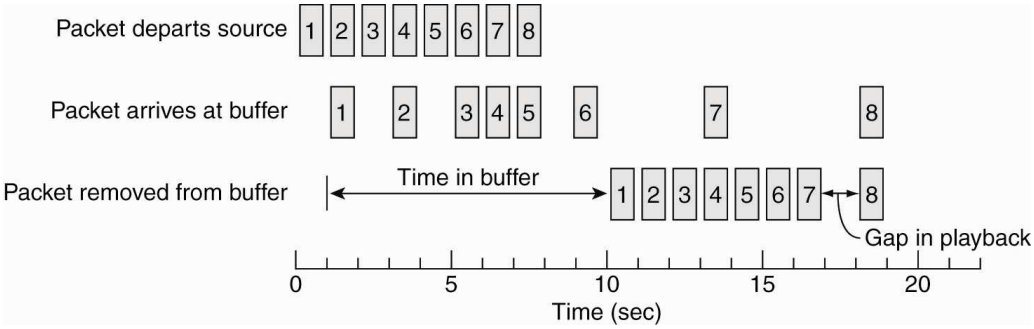


Figura 28: Esempio di utilizzo del buffer

Garantire la QoS Visto i problemi introdotti dall'utilizzo del protocollo IP il sistema distribuito può cercare di sopperire alla mancanza di qualità del servizio.
Prima di tutto l'IP permette di differenziare delle classi di dati per mezzo dei suoi **servizi differenziati**. Ad esempio un host può marcare i suoi pacchetti come appartenenti alla classe di **inoltro rapido** che indica che il pacchetto deve essere inoltrato dal router con priorità assoluta. Inoltre esiste una classe di **inoltro assicurato**.
Oltre a soluzioni a livello di rete, anche un sistema distribuito può usare alcuni accorgimenti per migliorare le prestazioni. Uno di questi è l'utilizzo di *buffer* per ridurre lo *jitter*; il principio è semplice ed è mostrato in Figura 28; il meccanismo prevede la memorizzazione dei pacchetti per un certo tempo massimo prima di passare i pacchetti all'applicazione in modo da avere sempre dei pacchetti a disposizione. Ovviamente può succedere che alcuni pacchetti arrivino con un ritardo maggiore del tempo di buffer questo causerà dei vuoti nello stream. Una soluzione è quella di aumentare il tempo di buffer ma questo causa un aumento del tempo di *set-up* dell'applicazione.
Un altro problema da affronta è la perdita di pacchetti, in quanto è inammissibile richiedere il rinvio dei pacchetti al mittente; è quindi necessario applicare delle tecniche di correzione degli errori. Una tecnica è quella di codificare i pacchetti in modo tale che qualsiasi k pacchetti persi tra gli n ricevuti siano sufficienti per ricostruire k pacchetti corretti. Un altro problema è quando un pacchetto contiene molti frame audio e video, in questo caso l'utente può percepire un notevole vuoto nella riproduzione. Tale vuoto può essere aggirato dall'uso di frame *interleaving* come mostrato nella Figura 29 in questo modo il vuoto viene distribuito su più frame e si nota meno; questo meccanismo tuttavia richiede un buffer più grande.

5.5.3 Sincronizzazione degli stream

Un elemento essenziale dei sistemi multimediali è che diversi stream eventualmente sotto forma di stream complesso, sono sincronizzati l'uno con l'altro. Sincronizzare più stream significa rispettare alcuni vincoli temporali tra gli stream.
Esistono diversi problemi di sincronizzazione come ad esempio la sincronizzazione di stream audio e video in un film il quale prende il nome di *lip synchronization*, un altro è il problema di sincronizzare i due substream che compongono uno stream per l'audio stereo.
La sincronizzazione avviene a livello delle unità di dati di cui è fatto uno stream. In altre parole possiamo sincronizzare due stream solo tra unità di dati. La scelta di che cosa sia un'unità di dati dipende molto dal livello di astrazione a cui è visto lo stream. Consideriamo uno stream audio di un CD; tale stream appare come una sequenza di campioni a 16bit con una frequenza di 44.1 kHz, la sincronizzazione con altri stream audio potrebbe avvenire ogni $23\mu sec$

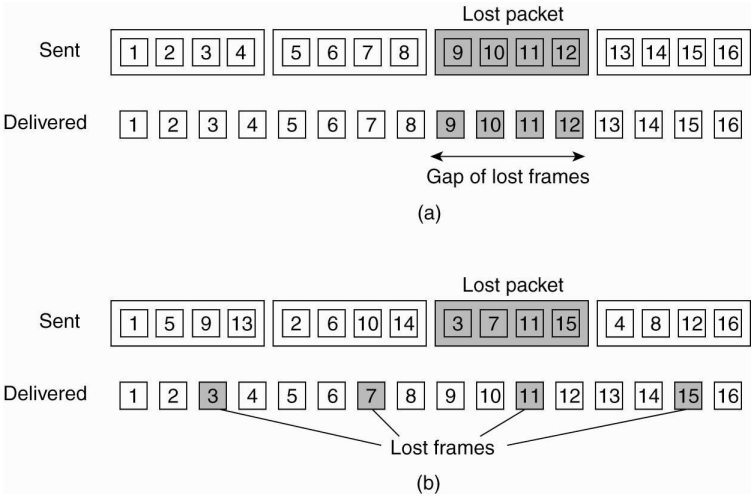


Figura 29: Applicazione del meccanismo di interleaving

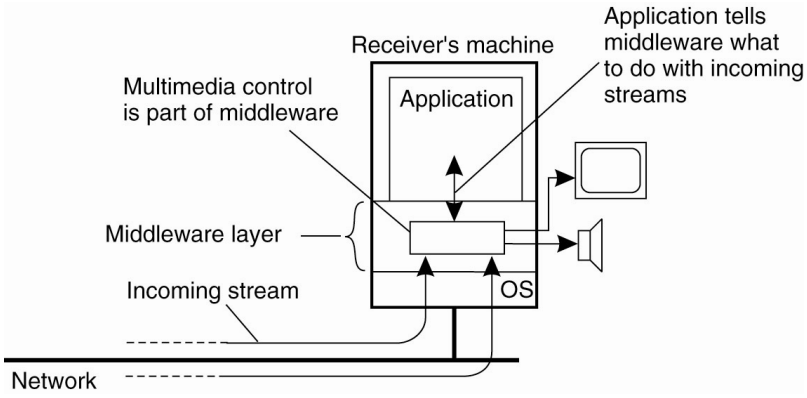


Figura 30: Sincronizzazione degli stream tramite interfacce

Meccanismi di sincronizzazione Prima di parlare di come sincronizzare due stream dobbiamo innanzitutto distinguere i meccanismi di base per la sincronizzazione di due stream dalla distribuzione di questi meccanismi in rete.
Come per la granularità sulle unità di dati anche i meccanismi di sincronizzazione possono lavora a diversi livelli di astrazione. A livello più basso la sincronizzazione viene fatta sulle unità di dati. In sostanza c'è un processo che esegue semplicemente delle operazioni di lettura e scrittura su molti stream assicurandosi che queste operazioni rispettino i determinati vincoli temporali. L'inconveniente di questo tipo di sincronizzazione è che l'applicazione è completamente responsabile dell'implementazione della sincronizzazione. Un approccio migliore è quello di fornire all'applicazione un'interfaccia che le consenta di controllare più facilmente lo stream come mostrato in Figura 30. Nei sistemi middleware per i media offrono una serie di interfacce per il controllo degli stream audio e video. Ogni dispositivo e ogni stream possiedono la loro interfaccia di alto livello inclusa quella per la notifica a un'applicazione di un cero evento.
Un altro aspetto da considerare è la distribuzione dei meccanismi di sincronizzazione. Nel caso sia il ricevente a dover sincronizzare uno stream complesso costituito da diversi substream esso deve sapere esattamente come procedere; in altre parole deve avere una *specifica della sincronizzazione* completa e disponibile localmente. In realtà le informazioni sulla sincronizzazione vengono fornite implicitamente inviando in *multiplexer* i diversi substream in un unico stream.

6 Naming

I nomi giocano un ruolo importante in tutti i sistemi di computer. Sono utilizzati per identificare, condividere e localizzare le diverse risorse. Un elemento importante del naming riguarda il fatto che un nome può essere risolto nell'entità a cui si riferisce, consentendo così ad un processo di accedere alla risorsa identificata da quel nome.
La differenza tra il *naming* nei sistemi distribuiti e nei sistemi non distribuiti è data dal modo in cui questi sistemi sono implementati. In un sistema distribuito molte volte anche il meccanismo di *naming* è distribuito per migliorarne efficienza e scalabilità.
Esistono diversi meccanismi di naming quelli che analizzeremo saranno quelli di tipo *human-friendly* come quelli dei file system o del World Wild Web. L'altro meccanismo di naming che analizzeremo è quello relativo al naming di dispositivi mobili dove i nomi human-friendly non sono adatti ma sono utilizzati meccanismi in cui i nomi sono identificatori indipendenti dalla posizione oppure quelli che utilizzano hash table distribuite. Infine analizzeremo quella tipologia di naming che esprimono le entità tramite varie caratteristiche attraverso l'utilizzo di attributi.

6.1 Nomi,identificatori ed indirizzi

Definiamo innanzi tutto che cos'è un nome. Un nome in un sistema distribuito è una stringa di bit o di caratteri utilizzata per riferirsi ad una entità. Un'entità può essere qualsiasi cosa, come host, stampanti, dischi o file; ma anche qualcosa che conosciamo meglio come processi, utenti, pagine web ecc.
Con tali entità noi possiamo interagire ma per fare ciò è necessario accedervi tramite un **punto d'accesso** (*access point*) che non è altro che un tipo particolare di entità il cui nome è anche chiamato **indirizzo.**
Un entità può avere più di un punto d'accesso e può cambiarlo nel corso del tempo. Un indirizzo perciò è un tipo speciale di nome che si riferisce al punto di accesso di un'entità. Visto che il punto di accesso è strettamente legato all'entità sarebbe opportuno utilizzare l'indirizzo come nome regolare per riferirsi all'entità. Questa tecnica non è però applicabile in quanto solitamente l'indirizzo non è di facile comprensione e nemmeno flessibile. Prendiamo ad esempio il caso in cui un sistema distribuito sia riorganizzato e che un servizio prima disponibile su di una macchina sia ora riassegnato ad un server differente supponiamo inoltre che sulla vecchia macchina venga messo in funzione un nuovo servizio; a questo punto se abbiamo utilizzato un indirizzo per riferirci all'entità nel momento in cui il punto di accesso cambia o viene riassegnato abbiamo un riferimento non valido.
Questo esempio illustra come per un entità un nome differente dal suo indirizzo sia più facile e più flessibile. Tale tipologia di nome è detta **indipendente dalla posizione**.
Oltre agli indirizzi ci sono altri tipi di nomi che meritano una piccola analisi e sono quelli utilizzati per identificare univocamente un'entità. Questi nomi sono detti **identificatori** e rispettano le seguenti proprietà:

- un identificatore si riferisce al massimo ad una entità.
- Ogni entità è referenziata da al massimo un identificatore.
- Un identificatore si riferisce sempre alla stessa entità (non è mai riusato)

Usando gli identificatori diventa più facile riferirsi a un'entità in maniera non ambigua.
Indirizzi ed identificatori sono due importanti categorie di nomi impiegati per due scopi molto diversi. In numerosi sistemi tali nomi sono sono rappresentati in una forma leggibile dalla

macchina ovvero sotto forma di bit. Al contrario i nomi ***human-friendly*** sono rappresentati sotto forma di stringhe di caratteri in modo da essere usate dalle persone.
Il problema principale del naming risulta a questo punto essere quello di risolvere nomi e identificatori in indirizzi. In linea di principio un sistema di *naming* mantiene un **collegamento nome-indirizzo** che nella sua forma più semplice è solo una tabella di coppie (*nomi, indirizzo*). Tuttavia in sistemi distribuiti su di un ampia area geografica o di grandi dimensioni una tabella centralizzata non può funzionare. Ciò che accade è che un nome viene scomposto in più parti e che la risoluzione del nome avviene in maniera ricorsiva

6.2 Flat naming

In precedenza abbiamo visto come gli identificatori sono adatti per rappresentare univocamente le entità tramite stringhe di bit molto semplici i quali per comodità vengono chiamati nomi non strutturati o semplici (*flat*). Tali nomi non contengono alcuna informazione su come localizzare il punto d'accesso all'entità.

6.2.1 Soluzioni semplici

Analizziamo come primo meccanismo due semplici approcci per localizzare le entità in una rete locale.

Broadcasting e multicasting Consideriamo una rete che offre funzionalità di *broadcasting* efficienti. Localizzare un'entità in tale ambiente è relativamente semplice, basta inviare un messaggio contenente l'identificatore dell'entità ad ogni macchina per verificare che tali macchine abbiano la risorsa. Solo le macchine che possono offrire un punto d'accesso per quella risorsa risponderanno al messaggio inviando l'indirizzo di quel punto d'accesso.
Questo principio è usato nel **protocollo di risoluzione degli indirizzi** (ARP *address resolution protocol*) per trovare l'indirizzo a livello del collegamento dati *(data-link)* di una macchina dato solo l'indirizzo IP.
In sostanza una macchina trasmette sulla rete locale un pacchetto richiedendo chi sia il proprietario di un determinato indirizzo IP. Quando un messaggio raggiunge una macchina controlla se ha l'indirizzo IP e in caso affermativo risponde al messaggio.
Tale meccanismo tuttavia diventa inefficiente man mano che la rete cresce, in quanto viene sprecata banda dalla grande quantità di messaggi inoltrati, inoltre molti host possono essere interrotti da richieste alle quali non possono dare una risposta. Una soluzione possibile è quella di passare al *multicasting* nel quale solo un ristretto numero di host riceve la richiesta. Il *multicasting* può essere usato per localizzare una risorsa nelle reti punto-a-punto, ad esempio, Internet supporta il *multicasting* a livello di rete consentendo a diversi host di unirsi ad uno specifico gruppo di *multicast*. Tali gruppi sono identificati da un **indirizzo multicast**. Quando un host invia un messaggio ad un indirizzo di multicast il livello di rete fornisce un meccanismo *best-effort* per consegnare questo messaggio a tutti i membri del gruppo.
Un modo per utilizzare un indirizzo multicast potrebbe essere il caso di un azienda nel quale un PC,che chiameremo *A*, può essere connesso alla rete. Quando A viene connesso gli viene assegnato un indirizzo IP ed entra a far parte di un gruppo multicast. Nel caso un altro PC volesse contattare A dovrebbe prima di tutto localizzarlo e per fare ciò potrebbe inviare un messaggio dov'è A? a tutto il gruppo multicast, se A è connesso risponde con il suo indirizzo attuale.
Un altro modo per utilizzare un indirizzo di multicast è quello di associarlo ad un'entità replicata ed utilizzare il multicast per localizzare la replica più vicina.

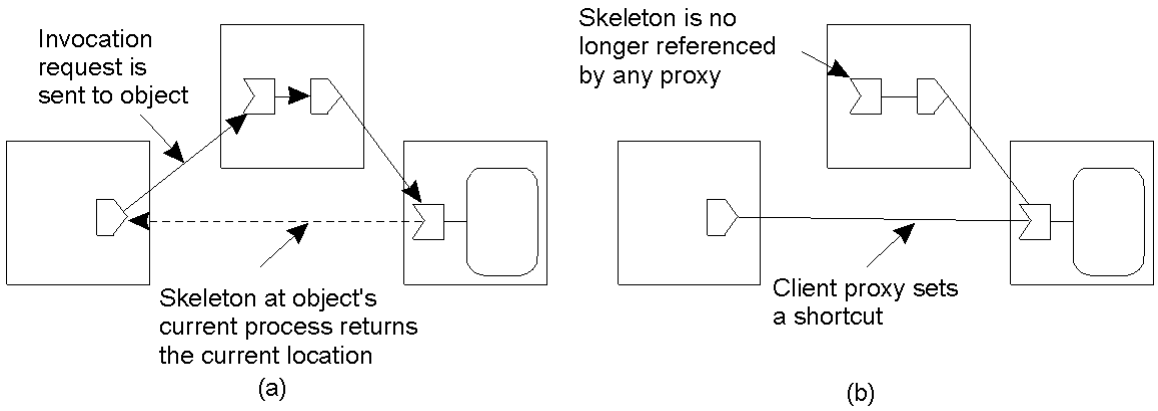


Figura 31: Esempio di utilizzo di puntatori forwarding

Puntatori forwarding Un altro approccio molto diffuso per localizzare un'entità mobile è quella di utilizzare dei puntatori *forwarding*. Il principio è abbastanza semplice, quando un'entità si sposta da A ad una nuova posizione B si lascia dietro un puntatore alla sua nuova posizione. Il vantaggio principale è la semplicità di realizzazione. Non appena viene trovata un'entità tramite *naming service* tradizionale, un client può cercare l'indirizzo attuale seguendo la catena dei puntatori.
Esistono però anche una serie di inconvenienti; primo fra tutti se non vengono prese le opportune contromisure la catena di puntatori per un'entità molto mobile può diventare estremamente lunga e la sua localizzazione diventare molto dispendiosa. Inoltre ogni nodo intermedio della catena deve mantenere la sua parte di puntatori finché è necessario. Infine, il sistema è molto vulnerabile alla perdita di comunicazione, se viene perso uno dei puntatori la risorsa non è più raggiungibile. È quindi fondamentale mantenere le catene di puntatori corte e i puntatori robusti.
Per capire il loro funzionamento prendiamo il caso dei puntatori *forwarding* applicati agli oggetti remoti ai quali si accede tramite chiamate a procedure remote. Seguendo l'approccio delle **catene SSP** ogni puntatore è implementato come coppia (*client stub, server stub*) come mostrato in Figura 31. Un server stub contiene un riferimento all'oggetto locale o un un riferimento ad un client stub remoto.

Quando un oggetto si sposta dallo spazzio degli indirizzi di A a quello di B lascia su A al suo posto un client stub che punta ad un server stub installato su B. Il punto focale è che la migrazione è completamente trasparente al client che contatta solo il client stub, gli è nascosto, invece, come e a quale posizione questo client inoltra le chiamate.
Per mantenere corta la catena una chiamata ad oggetto comporta l'identificazione del client che ha effettuato la chiamata tramite il suo indirizzo a livello di trasporto unito ad un numero generato localmente per identificare lo stub. Quando la chiamata arriva all'oggetto esso invia la risposta direttamente al client senza risalire la catena di puntatori, inoltre, insieme alla risposta viene inviata la posizione attuale dell'oggetto.
Si deve stabile un compromesso tra l'inviare una risposta direttamente al client stub oppure lungo la catena dei puntatori, nel primo caso la comunicazione è più veloce nel secondo invece è possibile aggiornare tutti i vari stub con la posizione aggiornata dell'oggetto.
Quando più nessun client fa riferimento ad un server stub, allora, quest'ultimo può essere rimosso. Tale operazione è strettamente collegata alla *garbage collection* distribuita.

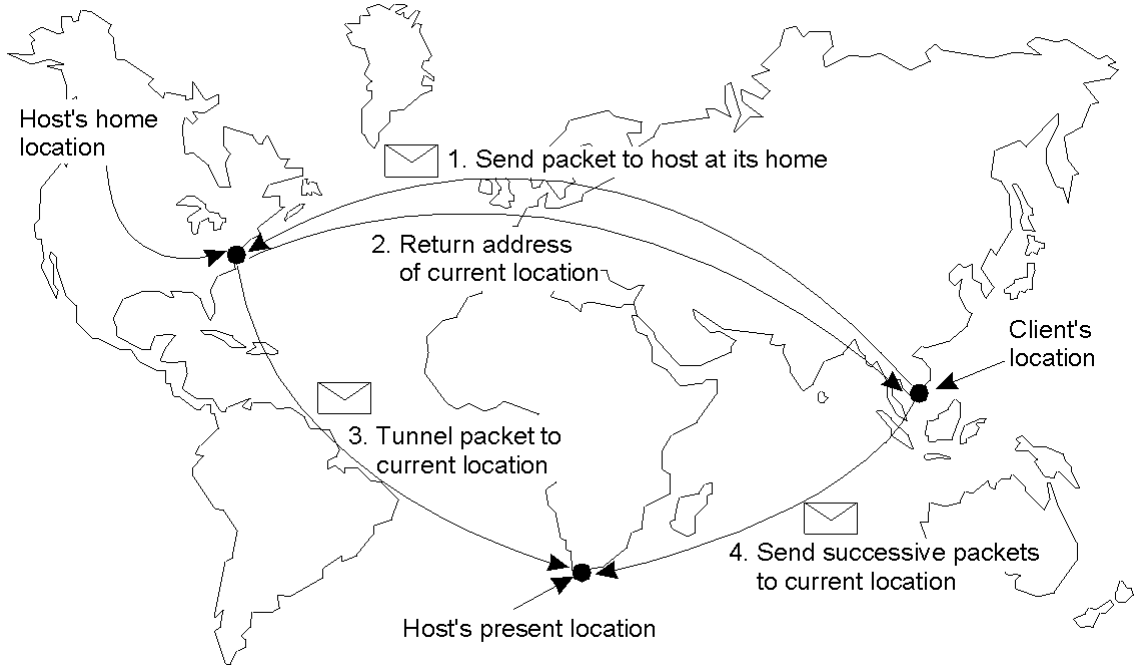


Figura 32: Il principio mobile IP

6.2.2 Approcci home-based

L'utilizzo del *broadcasting* o di puntatori *forwarding* comporta problemi di scalabilità. Un approccio diffuso per supportare entità mobili su reti di larga scala consiste nell'utilizzo della **home location** che tiene traccia della posizione attuale di un'entità. La home location di solito è il luogo dove è stata creata l'entità.
Il caso più comune nel quale si utilizzano gli approcci *home-based* è quello dei Mobile IP dove ogni host ha un indirizzo IP fisso. Tutte le comunicazioni verso questo indirizzo viene inizialmente diretta verso **home agent** dell'host mobile. L'home agent è posizionato nella rete locale corrispondente all'indirizzo IP dell'host. Quando l'host mobile si sposta in un altra rete richiede un indirizzo temporaneo; questo **care-of-address** viene registrato dall'*home agent*. Quando l'home agent riceve un pacchetto per l'host mobile cerca la posizione attuale dell'host, se l'host è nella rete locale allora il pacchetto semplicemente viene inoltrato, altrimenti, viene incanalato verso la posizione attuale dell'host e contemporaneamente il mittente del pacchetto viene informato sull'attuale posizione dell'host. Questo meccanismo è mostrato in Figura 32.
Un inconveniente di questo meccanismo è che un host può essere molto lontano dalla sua home, il risultato è un aumento dei tempi di latenza. Inoltre, bisogna assicurare che la home location esista sempre altrimenti risulta impossibile contattare la risorsa.

6.2.3 Hash table distribuite

I recenti sviluppi hanno portato a possibili risoluzioni di un identficatore nell'indirizzo di un elemento associato tramite l'utilizzo di *hash table* distribuite. Nella concezione di base i meccanismi basati su SHT non tengono conto della vicinanza della rete, questo può causare problemi di prestazioni.

Meccanismo generale Esistono diversi sistemi basati su DHT, il sistema *Chord* è rappresentativo di molti di loro anche se esistono importanti differenze riguardo la complessità di gestione e i protocolli di ricerca.
Come abbiamo visto nel Capitolo 2 Chord utilizza uno spazio degli identificatori a m bit per assegnare degli identificatori casuali ai nodi e alle chiavi di una specifica entità. Il numero m di bit è solitamente 128 o 160 a seconda della funzione di *hash* utilizzata. Un'entità con chiave k cade sotto la giurisdizione del nodo con il più piccolo identificatore $id \geq k$; questo nodo è chiamato il *successore* di k e indicato come $succ(k)$.
La questione principale in un sistema basato su DHT è quello di risolvere una chiave k nell'indirizzo di $succ(k)$. Un approccio semplice ma che purtroppo non scala bene è di lasciare che ogni nodo p tenga traccia del suo successore $succ(p+1)$ e del suo predecessore $pred(p)$. In questo caso quando un nodo p riceve una richiesta la inoltra semplicemente ad uno dei suoi vicini a meno che la chiave che sta cercando non rispetti $pred(p) < k \leq p$, in questo caso il nodo p deve restituire il suo indirizzo.
Diversamente da questo approccio lineare, ogni nodo Chord mantiene una **finger table** di al massimo m elementi. Se FT_p è la *finger table* del nodo p allora

$$FT_p[i] = succ(p + 2^{i-1})$$

Ovvero l'*i-esimo* elemento punta al primo nodo successivo a p di almeno 2^{i-1} posizioni. Questi riferimenti sono collegamenti a nodi realmente esistenti, dove la distanza del collegamento (*shortcutted distance*) dal nodo p cresce esponenzialmente man mano che aumenta l'indice della finger table.
Per individuare una chiave k il nodo p inoltra la richiesta al nodo q con indice j nella *finger table* di p dove

$$q = FT_p[j] \leq k < FT_p[j+1]$$

Per mostrare il funzionamento si rimanda alla Figura 33 dove si mostrano i passaggi per la ricerca di una chiave $k = 7$ inoltrata al nodo 1. Si può dimostrare che una ricerca richiede in genere $O(log(N))$ passi, dove N sono il numero di nodi del sistema.

In un sistema distribuito ci si può aspettare che l'insieme dei nodi cambi continuamente, non solo i nodi entrano ed escono volontariamente ma possono anche subire dei guasti per poi tornare a funzionare successivamente. Entrare in un sistema basato su DHT come *Chord* è relativamente semplice. Supponiamo che un nodo p voglia unirsi al sistema, esso contatta un nodo a caso del sistema e richiede la ricerca di $succ(p+1)$. Una volta identificato p può unirsi all'anello. La complessità sta nel mantenere le *finger table* aggiornate. La cosa più importante è che per ogni nodo q, $FT_q[1]$ (il successore) sia corretto. Per ottenere questo risultato ogni nodo q esegue periodicamente una semplice procedura che contatta il $succ(q+1)$ e gli richiede di restituire $pred(succ(q+1))$. Se $q = pred(succ(q+1))$ allora q è sicuro che le sue informazioni sono consistenti. Altrimenti nel sistema è entrato un nodo p con $q < p \leq succ(q+1)$ per cui q aggiornerà $FT_q[1]$ a p. A questo punto controllerà anche se p ha memorizzato q come suo predecessore. Se non fosse così allora è necessario un ulteriore aggiornamento di $FT_q[1]$.
Per aggiornare la finger table il nodo q non deve far altro che contattare il successore di $k = q + 2^{i-1}$ per ogni elemento i della sua *finger table*.

Sfruttare la vicinanza della rete Uno dei problemi principali del sistema Chord è che le richieste possono essere instradate in modo bizzarro, per minimizzare simili casi la progettazione di un sistema basato su DHT deve tener conto della rete sottostante.
Una prima soluzione potrebbe essere l'**assegnamento degli identificatori dei nodi basato**

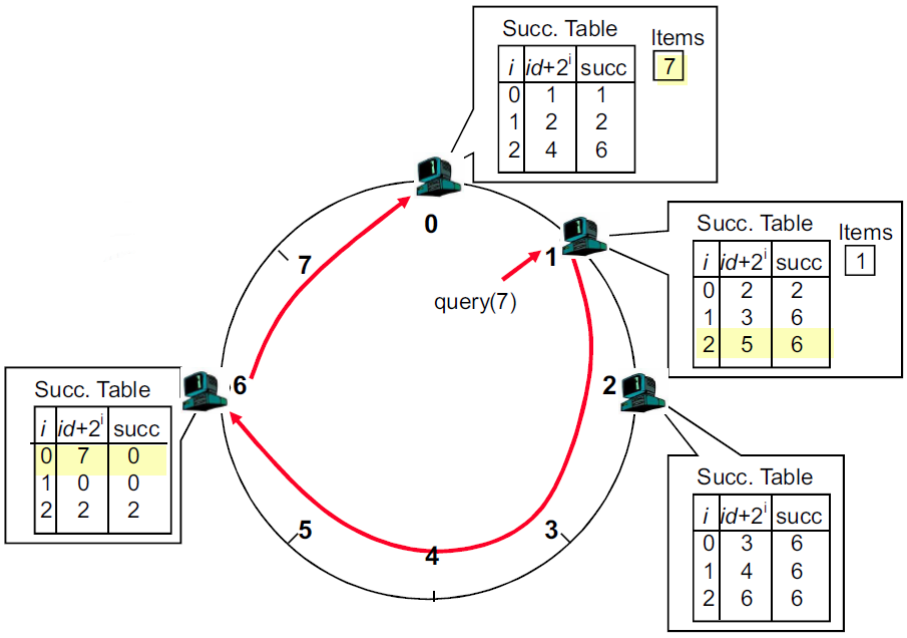


Figura 33: Esempio di esecuzione di due ricerche

sulla topologia, in altre parole l'idea è quella di assegnare gli identificatori in modo tale che nodi vicini abbiano anche identificatori vicini. Questo meccanismo porta con sè però molte problematiche tra cui il *mapping* di un anello logico su internet, che è un operazione non banale. Inoltre nel caso in cui una sottorete non diventi più raggiungibile si possono avere dei buchi consistenti negli identificatori che altrimenti avrebbero una distribuzione casuale.
Con l'**instradamento per vicinanza** (*proximity routing*) i nodi mantengono una liste di alternative a cui inoltrare una richiesta. Per esempio preso un indice della finger table non si tiene conto solo del successore ma anche di n successori nell'intervallo $[p + 2^{i-1}, p + 2^i - 1]$, ciò porta il sistema a poter scegliere a chi instradare una richiesta.
Infine nella **proximity neighbor selection** l'idea è quella di ottimizzare le tabelle di *routing* in modo tale che il nodo più vicino sia selezionato come *neighbor*. Non vi è molta differenza tra il *proximity routing* e il *proximity neighbor selection* in quanto nel proximity routing si scelgono r alternative mentre nel proximity neighbor si scelgono gli r vicini migliori.

6.2.4 Approcci gerarchici

L'approccio principale che presenteremo in questo paragrafo è basato sul servizio di localizzazione di *Globe*. Si tratta di un servizio di localizzazione generico rappresentativo di molti sistemi di localizzazione gerarchici.
In uno schema gerarchico una rete è suddivisa in un insieme di **domini**. Esiste un solo dominio *top-level* che comprende l'intera rete. Ogni dominio può essere suddiviso in molti sottodomini più piccoli, il dominio di livello più basso (*lowest-level*) viene chiamato **dominio foglia** e corrisponde di solito ad una rete locale (LAN).
Ogni dominio D ha un nodo directory associato $dir(D)$ che tiene traccia delle entità nel dominio. Questo meccanismo porta ad un albero di nodi directory con il nodo directory del dominio *top-level* chiamato anche **nodo radice** il quale conosce tutte le entità. Tale organizzazione è mostrata in Figura 34

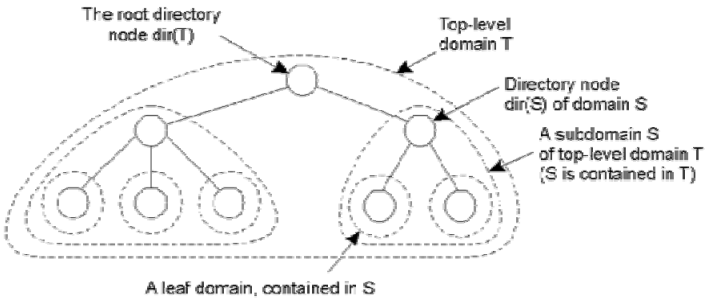


Figura 34: Esempio di organizzazione gerarchica

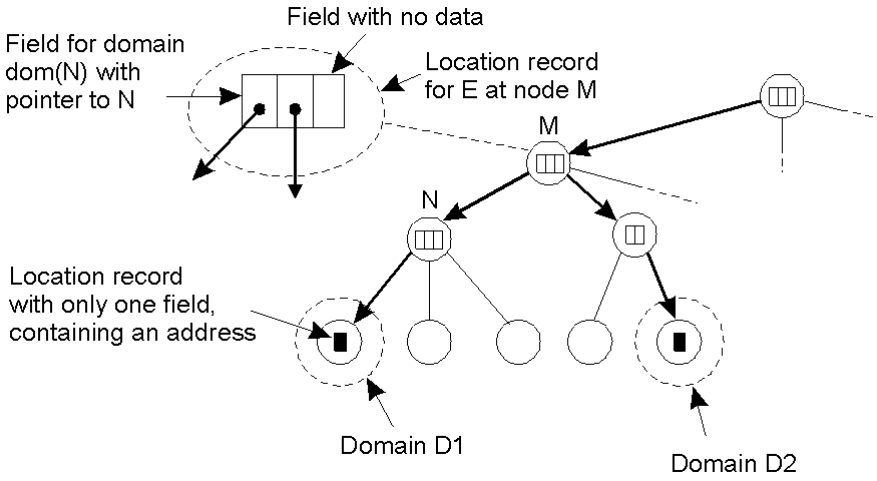


Figura 35: Esempio di memorizzazione delle informazioni nel caso di entità replicata

Ogni entità attualmente posizionata in un dominio D è rappresentata da un **location record** nel nodo directory $dir(D)$. Un *location record* dell'entità E nel nodo directory N del dominio foglia D contiene l'indirizzo attuale dell'entità in quel dominio. Mentre a livello superiore il *location record* di E conterrà un puntatore ad N.
Un'entità può avere molteplici indirizzi, ad esempio nel caso sia replicata. In tal caso il nodo directory del dominio più piccolo che contiene le due entità avrà due puntatori ai due sotto-domini per quell'entità coem mostrato in Figura 35 Per quanto riguarda la ricerca in un sistema gerarchico avviene anch'essa in modo gerarchico ma con un approccio *bottom-up*; se nel nodo directory non è presente un *location record* per quell'entità allora il nodo inoltra la richiesta al padre, se anche il padre non ha un *location record* per l'entità la richiesta risale l'albero fino a quando non trova un location record valido oppure non raggiunge la radice. Non appena una richiesta raggiunge un nodo directory M contenente un *location record* per l'entità E sappiamo che E si trova da qualche parte nel dominio $dom(M)$. A questo punto la richiesta viene inoltrata al nodo directory del sottodominio contente E e così via procedendo fino a raggiungere una foglia, questo indirizzo può finalmente essere restituito al client che ha inoltrato la richiesta. Questo procedimento è illustrato in Figura 36 Le operazioni di aggiornamento sfruttano lo stesso principio, l'inserimento avviene in un nodo foglia e il riferimento ad E è propagato verso l'alto fino a quando in un nodo directory non si incontra già un riferimento ad E, a questo punto il nodo che ha già il *location record* per e memorizza il nodo figlio dal quale è giunta la richiesta di inserimento, per installare la catena dei puntatori si possono applicare sia meccanismi *top-down*

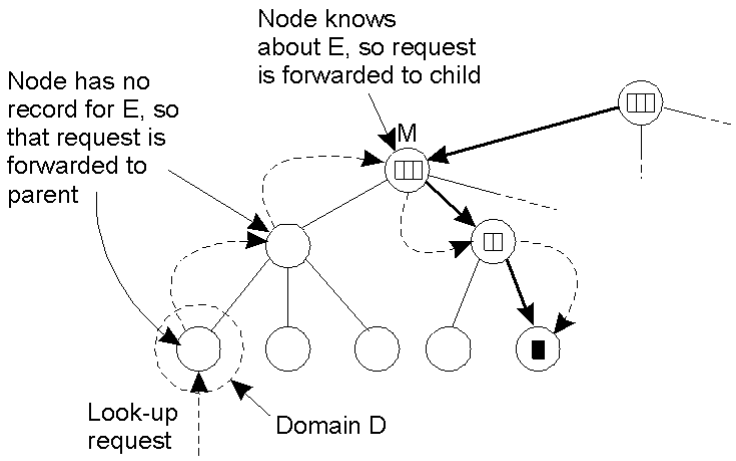


Figura 36: Esempio di ricerca in un sistema gerarchico

(una volta raggiunto un nodo con un location record per E si ridiscende l'albero impostando i diversi puntatori) oppure, un approccio *bottom-up* (a mano a mano che la ricerca del location record di E sale si instaurano i puntatori ad E). Nel secondo caso una risorsa diventa disponibile per un determinato dominio non appena la richiesta risale per il nodo directory di quel dominio.

6.3 Naming strutturato

I nomi semplici sono adatti per le macchine ma in generale non sono opportuni per gli uomini. Come alternativa i sistemi di *naming* supportano i nomi strutturati composti da diversi nomi semplici e leggibili dall'uomo. Un esempio sono i nomi dei file, i nomi degli host di Internet e altri ancora.

6.3.1 Name space

I nomi sono solitamente organizzati nel cosiddetti **spazi dei nomi** o *name space*. Uno spazio dei nomi può essere rappresentato come un grafo orientato nel quale esistono due tipi di nodi, i **nodi foglia** che rappresentano le diverse entità e può contenere soltanto informazioni come indirizzi, come nel caso di host oppure contenere direttamente tutto lo stato dell'entità come nel caso dei file system.
Esistono poi i **nodi directory** i quali hanno diversi archi in uscita ciascuno dei quali etichettato con un nome. Tuttavia questa è l'unica differenza in quanto in un grafo dei nomi ogni nodo è considerato come un'entità. Un nodo directory contiene una tabella in cui un arco in uscita è rappresentato da una coppia (*etichetta dell'arco, identificatore del nodo*), tale tabella è chiamata **directory table**.
Un nodo che ha solo archi in uscita e nessun arco in entrata viene detto **nodo radice**. Ogni percorso in un grafo dei nomi può essere chiamato tramite una sequenza di etichette corrispondenti agli archi del percorso ad esempio

$$N :< label_1, label_2, \ldots, label_n >$$

Tale percorso è detto **path name**. Se il primo nodo in un *path name* è la radice del grafo è detto **path name assoluto** altrimenti è detto **path name relativo**.
Un **nome globale** è un nome che denota la stessa entità a prescindere da dove sia usato il nome

all'interno del sistema. Diversamente un **nome locale** è un nome la cui interpretazione dipende da dove il nome viene usato.

6.3.2 Risoluzioni dei nomi

Gli spazi dei nomi sono un meccanisomo adatto a memorizzare e recuperare informazioni sulle entità per mezzo dei nomi. Dato un path name dovrebbe essere possibile ricercare qualunque informazione memorizzata nel nodo cui il path name si riferisce. Il processo di ricerca in base a un nome è chiamato **risoluzione del nome** o *name resolution*.
Per spiegare come funziona consideriamo il *path name* precedente

$$N :< label_1, label_2, \ldots, label_n >$$

La risoluzione di questo nodo comincia dal nodo N del grafo dei nomi, viene poi ricercato il nodo con nome $label_1$ nella *directory table*; ci si sposta poi in tale nodo e nella sua *directory table* si ricerca il riferimento al nodo con nome $label_2$ e così via. Supponendo che tale path sia un percorso valido allora la risoluzione terminerà nell'ultimo nodo indicato da $label_n$ con la resttuzione del contenuto di quel nodo.

Meccanismo di chiusura La risoluzione di un nome può avvenire soltanto se sappiamo come e da dove iniziare, nel nostro esempio il nodo iniziale era specificato e sapevamo di aver accesso alla sua *directory table*. Sapere come e da dove iniziare viene chiamato **meccanismo di chiusura**.
Un meccanismo di chiusura ha a che fare con la selezione del nodo iniziale in uno spazio dei nomi da cui deve iniziare la risoluzione. A volte però tali meccanismi sono di difficile comprensione in quanto molte volte impliciti e diversi gli uni dagli altri.
Per esempio nel file system di UNIX la risoluzione punta sul fatto che l'*inode* della *root directory* è il primo *inode* nel disco logico.

Linking e mounting Strettamente correlato alla risoluzione dei nomi è l'uso di **alias**. Un alias è un altro nome per la stessa entità. Ad esempio una variabile d'ambiente come `$HOME` è un esempio di alias. Ci sono fondamentalmente due modi per implementare un alias, il primo metodo è di consentire che molteplici path name assoluti si riferiscano allo stesso nodo come mostrato in Figura 37. Questo approccio nei file system UNIX è chiamato **hard link**. Il secondo

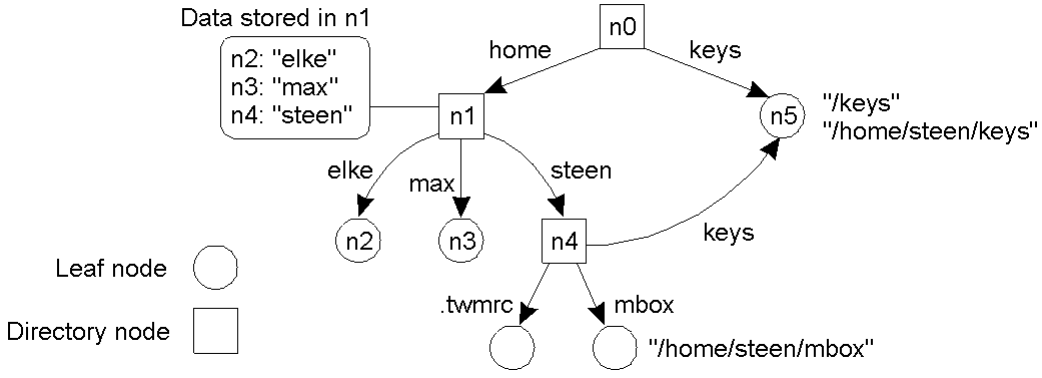


Figura 37: Esempio di hard link

metodo è quello di rappresentare un'entità tramite un nodo foglia il quale al posto di memorizzare

l'indirizzo o lo stato dell'entità memorizza un path name assoluto come mostrato in Figura 38. Questa volta, sempre riferendoci ai file system UNIX parliamo di **link simbolici**.

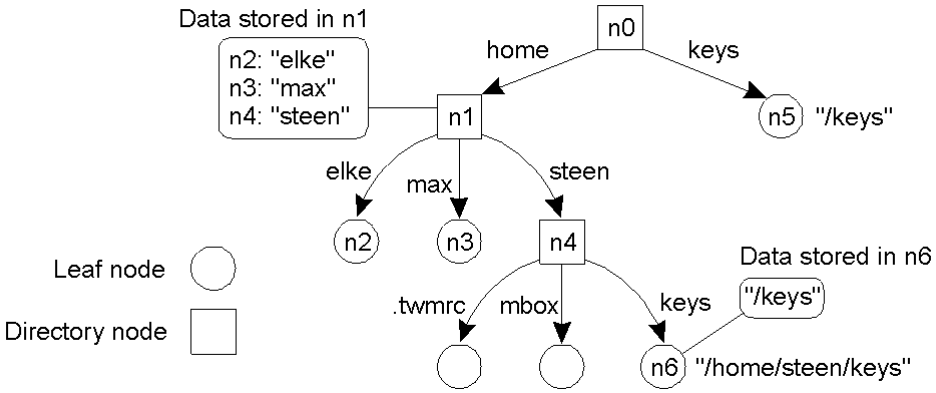


Figura 38: Esempio di hard link

6.3.3 Implementazione di uno spazio dei nomi

Lo spazio dei nomi come abbiamo visto è il cuore di un *naming service*. Un naming service è implementato da un name server. Se un sistema distribuito è limitato ad una rete locale è possibile implementare un name service mediante un unico name server centralizzato. Se invece il sistema ha molte entità ed è distribuito su larga scala allora è necessario distribuire il name server su più macchine.

Distribuzione dello spazio dei nomi Gli spazi dei nomi per sistemi distribuiti su larga scala sono solitamente organizzati gerarchicamente. Il **livello globale** è costituito dai nomi di più alto livello cioè il nodo radice ed i nodi directory logicamente più vicini ad essa. i nodi del livello globale sono spesso caratterizzati per la loro stabilità nel senso che le directory table cambiano raramente. Tali nodi possono rappresentare le aziende o i gruppi di aziende i cui nomi sono memorizzati nello spazio dei nomi.
Il **livello amministrativo** è costituito dai nodi directory gestiti da una sola azienda, questi nodi rappresentano gruppi di entità che appartengono alla stessa azienda o unità amministrativa, come un nodo per ogni per ogni dipartimento dell'azienda oppure uno utilizzato solo per gli utenti.
Il **livello gestionale** è il livello più basso ed è costituito da nodi che cambiano regolarmente. A questo gruppo appartengono gli host di una rete interna. Diversamente dagli altri livelli questi host sono gestiti non solo dagli amministratori ma anche dagli utenti finali.
La Figura 39 mostra una possibile suddivisione dello spazio dei nomi del DNS. Lo spazio dei nomi è diviso in parti non sovrapponibili chiamate **zone**. Una zona è uno spazio dei nomi che è implementato da un nameserver separato. Relativamente a disponibilità e prestazioni i name server dei diversi livelli devono rispettare diversi requisiti. I name server del livello globale devono essere altamente disponibili, in quanto se un name server si guasta un'ampia parte dello spazio dei nomi diventa irraggiungibile. Per quanto riguarda le prestazioni invece sono un po meno critiche in quanto cambiando raramente i risultati di ricerca possono essere memorizzati localmente dai client tramite meccanismi di cache. Il *throughtput* però deve essere elevato in quanto le richieste provengono da una grande quantità di utenti. I requisiti di disponibilità e

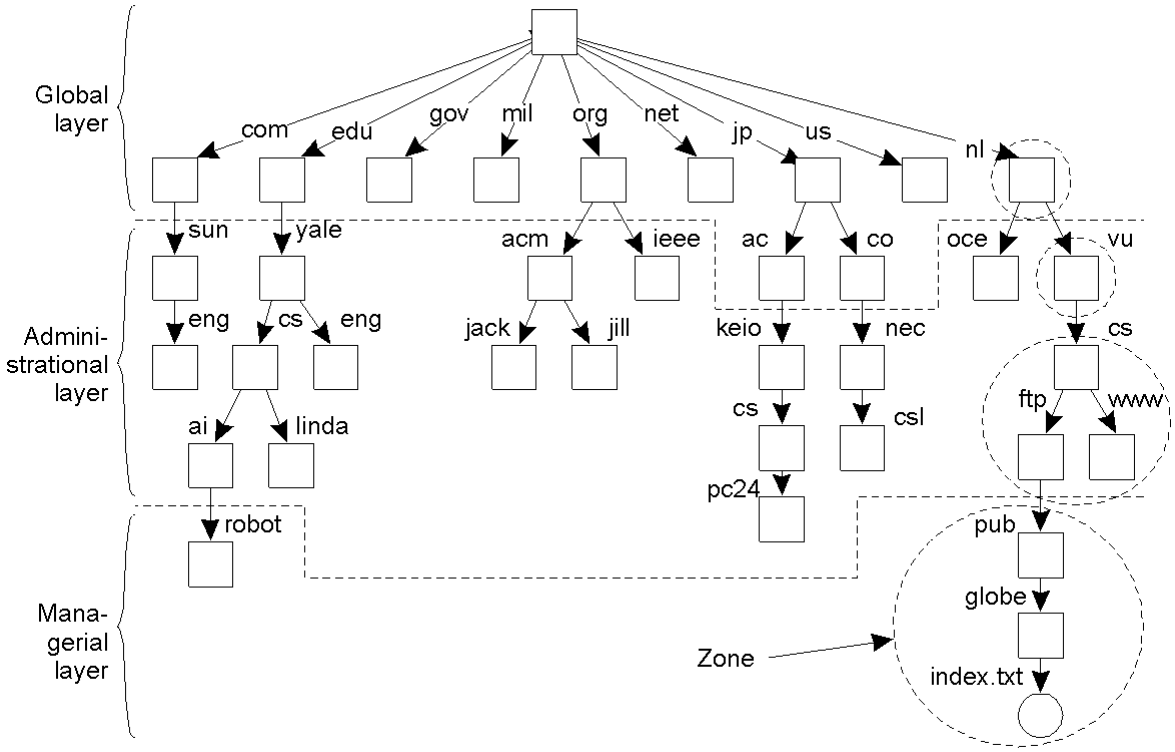


Figura 39: Esempio di suddivisione dei tre livelli nello spazio dei nomi del DNS

di prestazzioni per i server di livello globale possono essere soddisfatti replicando i server in combinazione con i meccanismi di cache.
La disponibilità dei name server a livello amministrativo è importante per i client all'interno della stessa azienda, infatti se si guasta le risorse all'interno dell'azienda diventano irraggiungibili. In merito alle prestazioni sono molto simili a quelle del livello globale in quanto anche qui i nodi non cambiano spesso, tuttavia, le tempistiche per i risultati delle ricerche devono essere nell'ordine di qualche millisecondo, anche gli aggiornamenti devono essere tempestivi, infatti, è inaccettabile che per l'attivazione di un account si debba aspettare qualche ora.
I requisiti prestazionali dei name server a livello gestionale sono molto più stringenti, la disponibilità non è importante ma le prestazioni sono una caratteristica molto importante in quanto gli utenti si aspettano che le operazioni avvengano immediatamente. Inoltre a causa dei continui aggiornamenti i meccanismi di cache lato client non sono efficaci e bisogna perciò interrogare sempre il nameserver.

Implementazione dello spazio dei nomi La distribuzione di uno spazio dei nomi su più *name server* influenza l'implementazione della risoluzione dei nomi. Per spiegare la loro implemnetazione partiamo perciò dal caso più semplice in cui i name server non siano replicati ne che abbiano meccanismi di cache.
Ogni client ha accesso ad un **name resolver** locale responsabile di portare avanti il processo di risoluzione dei nomi. Partiamo dall'esempio mostrato anche in Figura 39 e focalizziamoci sulla risoluzione del *path name* assoluto:

$$root :< nl, vu, cs, ftp, pub, globe, index.html >$$

che usando una notazione URL è possibile tradurre in *ftp://ftp.cs.vu.nl/pub/globe/index.html*.
Esisstono due modi per implementare la risoluzione dei nomi. Il primo metodo è quello della **risoluzione dei nomi iterativa**, un *name resolver* passa ak *root name server* il nome completo . Il *root server* risolverà il nome appena possibile e lo restituirà al client, nel nostro esempio il *root server* può risolvere solo l'etichetta *nl* per la quale restituirà l'indirizzo del name server associato. A questo punto il client passa il resto del path name (cioè $nl :< vu, cs, ftp, pub, globe, index.html >$) a quel name server. Questo server può risolvere solo l'etichetta *vu* e restituisce l'indirizzo del name server associato. Il *name resolver* continuerà così fino alla completa risoluzione del nome. Questo processo è mostrato in Figura 40 dove la notazione $\# < cs >$ indica l'indirizzo del server che si occupa della parte *cs* del nome.
In alternativa al meccanismo iterativo è possibile usare la ricorsione, invece di restituire ogni risultato intermedio al client un *name server* passa il risultato al name server successivo. Questo meccanismo mostrato in Figura 41 è chiamato **risoluzione dei nomi ricorsiva.** In questo caso quando il root name server trova l'indirizzo del server che implementa il nodo chiamato *nl* gli chiede di risolvere il path $nl :< vu, cs, ftp, pub, globe, index.html >$, il quale a sua volta individuerà il server che implementa *vu* e gli chiederà di risolvere il rimanente path.
L'inconveniente di questo tipo di risoluzione è che richiede ai name server livelli prestazionali maggiori rispetto alla versione iterativa. Tuttavia l'approccio ricorsivo presenta anche alcuni vantaggi, ad esempio, il cacheing è molto più efficace rispetto al caso iterativo inoltre, i costi di comunicazione sono ridotti al minimo.

6.3.4 DNS

Uno dei servizi di *naming* più diffusi oggi al mondo è il *domain name system* (DNS) di Internet. Il DNS è principalmente usato per ricercare gli indirizzi IP degli host e dei mail server.

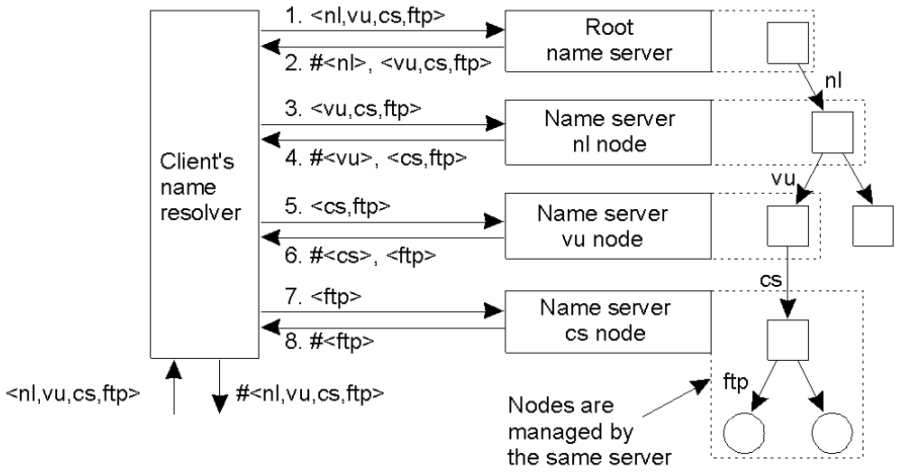


Figura 40: Risoluzione dei nomi iterativa

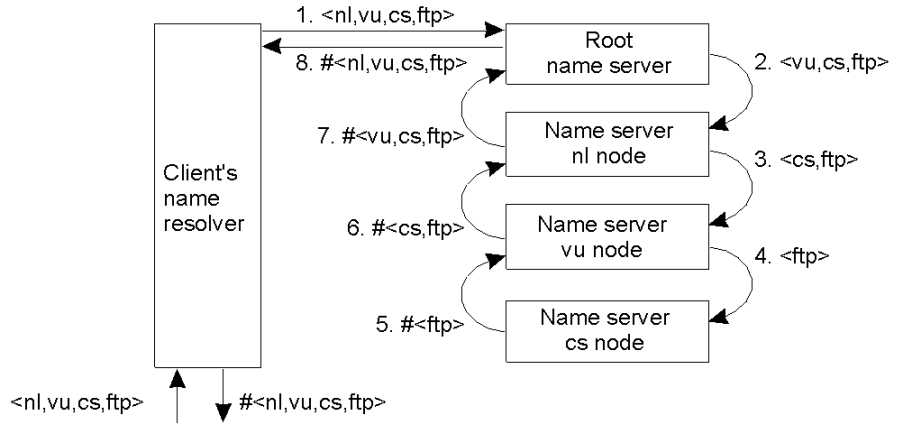


Figura 41: Risoluzione dei nomi ricorsiva

Spazio dei nomi del DNS Lo spazio dei nomi del DNS è organizzato gerarchicamente come un albero con radice. Un'etichetta è una stringa *case insensitive* costituita da caratteri alfanumerici con una lunghezza massima di 63 caratteri; la lunghezza di un path è limitata a 255 caratteri. La rappresentazione di un path name in forma di stringa parte dall'etichetta più a destra e separata da un punto (.). Anche la radice è rappresentata da un punto ma questo di solito viene omesso. Un esempio è il path name $root :< nl, vu, cs, flits >$ che rappresentato sotto forma di stringa diventa *flits.cs.vu.nl*.
Dato che un nodo nello spazio dei nodi ha esattamente un arco in entrata la sua etichetta viene usata come nome del nodo. Un sottoalbero è chiamato **dominio** ed il path name verso il suo nodo radice è chiamato **nome del dominio.** Il contenuto di un nodo è costituito da un insieme di **resource record** i quali possono essere di diverso tipo e sono illustrati in Tabella 4. Analiziamo ora alcuni campi dei resource record; un campo SOA contiene informazioni quali l'indirizzo mail dell'amministratore di sistema, il nome dell'host dal quale prelevare informazioni sulla zona ecc. Il resource A(*address*) rappresenta un particolare host in internet, il campo A contiene il suo indirizzo IP, in caso di host *multihomed* un nodo conterrà più campi A. I record

Tipo di record	Entità associata	Descrizone
SOA	Zona	Informazioni sulla zona rappresentata
A	Host	Contiene un indirizzo IP dell'host che questo nodo rappresenta
MX	Dominio	Si riferisce ad un mail server che gestisce le mail indirizzate a questo nodo
SRV	Dominio	Si riferisce ad un server che gestisce un particolare servizio
NS	Zona	Indica il name server che implementa la zona rappresentata
CNAME	Nodo	Link simbolico con il nome primario del nodo rappresentato
PTR	Host	Contiene il nome canonico dell'host
HINFO	Host	Mantiene informazioni sull'host che questo nodo rappresenta
TXT	Qualunque tipo	Contiene qualunque tipo di informazione

Tabella 4: Tipi importanti di resource record

MX (*mail exchange*) sono contengono un link simbolico a un nodo che rappresenta il mail server per quella zona.
Un esempio di tali record sono rappresentati in Figura 42

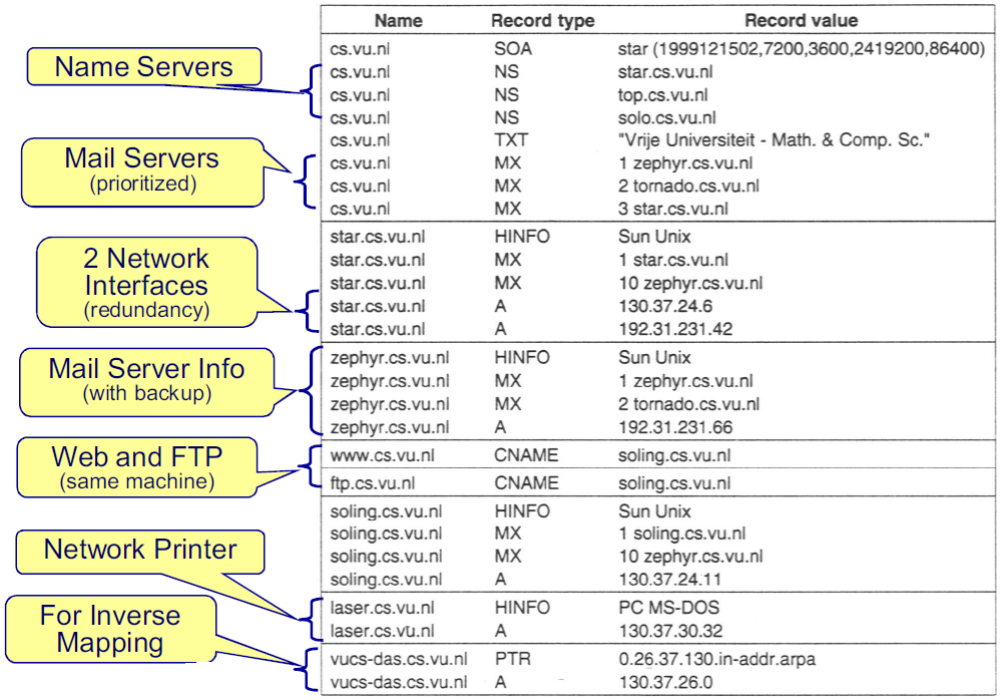
Implementazione del DNS In sostanza , lo spazio dei nomi del DNS è suddivisibile in un livello globale ed uno amministrativo come mostrato in Figura 39. Il livello gestionale solitamente è gestito a livello locale e quindi non fa parte del DNS. Ogni zona è implementata da un name server replicato per garantirne la disponibilità. Gli aggiornamenti della zona sono solitamente gestiti dal server primario. Gli aggiornamenti vengono propagati ai server secondari solo tramite richiesta di quest'ultimi, questo procedimento è chiamato **trasferimento di zona**.
Una base di dati del DNS è un insieme di file contenete diversi *resource record* tra questi file uno in particolare contiene gli identificatori di tutti i nodi di una particolare zona in modo da consentire l'identificazione di tutti i nodi semplicemente mediante il nome del loro dominio.

6.4 Naming basato sugli attributi

I nomi semplici e quelli strutturati offrono la possibilità di far riferimento ad un'entità in modo indipendente dalla sua posizione. Inoltre i nomi strutturati sono progettati per essere relativamente *human-friendly*. Ma a volte non interessa il nome dell'entità ma un utente vorrebbe ricercare una risorsa in base ad una serie di attributi specificati.
Un modo assai diffuso per effettuare questa ricerca è usare un sistema di **naming basato sugli attributi**, il quale consiste nel descrivere un'entità in termini di coppie (*attributo, valore*) e ad ogni entità possono essere associati più attributi diversi.

6.4.1 Directory service

I sistemi di naming basati sugli attributi sono anche conosciuti come **directory service**. Con questo meccanismo le entità hanno associato un insieme di attributi utilizzabili per le ricerche. Ad esempio in un sistema di mail i messaggi possono essere etichettati tramite attributi relativi al mittente al destinatario all'oggetto e così via; quando però si vuole ampliare il meccanismo



Name	Record type	Record value
cs.vu.nl	SOA	star (1999121502,7200,3600,2419200,86400)
cs.vu.nl	NS	star.cs.vu.nl
cs.vu.nl	NS	top.cs.vu.nl
cs.vu.nl	NS	solo.cs.vu.nl
cs.vu.nl	TXT	"Vrije Universiteit - Math. & Comp. Sc."
cs.vu.nl	MX	1 zephyr.cs.vu.nl
cs.vu.nl	MX	2 tornado.cs.vu.nl
cs.vu.nl	MX	3 star.cs.vu.nl
star.cs.vu.nl	HINFO	Sun Unix
star.cs.vu.nl	MX	1 star.cs.vu.nl
star.cs.vu.nl	MX	10 zephyr.cs.vu.nl
star.cs.vu.nl	A	130.37.24.6
star.cs.vu.nl	A	192.31.231.42
zephyr.cs.vu.nl	HINFO	Sun Unix
zephyr.cs.vu.nl	MX	1 zephyr.cs.vu.nl
zephyr.cs.vu.nl	MX	2 tornado.cs.vu.nl
zephyr.cs.vu.nl	A	192.31.231.66
www.cs.vu.nl	CNAME	soling.cs.vu.nl
ftp.cs.vu.nl	CNAME	soling.cs.vu.nl
soling.cs.vu.nl	HINFO	Sun Unix
soling.cs.vu.nl	MX	1 soling.cs.vu.nl
soling.cs.vu.nl	MX	10 zephyr.cs.vu.nl
soling.cs.vu.nl	A	130.37.24.11
laser.cs.vu.nl	HINFO	PC MS-DOS
laser.cs.vu.nl	A	130.37.30.32
vucs-das.cs.vu.nl	PTR	0.26.37.130.in-addr.arpa
vucs-das.cs.vu.nl	A	130.37.26.0

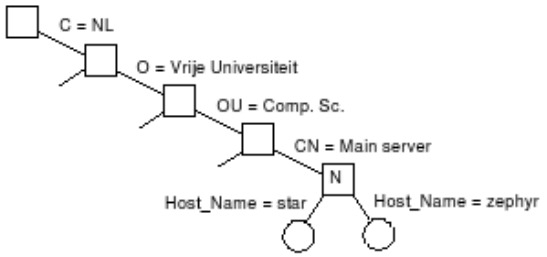
Figura 42: Esempio di resource record estratto dalla base di dati per la zona *cs.vu.nl*

dei descrittori risulta un po più difficile, in quanto la maggior parte delle volte tale meccanismo viene impostato manualmente.
Per attenuare alcune problematiche sono state introdotti diversi framework tra cui il **resource description framework** (**RDF**) il quale basa la sua descrizione su una tripletta formata da soggetto, predicato e oggetto i quali possono essere delle risorse oltre che degli attributi come nel caso della tripletta (*Persona, nome, Alice*) dove si fa riferimento ad una risorsa di tipo *Persona* il cui *nome* è *Alice*.
A differenza dei sistemi di naming strutturati la ricerca dei valori in un sistema di naming basato sugli attributi richiede di effettuare una ricerca esaustiva su tutti i descrittori.

6.4.2 LDAP

Un approccio diffuso di affrontare il problema dei directory service è quello di combinare il naming strutturato con quello basato sugli attributi, questo approccio è stato ampiamente usato nel servizio *Active Directory* di Microsoft ed in altri sistemi. Molti di questi sistemi utilizzano o si basano sul **light directory access protocol** comunemente chiamato anche **LDAP**.
Un directory service LDAP consiste in un certo numero di record di solito chiamati elementi della directory. Un elemento della directory è paragonabile ad un resource record del DNS. Ogni record è composto da una coppia (*attributo, valore*) in cui ogni attributo ha un tipo associato. L'insieme di tutti gli elementi di un directory service LDAP è chamato **directory information base** (**DIB**). Un aspetto importante di un DIB è che ogni record ha un nome univoco in modo da renderlo identificabile, tale nome appare in ogni record come la sequenza di attributi di naming. Ogni attributo di naming è chiamato **relative distinguished name** o in breve **RDN**. Come nel caso dei nomi univoci strutturati l'uso dei nomi globali ottenuti tramite l'elenco degli RDN in sequenza genera una gerarchia degli elemnti della directory chiamata anche **directory information tree** (**DIT**). Un DIT non è altro che un grafo dei nomi in un sistema basato su

LDAP in cui ogni nodo è un elemento della directory. In tal senso alcuni nodi possono agire sia da elemento sia da directory nel senso tradizionale del termine. Per accedere a tali elementi possiamo utilizzare due distinte funzioni di interrogazione, la `read` che è utilizzata per leggere il contenuto di un elemento, e la `list` utilizzata per elencare i diversi archi in uscita come mostrato in Figura 43(b). Quando si ha progetta un sistema basato su LDAP di larga scala il DIT viene



(a)

Attribute	Value
Country	NL
Locality	Amsterdam
Organization	Vrije Universiteit
OrganizationalUnit	Comp. Sc.
CommonName	Main server
Host_Name	star
Host_Address	192.31.231.42

Attribute	Value
Country	NL
Locality	Amsterdam
Organization	Vrije Universiteit
OrganizationalUnit	Comp. Sc.
CommonName	Main server
Host_Name	zephyr
Host_Address	137.37.20.10

(b)

Figura 43: Esempio di DIT 43(a) e di valori restituiti da un operazione di list e di read 43(b)

suddiviso su più server chiamati **directory server agent (DSA)** mentre i client sono chiamati **directory user agent** o DUA i quali sono simili ai name resolver nel sistema DNS.

6.5 Removing

I sistemi di naming fin qui descritti permettono l'accesso a delle entità distribuite molte volte in maniera globale. Ma quando queste entità non vengono più referenziate devono essere eliminate, per fare ciò solitamente si utilizza un garbage collector ma in un ambiente distribuito le cose si complicano alquanto.
Esistono tuttavia diversi meccanismi per capire se un entità è ancora differenziata.

6.5.1 Reference counting

In questo meccanismo gli oggetti tengono conto di quanti altri oggetti possiedono un loro riferimento, il sistema risulta molto efficiente se il conteggio è fatto tramite l'invio di un solo messaggio come mostrato in Figura 44 (a) ma purtroppo si potrebbero verificare dei problemi di *race condition* quando si passano dei riferimenti tra processi. Una tecnica è quella di comunicare all'oggetto anche il passaggio di riferimento ad altri processi come mostrato in Figura 44 (b) questo meccanismoo purtroppo degrada le prestazioni in quanto i messaggi scambianti diventano tre.
Un meccanismo simile ma che evita la race condition è quello che utilizza dei pesi per gli oggetti e comunica soltanto il decremento di tali pesi come mostrato in Figura 45 dove il peso di un oggetto viene diviso tra i due processi che usano tale oggetto, in questo caso al processo P_2 il riferimento viene passato dal processo P_1. Quando il peso totale e il peso dell'oggetto tornano a pareggiarsi l'oggeto può essere rimosso in quanto non più referenziato. Il problema di questa tecnica è che sono possibili solo un numero limitato di riferimenti. La soluzione è quella di concatenare i riferimenti come mostrato in Figura 46; questo risolve il problema dei riferimenti limitati ma aggiunge un hop per l'accesso all'oggetto.

6.5.2 Reference listing

Un altro meccanismo molto utilizzato è quello del reference listing che non tiene traccia del numero di referenze ma soltanto dell'identità di chi richiede un riferimento, il vantaggio è che

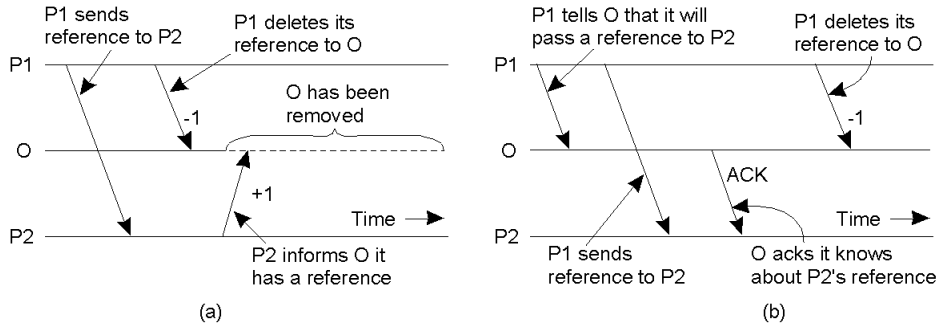


Figura 44: Esempio di reference counting con lo scambio di uno (a) e tre (b) messaggi

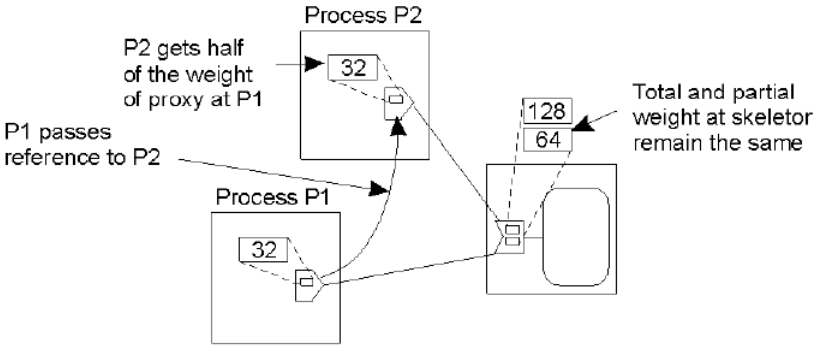


Figura 45: Esempio di reference counting con il meccanismo dei pesi

l'inserimento o l'eliminazione di un proxy (inteso come processo che richiede il riferimento) è idempotente ovvero inserimento e cancellazione di un proxy richiedono un messaggio di ack ma le richieste multiple non hanno effetto sul carico di rete. Il secondo vantaggio è che le risorse orfane possono essere individuate facilmente pingando periodicamente i client presenti nella lista dei riferimenti. Tuttavia anche questo meccanismo soffre del fenomento di race condition quando viene copiato un riferimento.

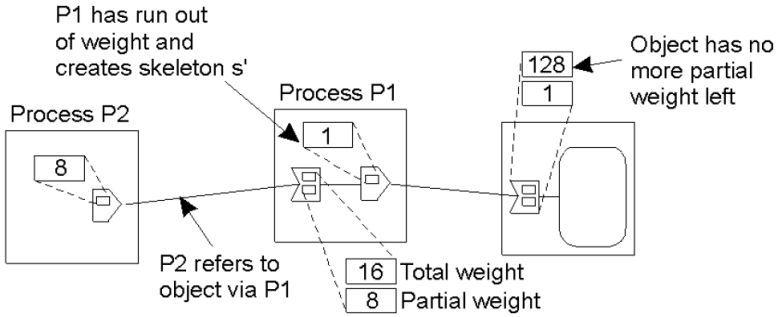


Figura 46: Esempio di reference counting con il meccanismo dei pesi e la concatenazione

6.5.3 Distributed mark-and-sweep

Il mark and sweep permette di tenere traccia di quelle entità orfane. Nel caso di sistema uniprocessore tale meccanismo si suddivide in due fasi, nella prima fase vengono marchiate tutte le entità accessibili da qualche tipo di referenza. Tutti i nodi partono colorati di bianco, un nodo viene colorato di grigio quando è possibile raggiungerlo dalla root ma alcune delle sue referenze non sono ancora state valutate, infine viene marcato di nero se tutte le sue referenze sono marchiate di grigio.
Nella seconda fase il garbage collector elimina tutti i nodi marchiati di bianco.
Nel caso di sistema distribuito il garbage collector entra in funzione su ogni nodo, partendo da un nodo P una risorsa viene marchiata di grigio se è possibile raggiungerla dalla root del nodo P quando un proxy q è marchiato di grigio si invia un messaggio a tutti i suoi riferimenti. Quando tutti i riferimenti rispondono con un ack il proxy q viene marchiato di nero.

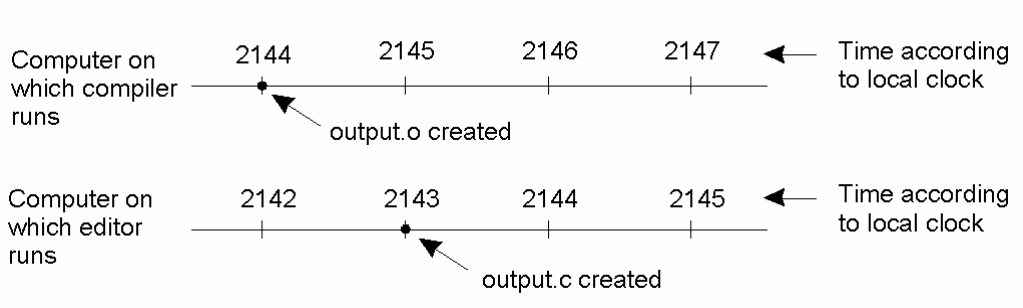


Figura 47: Assegnamento di istanti di modifica su due macchine diverse

7 Sincronizzazione

Nei capitoli precedenti abbiamo visto come avviene la comunicazione tra diversi processi, tuttavia, anche se questa gioca un ruolo importante nei sistemi distribuiti non è tutto. Strettamente correlato alla comunicazione è il modo in cui i processi cooperano tra loro e si sincronizzano.
La cooperazione è ottenuta parzialmente tramite il naming che permette ai processi di condividere le risorse o comunque le entità. In questo capitolo ci concentreremo su come i processi possono sincronizzarsi . Partiremo dalla sincronizzazione basata sul tempo reale per poi proseguire con la sincronizzazione in cui conta solo l'ordine relativo. Infine analizzeremo come un gruppo di processi possa eleggere un coordinatore per mezzo di algoritmi di elezione.

7.1 Sincronizzazione nei sistemi distribuiti

Il problema della sincronizzazione si presenta anche nei sistemi non distribuiti, tuttavia la distribuzione complica molto le cose in quanto vi sono alcune caratteristiche che nei sistemi centralizzati o monoprocessore non si presentano come:

- L'assenza di un clock fisico globale.
- L'assenza di un area di memoria condivisa.
- La possibilità di avere dei fallimenti parziali

In un sistema centralizzato il tempo non è ambiguo, quando un processo vuole conoscere data e ora attuali esegue una chiamata di sistema e il kernel glie la indica. Se un processo A richiede la data e l'ora e un processo B la richiede poco dopo il valore che B otterrà sarà leggermente superiore di quello del processo A. In un sistema distribuito invece non è semplice mantenere un'ora e una data comuni a tutte le macchine.
Come esempio prendiamo le implicazioni di un orario globale per la funzione *make* di UNIX. Di solito i programmi molto grandi sono divisi in più file sorgenti, per non dover ogni volta compilare tutti i file il *make* esamina la data e l'ora dell'ultima modifica del file sorgente e di quello oggetto. Se il file sorgente *input.c* è stato modificato all'istante 2051 mentre il file oggetto *input.o* è stato modificato all'istante 2050 il *make* sa che il file sorgente è stato modificato e lo deve quindi ricompilare.
Supponiamo ora di spostarci in un sistema distribuito in cui non ci siano una data ed un'ora globali. Supponiamo che il file *output.o* abbia associato l'istante 2144; il file *output.c* viene modificato su una macchina diversa che ha il clock leggermente in ritardo rispetto a quella dove risiede *output.o* e perciò gli viene assegnato l'istante 2143 come mostrato in Figura 47. Se ora

eseguiamo il *make* esso non richiamerà il compilatore in quanto l'istante di modifica del file sorgente e precedente a quello del file oggetto procurando notevoli problemi al programmatore.

7.1.1 Orologi fisici

Praticamente tutti i computer hanno un circuito per tener traccia del tempo; nonostante l'ampio uso che si fa della parola clock per indicare questi dispositivi non si tratta realmente di orologi ma si tratta più di **timer**. Tale timer di solito è un cristallo di quarzo che oscilla a una frequenza ben definita se sottoposto a tensione, che dipende dal tipo di cristallo, da come è tagliato e dal livello di tensione applicato. Ad ogni cristallo inoltre sono associati un **contatore** *(counter)* ed un **registro di mantenimento** (*holding register*). Ogni oscillazione del cristallo decrementa il contatore di uno. Quando il contatore raggiunge lo zero viene generato un interrupt e il contatore viene resettato con il valore contenuto nell'holding register. Ogni interrupt è chiamato **colpo di clock** (*clock tick*).
Con un solo computer non importa se questo orologio sia leggermente in ritardo o leggermente in anticipo in quanto tutti i processi sulla stessa macchina utilizzano lo stesso clock, infatti ciò che conta veramente sono gli istanti relativi.
Nel momento, invece, in cui si utilizzano più CPU, ciascuna con il proprio clock la situazione cambia radicalmente. Anche se la frequenza a cui oscilla un cristallo è abbastanza stabile è impossibile garantire che cristalli su computer diversi oscillino alla stessa frequenza; ciò porterà gli orologi logici ad andare lentamente fuori sincrono. Questa variazione nei valori della data e dell'ora è chiamata **disallineamento dei clock**.
Per alcuni sistemi, come quelli *real-time* è importante il valore reale dell'orologio e perciò si utilizzano degli orologi fisici esterni. Per motivi di affidabilità è però più sicuro utilizzare più di un orologio fisico il che porta a problematiche di sincronizzazione sia tra i diversi clock sia con gli orologi del mondo reale.
Prima di rispondere a queste domande analiziamo brevemente come funziona il tempo reale. A partire dal diciassettesimo secolo il tempo era calcolato in termini astronomici, ovvero la durata di un giorno era calcolata come l'intervallo tra due **culmini del sole**, tale intervallo è chiamato **giorno solare**. Dato che un giorno è composto da 86400 secondi un **secondo solare** è definito come 1/86400*esimo* di un giorno solare.
Con l'invenzione dell'orologio atomico nel 1948 è divenuto possibile misurare il tempo tramite più accuratamente e in maniera indipendente dai movimenti della terra. Attualmente molti laboratori nel mondo possiedono un orologio al Cesio 133 e periodicamente comunicano ad un laboratorio centrale (BIH) il loro numero di tic. Questo laboratorio calcola una media di questi valori e definisce il **tempo atomico internazionale** abbreviato con **TAI** come il numero mendio di scatti degli orologi al Cesio 133. Tuttavia il TAI è leggermente più breve di un secondo solare e ogni 86400 secondi TAI sono 3 msec in meno di un giorno solare medio. Per risolvere questo problema il BIH introduce periodicamente dei **secondi intercalari** conosciuti come **leep seconds** ogni volta che la discrepanza tra TAI e tempo solare raggiunge gli 800ms.

7.1.2 Global positioning system

Come passo per arrivare alla vera sincronizzazione dei clock consideriamo un problema che la riguarda, vale a dire la determinazione della posizione geografica di una persona o di un oggetto sulla terra. Questo problema è risolto mediante l'utilizzo del **global position system** o **GPS**. Il GPS è un sistema distribuito basato su un sistema di 29 satelliti in orbita a 20.000 km di altezza. Ogni satellite possiede fino a quattro orologi atomici, regolarmente calibrati da speciali stazioni sulla terra.

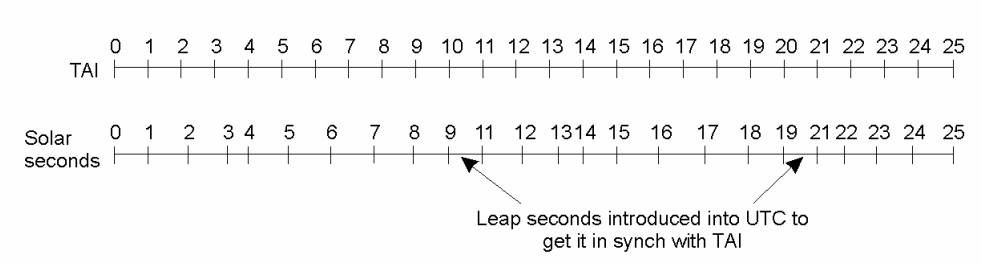


Figura 48: Paragone tra secondi TAI e secondi solari

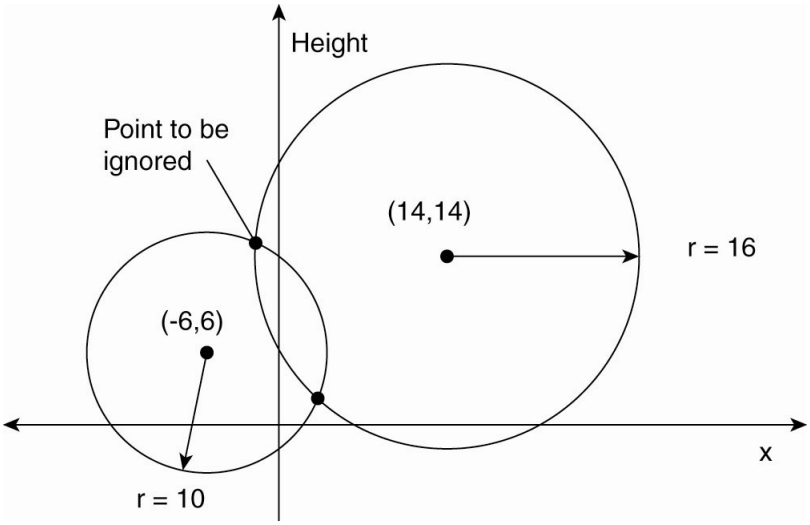


Figura 49: Calcolo della posizione in uno spazio bidimensionale

Ogni satellite diffonde continuamente la sua posizione e contrassegna ogni messaggio con il suo *timestamp* locale. Questa diffusione consente ad ogni soggetto sulla terra di calcolare con accuratezza la sua posizione.
Partiamo dal caso bidimensionale mostrato in Figura 49 in cui sono segnati due satelliti e le circonferenze che rappresentano i punti alla stessa distanza rispetto al satellite. L'asse y rappresenta l'altezza mentre l'asse x rappresenta una linea retta sulla terra a livello del mare. Ignorando il punto più in alto vediamo come l'intersezione delle circonferenze porti ad un punto univoco. Il principio dell'intersezione delle circonferenze può essere esteso a 3 satelliti per conoscere longitudine latitudine e altitudine. Ora supponiamo che gli orologi dei satelliti non siano perfettamente sincronizzati. Dobbiamo tener conto inoltre anche di due caratteristiche:

- ci vuole un certo tempo perchè i dati sulla posizione del satellite raggiungano il soggetto;
- l'orologio del soggetto in genere non è sincronizzato con quello del satellite.

Supponiamo che il *timestamp* del satellite sia assolutamente accurato. Definiamo ora Δ_r come la deviazione dell'orologio del soggetto rispetto al tempo reale, T_i è il timestamp del messaggio ricevuto dal satellite i e il tempo di trasferimento di tale messaggio è indicato come Δ_i che viene misurato dal soggetto ed è dato dal tempo di trasferimento reale e dalla sua deviazione. Abbiamo quindi che:

$$T_{now} = T_r - \Delta_r$$
$$\Delta_i = T_r - T_i$$

$$\Delta_i = (T_{now} - T_i) + \Delta_r$$

Dato che i segnali viaggiano alla velocità della luce c la distanza del satellite è data da $c\Delta_i$:

$$d_i = c\Delta_i = c \times (T_now - T_i) + c \times \Delta_r$$

La distanza calcolata dalla prima parte dell'espressione può essere rappresentata come

$$\sqrt{(x_i - x_r)^2 + (y_i - y_r)^2 + (z_i - z_r)^2}$$

Se prendiamo in considerazione quattro satelliti otteniamo un sistema di quattro equazioni e quattro incognite $(x_r, y_r, z_r, \Delta_r)$

7.1.3 Algoritmi di sincronizzazione dei clock

Se una macchina ha un ricevitore WWV l'obiettivo è quello di mantenere le altre macchine il più possibile allineate a questa. Sono stati proposti diversi algoritmi che tuttavia si basano sullo stesso sistema. Si suppone che ogni macchina abbia un timer che solleva un interrupt H volte al secondo. Quando il timer scade, il gestore dell'interrupt aggiunge 1 all'orologio software. Chiamando C questo valore e t il valore dell'UTC in un determinato istante avremo che in un mono perfetto $C_p(t) = t$ dove il pedice p indica una determinata macchina. Definito inoltre $C'_p(t) = dC/dt$ la frequenza del clock di p al tempo t il **disallineamento** del clock viene definito come $C'_p - 1$ e denota qual'è lo scostamento della frequenza dal clock perfetto. L'**offset** relativo ad un determinato istante t è dato da $C_p(t) - t$.

Algoritmo Cristian Un approccio comune a molti protocolli è quello di lasciare che i client contattino un *time server*. Quest'ultimo può fornire la data e l'ora attuali con precisione. Il problema principale di questo algoritmo è che il tempo di trasferimento dei messaggi rende obsolete data e ora. Il trucco è di trovare una stima valida per questo tempo di trasferimento. Considerando la Figura 50. In questo caso A invia una richiesta a B e la contrassegna con il *timestamp* T_1, B a sua volta memorizza l'istante di ricezione T_2 e restituisce una risposta contrassegnata con il timestamp T_3 che oltre all'orario si porta dietro anche il valore di T_2. Infine A memorizza l'istante di arrivo T_4; supponendo che il tempo di trasmissione da A a B sia pressoché uguale a quello da B ad A il che significa che $T_2 - T_1 \approx T_4 - T_3$ allora si può stimare l'offset relativo a B come

$$\theta = T_3 - \frac{(T_2 - T_1) + (T_4 - T_3)}{2} = \frac{(T_2 - T_1) + (T_3 - T_4)}{2}$$

Se il clock di A è veloce otterremo un $\theta < 0$ questo significa che il clock di A teoricamente dovrebbe tornare indietro. Questo però è impossibile in quanto provocherebbe potrebbe provocare parecchi problemi. Il cambiamento perciò deve essere introdotto gradualmente.
Nel **network time protocol** o NTP, che è un algoritmo basato su quello di Cristian si utilizzano otto stime dell'ora e si sceglie quello che ha un tempo medio di trasferimento minimo.

Algoritmo di Berkeley In molti algoritmi il *time server* è passivo, sono le altre macchine che periodicamente chiedono l'ora. Tutto ciò che esso fa è rispondere alle richieste. In UNIX Berkley si esegue l'approccio esattamente opposto. In questo caso il *time server* è attivo e di tanto in tanto richiede data e ora a tutte le macchine. In base alla risposta esso calcola un tempo medio e lo comunica alle altre macchine le quali si devono adeguare. Questo metodo è adatto per quelle macchine che non hanno un ricevitore WWV ma in questo caso la data e l'ora

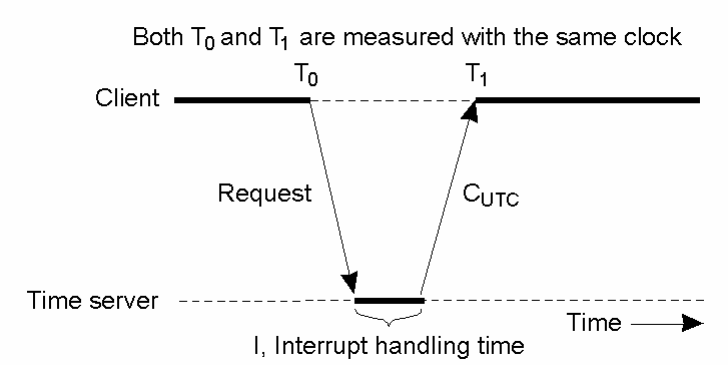


Figura 50: Esempio di scambio dei messaggi nell'algoritmo cristian

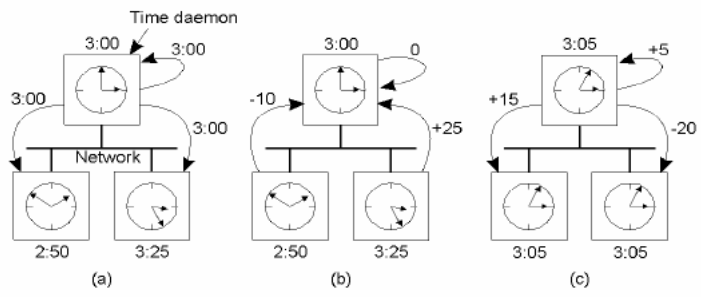


Figura 51: Esempio di algoritmo di berkley in (a) il server richiede l'ora, calcola una media (b) e invia le correzzioni (c)

dei *time daemon* devono essere impostate manualmente dall'operatore. Un esempio di questo algoritmo è mostrato in Figura 51 In questo sistema non è essenziale che il tempo corrisponda a quello reale, se la rete è chiusa ovvero non ci sono comunicazioni con altri computer su internet non vi sarebbe alcun danno.

7.2 Orologi logici

Fino ad ora abbiamo supposto che la sincronizzazione sia naturalmente correlata con il tempo reale. Tuttavia abbiamo anche notato che la cosa importante è che i nodi concordino sulla data e sull'ora attuali senza che esse corrispondano a quelli reali (Berkley).
Riprendiamo ora il caso del *make*, l'importante è che i nodi concordino sul fatto che il file *input.o* sia diventato obsoleto a causa di una nuova versione del file *input.c*. In questo caso l'unica cosa importante è tener traccia dei reciproci eventi. Per questi algoritmi si parla di **orologi logici**.

7.2.1 Clock scalari

Per sincronizzare gli orologi logici, Lamport ha definito una **relazione di precedenza** (*happens-before*) che viene indicata da $a \to b$ e si legge come a precede b e sta ad indicare che tutti i processi concordano sul fatto che prima accade l'evento a e poi accade l'evento b. Esistono alcune situazioni in cui la relazione di precedenza può essere osservata:

1. Se a e b sono eventi dello stesso processo e a precede b allora $a \to b$ è vera.

2. Se a è l'evento invio di un messaggio da parte di un processo e b è l'evento di ricezione del messaggio da parte di un altro processo, allora $a \to b$ è di nuovo vera. Infatti è impossibile che un messaggio sia ricevuto prima di essere inviato.

La relazione di precedenza è transitiva, per cui se $a \to b$ e $b \to c$ allora $a \to c$. Se x e y accadono in processi differenti che non si scambiano messaggi allora $x \to y$ non è vera ma non lo è nemmeno il suo opposto $y \to x$. In questi casi si dice che questi eventi sono **concorrenti** il che significa che non si può dire nulla su quanto accade.

8 Tolleranza ai guasti

8.1 Introduzione alla tolleranza ai guasti

8.1.1 Concetti base

8.1.2 Modelli di guasto

8.1.3 La ridondanza

8.2 Comunicazione client server affidabile

8.2.1 Comunicazione punto-a-punto

8.2.2 RPC in presenza di fallimenti

8.3 Protezione contro i fallimenti

8.3.1 Elementi di progettazione

8.3.2 Mascheramento dei guasti e meccanismi di replica

8.3.3 Accordo nei sistemi guasti

8.3.4 Rilevamento dei guasti

8.4 Comunicazione affidabile nei gruppi

8.4.1 Multicasting affidabile

8.4.2 Scalabilità del multicasting affidabile

8.4.3 Multicasting atomico

8.5 Commit distribuiti

8.5.1 Commit a due fasi

8.5.2 Commit a tre fasi

8.6 Tecniche di ripstistino

8.6.1 Introduzione

8.6.2 Creazione di checkpoint

8.6.3 Logging dei messaggi

9 Consistenza e replicazione

9.1 Introduzione

9.2 Modelli di consistenza data-centrici

9.2.1 Consistenza sequenziale

9.2.2 Consistenza causale

9.2.3 Consistenza *release* e *entry*

9.3 Modelli di consistenza client-centrici

9.3.1 Eventual consistency

9.3.2 Monotonic read

9.3.3 Monotonic write

9.3.4 Read your writes

9.3.5 Write follow reads

9.4 Gestione delle repliche

9.4.1 Repliche server-initiated

9.4.2 Repliche client-initiated

9.4.3 Protocolli pull e protocolli push

9.5 Protocolli di consistenza

9.5.1 Protocolli primary-based

Protocolli remote write

Protocolli local write

9.5.2 Protocolli replicated-write

Active replica

Protocolli quorum-based