MSA - 3.03.2021

Abbiamo iniziato ad affrontare le tematiche del corso, dando un punto di riferimento, il livello 0 che useremo come termine di paragone dei discorsi che faremo. Un sistema fisso, l'architettura definita nei due decenni che vanno dagli anni '70 agli anni '90, caratterizzato da un insieme di componenti che hanno una componente fissa e che utilizzano cavi per realizzare interazioni tra le componenti del sistema. Come prima cosa, proviamo a ripercorrere lo stesso cammino dal punto di vista del sistema di cui ci occuperemo. Caratteristiche fondamentali degli elementi che compongono le infrastrutture di cui ci occuperemo e le soluzioni che potremo trovare per le varie problematiche che discuteremo. I nodi di elaborazione, hanno ua grande varianza di potenza e di memoria RAM. Potremmo avere a che fare con nodi che rispetto a questi due parametri si collocano a distanza amplia da questi due parametri. Stesso discorso per le modalità di interazione di noi umani con i dispositivi. A seconda delle caratteristiche, e dimensioni consentite, queste interazioni possono essere molto primitive, unicamente di tipo testuale per esempio, fino a modalità sofisticate corrispondenti a quelle che siamo abituati ad aspettarci da dispositivi desktop.

Un aspetto fondamentale che distingue molte delle entità computazionali con cui avremo a che fare è che la disponibilità di energia non può più essere considerata illimitata. Tipicamente questi dispositivi fanno affidamento su batterie. Per dare un'idea tangibile delle diverse capacità dei nodi con cui abbiamo a che fare, possiamo avere nodi di tipo sensore, smartphone o laptop, esempi mediati rispetto alle tecnologie attualmente disponibili. Non sono dati relativi specificamente ad un particolare dispositivo, ma ispirati a dispositivi attuali. Come vediamo la variabilità dei parametri di CPU, memoria, capacità di comunicazione, alimentazione ecc. sono estremamente differenti. Dal punto di vista della disponibilità di energia, con cui dovremo fare i conti, questa figura semplifica come si sono evolute le capacità delle batterie rispetto agli altri parametri. Il grafico, su scala logaritmica, indica che l'aumento di capacità di batteria è molto lento nel tempo. Mentre le capacità di calcolo si sono raddoppiate ogni anno e mezzo, corrisponde alla crescita esponenziale nel grafico, invece le capacità delle batterie sono, se mai, raddoppiate ogni 10 anni. Il ricercatore raggiunge che le aspettative che abbiamo rispetto a ciò che un dispositivo mobile dovrebbe essere in grado di fare si colloca un po a metà tra queste due, anche se l'ultima affermazione non è tutta vera. In ogni caso il problema della disponibilità di energia all'interno dei sistemi che andiamo a considerare è un problema rilevante.

Questo dal punto di vista dei nodi di elaborazione. Andiamo a guardare cosa succede rispetto alla capacità delle modalità di interazione, tipicamente wireless. Utilizzano una certa porzione dello spettro elettromagnetico come supporto fisico. Quello che possiamo osservare è anche qui una grande variabilità nella capacità di comunicazione che queste tecnologie offrono. Caratteristica comune per gran parte delle tecnologie è di soffrire di latenze più alte, mediamente, e tassi di errore fisici notevolmente più alti rispetto alle comunicazioni basate su cavo.

Al di là di questi parametri che sono peggiorativi rispetto ai parametri medi di una comunicazione su caso, una caratteristica di una rete wireless è, se noi ci focalizziamo su una singola tecnologia wireless, possiamo osservare che la capacità di comunicazione non è costante nel tempo, ma varia anche notevolmente, come mostra il grafico su un'osservazione su un arco di tempo di una decina di secondi. Il throughput osservato varia nell'arco di poche decine di secondi. Giusto per dare un'idea di come questi parametri hanno un impatto sulle prestazioni che una rete wireless può offrire, abbiamo tre esperimenti molto "casarecci", fatti da un ricercatore. Sono simulazioni su una rete wired, ma servono in modo molto immediato a rendere visibile l'impatto che i parametri peggiorativi di una rete wireless hanno sulle prestazioni finali che l'utilizzatore osserva. Il ricercatore ha provato a misurare quanti download di un file di 130Kb riesce a fare in 5 minuti, partendo da una situazione base in cui utilizza la rete wired al meglio delle sue capacità, ed introducendo dopo delle latenze via via più alte. Osserviamo che man mano che la latenza viene fatta crescere, il throughput diminuisce in modo esponenziale. Un'altra causa di peggioramento è dovuto al maggiore tasso di errore, di cui una rete wireless soffre rispetto ad una rete wired. Stesso esperimento fatto in casa, con percentuali di scarto dei pacchetti ricevuti, anche in questo caso il throughput decresce drammaticamente.

Altro aspetto rilevante, ritorna il problema dell'alta variabilità. Osserviamo nel grafico il tempo di risposta osservato da un dispositivo che comunica con il resto del mondo tramite un canale wireless. Qui vediamo che il tempo di risposta osservato ha una varianza molto alta, da 0.5 a 3 secondi nell'arco di una 10ina di secondi. Il grafico fa anche vedere che dopo una prima fase in cui si osserva questa amplia variazione dei tempi di risposta, da un certo punto la situazione si stabilizza naturalmente. Che cosa è successo? "E' pronto il the". Cosa c'entra? Il te era in preparazione in un forno a microonde, che emette radiazioni elettromagnetiche che evidentemente interferiva con le capacità trasmissive del canale wireless utilizzato. Questo ci dice che una comunicazione wireless è molto più suscettibile di disturbi da parte dell'ambiente circostante piuttosto di una comunicazione su cavo, in buona misura schermata da eventuali disturbi dall'ambiente circostante.

A queste problematiche intrinseche si aggiunge quelle dovute all'architettura attualmente utilizzata. In buona misura, attualmente le reti cellulari sono concepite in questo modo, anche se proprio a causa delle problematiche di questo tipo di architettura si sta cercando nelle ultimissime generazioni di cambiare l'approccio architetturale. Giusto per dare un'idea di quale impatto hanno queste scelte architetturali, sempre tramite una rilevazione, un ricercatore si trovava a Las Vegas e per interagire con un altro dispositivo mobile sempre nella stessa area, il viaggio che facevano i pacchetti è quello presente nell'immagine. Da Las Vegas arrivavano in California e dopo tornavano indietro, introducendo una latenza notevole. L'effetto risultate di questi fattori è che le infrastrutture mobili sono state e continuano a stare un passo indietro rispetto a quello che offrono i sistemi fissi contemporanei. Lo si è osservato nel passato e continua a valere nel presente, ad esempio prendendo un esempio con rete disponibili 3G. In questo esperimento è stato osservata la mediana del tempo di caricamento delle pagine web di 200 siti più popolari di ecommerce negli USA. Quello che vediamo è che due tipologie di smartphone, confrontate con le prestazione di un desktop con tecnologia DSL alla rete internet, vediamo che abbiamo una differenza di 4 e più secondi.

Altro dato che evidenzia questo gap, in questo lavoro si confrontano le esigenze computazionali fatte da alcune successive generazioni di videogiochi, passando da 8 a circa 26 GPixel per secondo, nell'arco di 4 anni. Le aspettative degli utenti sono cresciute in modo più che lineare, ma le capacità di un dispositivo mobile sono cresciute, ma la crescita della richiesta è stata più veloce della crescita dell'offerta della capacità di computazione grafica.

Altra evidenza, che si muove nello stesso ambito come applicazioni di riferimento presi videogiochi online, viene confrontato il tasso di refresh, **fps**, che può essere offerto da un dispositivo di tipo desktop rispetto ad un mobile di pari generazione. Quello che osserviamo è che mentre un desktop offre tassi di refresh ampliamene superiore ai 60fps, per tutti i livelli di risoluzione considerati, un dispositivo mobile non è quasi mai in grado di garantire un tasso di refresh maggiore di 30. Ultima evidenza, guardando dietro le quinte, sono state messe a confronto generazioni diverse di dispositivi mobili e computer desktop differenti, usando vari parametri. Il punteggio ottenuto facendo eseguire varie tipologie di benchmark, e vediamo che i computer dekstop a parità di generazione offrono prestazioni superiori rispetto ai contemporanei dispositivi mobili. Altro aspetto distintivo in negativo di un sistema mobile rispetto ad un sistema fisso è l'aspetto della sicurezza. Sia dal punto di vista delle comunicazioni che dal punto di vista dei dispositivi. Un canale wireless è un canale aperto, il segnale si espande nello spazio tridimensionale. Chiunque si trovi nei pressi di una sorgente wireless, può catturare questa comunicazione, a differenza della comunicazione su cavo. Dal punto di vista dei nodi per evidenti ragioni di portabilità, che implica maggiore possibilità di prendere direzioni indesiderate per i nostri dispositivi.

Se vogliamo riassumere tutto questo, riguardo ai parametri che andiamo a guardare per avere un'idea sulle prestazioni di una sistema, i parametri sono sempre gli stessi, ma i valori per i parametri sono mediamente inferiori per un'infrastruttura mobile. Inoltre la variabilità dei valori è molto più amplia in un sistema mobile rispetto ad un sistema fisso. Giusto per ricapitolare la linea evolutiva partendo dal nostro livello base, abbiamo detto che nei decenni dagli anni 70 ai 90 si è consolidata la problematica di computazione distribuita fissa. Dagli anni '90 si è iniziato ad affrontare la problematica della computazione mobile. A partire dagli anni 2000 si è cercato di avvicinarsi un po di più alla visione del pervasive computing. Continuando sulla scala temporale, le tecnologie si sono evolute ed a partire da "oggi" si parla di collective computing, una terminologia per caratterizzare la situazione esistente. Viene messo l'accento sul fatto che allo stato attuale possiamo vedere che comincia ad essere una realtà la disponibilità massiccia di risorse di calcolo che vanno al di là delle capacità di un singolo dispositivo mobile. La possibilità che mettendo a comune tutte queste risorse si possono creare funzionalità che vanno molto al di là che un singolo dispositivo è in grado di offrire.

Giusto per chiudere questa panoramica temporale, di nuovo una citazione dell'editoriale con cui ho aperto la lezione di ieri, che fa il punto sullo stato dell'arte attuale e su cosa possiamo aspettarci nel decennio che si è appena aperto. Nel decennio appena chiuso, dal punto di vista dei sistemi di cui andiamo ad occuparci è stato un anno con poche novità, solo miglioramenti incrementali di ciò che già c'era un decennio fa. La previsione che fa questo ricercatore è che siamo nelle condizioni di aspettarci un salto di qualità, per cui può essere che alla fine di questo decennio la situazione sarà paragonabile a quella che abbiamo

adesso quando confrontiamo uno smartphone di generazione attuale con i Nokia di anni fa. La previsione è che grazie a queste tecnologie evidenziate in grassetto, fra 10 anni avremo osservato un salto qualitativamente paragonabile.

Ultima nota, prima di iniziare discorsi più tecnici, torniamo un attimo ad un problema che è bene non trascurare. Torniamo a quella visione di cui vi parlavo ieri, espressa da questo ricercatore, sul fatto che l'ambizione a cui dovremmo aspirare come ingegneri informatici è quella di costruire ambienti che siano saturati in un certo senso di capacità di computazione e comunicazione, in modo non invasivo per gli essere umani ma integrati in maniera piacevole con le attività umane. L'obiettivo è di rendere questa mole di informazioni che potenzialmente si possono aggiungere altrettanto piacevole e soddisfacente di quella che proviamo quando passeggiamo in un bosco. Un'idea di quello ceh significhi integrare queste entità all'interno dell'ambiente umano. Quello che osserva questo ricercatore è che riguardo al cercare di saturare i nostri ambienti con capacità di computazione abbiamo fatto un grande lavoro, però se su questo primo punto abbiamo avuto successo, riguardo l'altro punto, l'integrazione piacevole di tutta questa ricchezza con l'attività umana, forse non possiamo ancora ritenerci completamente soddisfatti.

Uno dei punti cruciali in questa prospettiva è che se le capacità di calcolo sono cresciute enormemente, che hanno sempre più saturato alcuni ambienti in cui ci possiamo trovare a spendere una parte della nostra esistenza, a fronte di questa crescita c'è una risorsa che dal nostro punto di vista umano non è mai cresciuta e non crescere mai, ed è la nostra capacità di attenzione. Dai primordi dell'umanità fino ad oggi, le nostre capacità di prestare attenzione sono rimaste immutate. Sono cambiate a cosa prestare attenzione, ma la nostra capacità di concentrarci su qualcosa non è aumentata. Dobbiamo selezionare quali sono le cose meritevoli a cui dedicare la nostra limitata capacità di attenzione. Se aumentiamo il potenziale flusso di informazioni che ci raggiunge, dobbiamo trovare modi di farlo arrivare in maniera selettiva ed accurata. Da alcuni anni a questa parte nel mondo della ricerca informatica c'è una conferenza specifica dedicata sull'attenzione umana.

A questo punto la panoramica di tipo generale è terminata. Iniziamo a guardare con maggiore precisione i tempi di cui ci occuperemo nel nostro percorso. Come dicevo ieri, non saremo in grado di coprire tutto ciò che sarebbe interessante. Ci occuperemo alle tematiche relative alla creazione di infrastrutture wireless. Parleremo di problematiche relative a come realizzare la comunicazione tra dispositivi che hanno una collocazione non fissa nello spazio. In un'altra parte ancora ci muoveremo nel lato applicativo, affrontando alcuni principi che ci dovrebbero guidare nella progettazione di applicazioni che poggiano sull'infrastruttura attuale o quella del futuro prossimo. Per tutte le problematiche che affronteremo, dove possibile vedremo soluzioni consolidate in forma di standard reali, formalizzati da organismi di riferimento o standard di fatto, ormai accettati dalla comunità di riferimento. Come temi trasversali presteremo attenzione al risparmio energetico ed all'adattamento all'ambiente circostante. Piccolo discorso sul tipo di architettura di rete a cui faremo riferimento. Questa che vedete a destra sono il classico disegno della stratificazione architetturale su vari livelli, quella originaria a cui si è aggiunto il livello ulteriore del middleware. Questo disegno veicola con se un'informazione importante, però non ci da un'idea precisa di qual'è il ruolo ed il modo in cui si articola ognuno di questi livelli.

Da questo punto di vista, una figura grafica, un modo grafico di rendere più preciso il ruolo svolto da ognuno di questi livelli all'interno di un'infrastruttura generale è il modello a clessidra. I vari livelli sono stati espansi in modi diversi, per evidenziare una distinzione tra i ruoli svolti dai vari livelli. Ai livelli bassi osserviamo la coesistenza di un'enorme varietà di infrastrutture, varie tecnologie, sia dal punto di vista fisico che dei protocolli. Varietà anche ai livelli superiori. Il punto è che questi livelli devono interagire tra di loro. Come fare interagire in maniera efficace livelli dentro cui si annida un'eterogeneità così amplia? La soluzione di definire tanti modi diversi per quante sono le possibili coppie, si, potrebbe essere un'idea ma da un punto di vista dell'ingegnerizzazione sarebbe una follia. Crescerebbe in modo esponenziale la quantità di modelli da progettare. C'è bisogno di definire un punto di unificazione, qualcosa a cui convergono tutti i livelli superiori ed inferiori, in cui tutte le diversità si congiungono e poi si espandono nuovamente. Allo stato attuale nell'infrastruttura distribuita globale attuale questo punto di unificazione è definito dalla rete Internet, il protocollo IP.

Tutto ciò che c'è sopra e sotto si parlano tra di loro grazie alla presenza di questo protocollo unificante. Potrebbe sembrare un dato di fatto, ma non lo è. Il mondo non era così inizialmente, IP era uno tra i tanti possibili protocolli che si potevano proporre come punto di unificazione. Allo stato attuale c'è chi sta iniziando, anche da qualche tempo, a ragionare se sia il caso di continuare a mantenere questo punto unificante, facendolo eventualmente evolvere. Il protocollo si è adattato al mutare delle circostanze, ma c'è chi da qualche tempo si sta chiedendo se è il caso di continuare a puntare su questa via o se è il caso di riprogettare radicalmente l'infrastruttura del nostro sistema distribuito globale cambiando il punto di unificazione. Una lezione che dobbiamo tenere sempre presente è che non è detto che le soluzioni tecnicamente migliori siano poi quelle che risultano vincenti.

C'è questa figura che veniva dal mondo delle società di comunicazioni dell'epoca, che può essere interpretata come la resa senza volerlo ammettere delle grandi società di telecomunicazioni rispetto all'avvento del protocollo IP come protocollo di unificazione del sistema globale. Descrive nella visione delle società di telecomunicazioni è il sistema globale di comunicazione, dove la rete internet ha una posizione accessoria, come se fosse uno tra i tanti membri. La parte del leone è giocata dalle infrastrutture cellulari, quelle su cui all'epoca le società di comunicazione pensavano di avere il controllo. Quello che vediamo che ciò che consente alle varie reti cellulari di interagire tra loro è un core IP-based, che è un nome diverso per dire una rete che utilizza il protocollo IP. Loro non l'hanno voluta chiamare così, ma alla fine questo è internet. Ci da un po il segno della resa. Dopo aver proposto protocolli proprietari anche all'interno delle infrastrutture proprietarie il protocollo IP è stato utilizzato per fare interagire i vari componenti della rete.

Questo punto di unificazione è quello che consente a tutte le varie porzioni del sistema globale, che possiamo individuare su varie scale, tutti questi ambiti colloquiano tra di loro tramite l'infrastruttura globale che è la rete Internet. Visto che stiamo in un regime di mobilità come spostamenti da un ambito all'altro, possono essere gestiti in modo tale da mantenere la continuità delle interazioni.

Come si collocano rispetto al modello a clessidra i mattoni di cui ci occuperemo? Le tematiche relative a comunicazione wireless riguarda la parte inferiore, mentre la mobilità è affrontabile a vari livelli. Invece le due tematiche trasversali dell'adattamento e risparmio energetico, e l'architettura software che riguarda il modello superiore. Per ogni mattone abbiamo un brevissimo indice degli argomenti che tratteremo.

Ultima nota prima di iniziare ad entrare nel merito dei discorsi, il sistema di cui andremo a parlare è estremamente complesso. Progettare qualcosa che si poggia su questo sistema è un compito non banale. Non c'è mai un'unica soluzione per i problemi di cui discuteremo. Una parte fondamentale del lavoro è, a fronte di possibili soluzioni, confrontarle e scegliere la soluzione migliore. Fare questo significa riuscire a valutare rispetto a parametri e misure dove si collocano le soluzioni che abbiamo davanti. Ci sono vari strumenti con cui questo può essere fatto. L'esempio da mostrare è che a volta bastano strumenti semplici per ottenere comunque visioni anche in profondità.

L'esempio è preso da un lavoro fatto da un ricercatore. Il problema nell'esempio è quello di realizzare una comunicazione wireless tra due nodi, A e B. Questa tecnologia, qualunque, garantisce un certo ritardo per trasmettere informazioni da A a B, che dipende dalla distanza tra i nodi A e B, che chiamiamo R. Il ritardo possiamo immaginarlo funzione della distanza, che sia un ritardo che cresce con la distanza, che possiamo immaginare come una funzione non decrescente ma tipicamente quale sia la tecnologia wireless, possiamo immaginare una funzione convessa, polinomiale o esponenziale. Possiamo immaginare che questo tempo di risposta cresce in questo modo perché essendo la tecnologia wireless, crescendo la distanza, aumenta la probabilità di interferenze, come minimo quadratica.

L'idea è quella che quanto maggiore è la distanza, tanto maggiore è l'interferenza. Piuttosto che soffrire di questo aumento della probabilità di interferenza, potrei pensare di dividere la mia comunicazione in tappe successive, in modo che la probabilità di interferenza sia minore per ogni tappa. Il punto è: qual'è la lunghezza ottimale delle tappe che devo percorrere per trasmettere da A a B, sapendo che il tempo di risposta è una generica funzione crescente convessa della distanza, ed in assenza di qualunque altra informazione? La risposta è si, si può fare, usando strumenti già acquisiti.

Il ritardo che la trasmissione subirà sarà pari al numero di tappe (D/R) per il tempo necessario per coprire ogni tappa. Qual'è il valore di R che minimizza il tempo complessivo? Anche se non so come è fatta T, arriviamo a definire rapidamente una relazione. Qualsiasi T(R), crescente e convessa, deve soddisfare l'equazione. Pensiamo di avere un plot della funzione T(R), prendo un bel righello, traccio la curva tangente che parte dall'origine alla funzione T(R), ed il punto in cui la retta tocca la curva è il punto R*. Questo vale qualunque sia la funzione T(R), qualunque sia la tecnologia wireless. Questo per dire che, semplicissimo esempio, a volte basta decidere di farlo, non considerarlo una perdita di tempo, a volte basta poco. Questo tipo di valutazione non è mai una perdita di tempo.