Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea in Informatica

GREEN CLOUD COMPUTING:
una rassegna comparativa

Tesi di Laurea in Reti di Calcolatori

Relatore:
Chiar.mo Prof.
FABIO PANZIERI

Presentata da:
ALESSANDRO FRANCI

Correlatore:
Ill.mo Dott.
MAURO MINELLA

Sessione I
Anno Accademico 2009/2010

Questa tesi è stata stampata su carta riciclata.
Ogni copia ha risparmiato la produzione di 150g di CO_2.

Indice

1. Introduzione

Da diversi anni stiamo assistendo a una crescita esponenziale di Internet e degli host ad esso associati. Questa crescita ha portato alla nascita di nuovi tipi di tecnologie e di servizi offerti sul web, determinando così un inevitabile aumento dei datacenter: il numero di server è aumentato di sei volte in dieci anni [1,2]. Si pensi ad esempio alla nascita di tecnologie di memorizzazione remota di files e alla gestione serverside di documenti, o al graduale abbandono di protocolli quali POP3 in favore di soluzioni serverside quali IMAP, che rendono sempre di più il client un thin-client. In particolare, da pochi anni sta aumentando l'interesse verso soluzioni di Cloud Computing, ospitate da datacenter di grosse dimensioni.

Tuttavia l'aumento di richieste verso servizi di Cloud Computing ha incrementato anche la domanda di energia dei datacenter che ospitano tali servizi: questo aumento di richiesta energetica si pone come limite alla scalabilità di tali datacenter poiché, se da un lato la fortissima evoluzione degli ultimi decenni ha portato a processori sempre più piccoli e più veloci, dall'altro ha indotto un forte aumento della potenza dissipata per il calcolo: mentre un 486 dissipava circa 10 W, un Pentium IV ne dissipa 120, con un consumo energetico aumentato di un ordine di grandezza. Per avere un'idea dell'entità dell'energia consumata dai sistemi IT è sufficiente considerare che un moderno server blade consuma circa 1 kW, tanto quanto un condizionatore domestico acceso alla massima potenza. Conseguentemente, un rack di server blade, per esempio, formato da 5 scaffali con 8 unità ciascuno, consuma 40 kW, l'equivalente di una palazzina. Un data center di medie dimensioni consuma circa 250 kW, come un quartiere, mentre i grandi data center, che per esempio ospitano soluzioni di Cloud Computing, possono arrivare a consumare 10 MW, l'equivalente di una cittadina. L'alto consumo di

energia delle apparecchiature informatiche sta diventando un limite alla scalabilità dei data center per il Cloud dislocati in aree ad alta densità abitativa. La potenza elettrica richiesta sta crescendo dell'8-10% all'anno e i gestori della rete elettrica rischiano di non essere più in grado di convogliare così tanta energia in un'area ristretta di un centro urbano.

Oltretutto, come è facile intuire, questa richiesta energetica si traduce in forti impatti sui costi di gestione, anch'essi limiti per la scalabilità ai datacenter che ospitano soluzioni di Cloud Computing.

Questa crescita esponenziale ha tuttavia un effetto collaterale non trascurabile: l'aumento del numero di host e server implica un aumento della quantità di gas serra che essi producono indirettamente. Nel 2007 il carbon footprint complessivo del settore ICT è stato di 830 Mt di CO_2: un contributo analogo a quello del settore aeronautico civile. Ogni PC in uso produce all'incirca una tonnellata di CO_2 l'anno e un server consuma energia per la cui produzione viene emessa la stessa quantità di CO_2 prodotta da un SUV in 25 km [2]. Il dato è significativo poiché la crescita del settore ICT ha un tasso sicuramente maggiore di quella del settore aeronautico.

La CO_2 è un gas serra che intrappola i raggi del sole nella bassa atmosfera ed è di fondamentale importanza poiché in questo modo permette un clima adatto per la vita. Più la concentrazione di tale gas aumenta, più calore è intrappolato nell'atmosfera e meno ne può uscire verso lo spazio. L'eccessivo calore della bassa atmosfera tuttavia porta a rilevanti alterazioni delle condizioni climatiche, tra cui l'innalzamento del livello dei mari, lo scioglimento dei ghiacci polari e l'aumento della frequenza e dell'intensità delle precipitazioni in diverse regioni del globo.

Ad oggi, la concentrazione di CO_2 presente nell'atmosfera è la più alta che sia mai stata registrata negli ultimi 650,000 anni, cresciuta del 35% dalla rivoluzione

industriale e del 18% dal 1960 [2]. Molti scienziati sostengono che per stabilizzare la concentrazione di anidride carbonica nell'atmosfera - e quindi le temperature mondiali - bisogna ridurne le emissioni del 70-80% [2]. Nonostante ciò, l'International Energy Agency stima che nel 2030 la richiesta mondiale di energia sarà aumentata del 53%, portando ad un incremento del 55% delle emissioni di CO_2 [3].

La crescente consapevolezza che le emissioni di gas serra da parte dell'uomo sono un fattore importante per il riscaldamento globale, ha portato davanti alle imprese, ai governi e in generale alla società un nuovo importante ordine del giorno: affrontare le questioni ambientali e adottare pratiche ecocompatibili. Rendere "verdi" i nostri prodotti, applicazioni e servizi IT non è solo una pratica economica, ma anche un imperativo ambientale che viene dalla nostra responsabilità sociale.

Nasce perciò in questo contesto il "Green Computing", che si riferisce ad un'informatica ecologicamente sostenibile, detta appunto Green IT. Esso si occupa dello studio e della messa in pratica di tecniche di progettazione e realizzazione di computer, server, e sistemi connessi – come monitor, stampanti, dispositivi di archiviazione, reti e sistemi di comunicazione – efficienti e con impatti ambientali limitati. La Green IT si pone un duplice obiettivo: il raggiungimento di un tornaconto economico e di buone prestazioni tecnologiche, rispettando le nostre responsabilità sociali ed etiche; essa riguarda perciò la sostenibilità ambientale, l'efficienza energetica, il costo totale di proprietà (che comprende il costo di smaltimento e riciclaggio). La Green IT si occupa quindi di tutti i problemi legati all'impatto ambientale e al consumo energetico dei sistemi informatici, studiandone le tecnologie per renderle il più possibile efficienti.

Le origini del Green Computing risalgono al 1992, quando la US Environmental Protection Agency ha lanciato Energy Star, un programma di etichettatura volontaria con lo scopo di promuovere e riconoscere l'efficienza energetica nei monitor, apparecchiature di controllo del clima e altre tecnologie. Il termine "Green Computing" è stato probabilmente coniato poco dopo che il programma Energy Star ha avuto inizio.

Ad oggi, un numero crescente di fornitori IT e di utenti si stanno muovendo verso la Green IT aiutando in tal modo la costruzione di una società e un'economia ecologicamente sostenibili.

Il rapporto fra il consumo elettrico globale del data center e quello dei server è indicato come PUE (Power Usage Effectiveness); PUE uguale a 1 significa che tutta l'energia assorbita dall'impianto viene utilizzata dagli apparati IT.

Se il rapporto tra il consumo per il raffreddamento e il consumo effettivo del server aumenta il PUE cresce. Oggigiorno il PUE medio di un data center è 2.5 e raggiungere un PUE minore di 2 è considerato un ottimo risultato; ovviamente la condizione ottimale si verificherebbe con un PUE uguale a 1.

Adottando alcune politiche e soluzioni vincenti, Google Inc. nel suo Cloud App Engine ha raggiunto in uno dei suoi datacenter PUE di 1.13: per 1 watt usato per il calcolo, l'impianto ne assorbe 0.13. Anche Microsoft Corporation, nel suo Cloud Azure, ha totalizzato un ottimo PUE di 1.22, adottando alcune politiche simili a quelle in App Engine [5].

Negli attuali datacenter che ospitano soluzioni di Cloud Computing, l'energia complessiva è usata per il 55% dai sistemi di alimentazione e raffreddamento e per il 45% dal carico IT; in altre parole, per ogni Watt usato effettivamente per il calcolo, il processore consuma 5W, il server 16W e il data center nel suo insieme

27W [4]. Questo porta inevitabilmente ad uno spreco energetico, nonché a un'inutile emissione di CO_2 e ad un aumento dei costi di gestione del datacenter.

La tesi è strutturata in tre parti fondamentali: la prima parte sarà di tipo introduttivo e si concentrerà sui concetti di Green Computing e Cloud Computing; nella seconda parte verranno studiate alcune soluzioni alle problematiche energetiche esposte nella prima parte; infine la terza parte sarà costituita dallo studio e classificazione di alcune soluzioni di Cloud Computing esistenti.

Più in particolare, nella prima parte, vengono introdotti i concetti di Green Computing e Cloud Computing. Per quanto riguarda il Green IT sarà analizzato l'inquinamento totale dovuto all'IT e i costi energetici di quest'ultimo. Il Cloud Computing sarà introdotto descrivendo le categorie in cui questo paradigma si divide e gli eventuali vantaggi o svantaggi nell'adozione di una soluzione di tipo Cloud; verrà poi studiata l'architettura di una tipica Cloud Facility, ovvero l'architettura di un datacenter che ospita un Cloud, e i costi energetici all'interno di essa. Successivamente ne saranno studiate alcune delle più rilevanti problematiche energetiche. Infine verranno introdotti e studiati due standard di misurazione dell'efficienza energetica dei datacenter.

Nella seconda parte della tesi verranno studiate ed approfondite alcune politiche e metodologie per il miglioramento energetico all'interno del Cloud Computing e per ridurre i problemi di scalabilità: saranno analizzate sia soluzioni software (per esempio di bilanciamento del carico) che soluzioni fisiche (ad esempio la disposizione dei rack di server o la locazione del datacenter in luoghi geografici più freddi).

Infine nella terza parte della tesi saranno prese in esame ed analizzate alcune soluzioni di Cloud Computing esistenti, ovvero App Engine di Google Inc., Azure di Microsoft Corporation e EC2 di Amazon.com Inc.. Tali soluzioni verranno

confrontate fra di loro, creando una tassonomia sulle basi di alcune caratteristiche importanti per l'adozione di politiche Green volte al risparmio energetico, nell'ottica di un Cloud Computing "verde".

La metodologia usata per lo studio di questa tesi è basata sulla ricerca e l'analisi di articoli, studi e documenti scientifici, consultazione di riviste del settore; l'autore ha inoltre partecipato a una conferenza tenuta da Microsoft Corporation sul Cloud Azure ed ha avuto accesso a documenti di Microsoft Corporation riguardanti Azure.

2. Green Computing

Il Green Computing è una disciplina emergente che studia come rendere l'IT ecologicamente sostenibile; esso viene quindi definito come lo studio e la pratica nel design, produzione, uso e disposizione di computer, server e sottosistemi associati (come per esempio monitor, stampanti, device di archiviazione, device di rete) in modo efficiente con impatto sull'ambiente minimo o nullo; il Green Computing inoltre ambisce a migliorare l'attuabilità economica, l'uso e le performance dei sistemi, rispettando le nostre responsabilità sociali ed etiche.

Con la crescente consapevolezza che l'emissione dei gas serra prodotti dall'uomo sono la causa principale del riscaldamento globale del pianeta, imprese, governi e società hanno ora un nuovo importante ordine del giorno: affrontare i temi ambientali adottando buone pratiche e misure di prevenzione. Rendere più “verdi” i nostri prodotti IT, applicazioni e servizi ha un vantaggio sia dal lato ambientale e sociale, ma ha anche un grande vantaggio economico; per questo, un sempre maggior numero di fornitori e di utenti si stanno muovendo verso il Green Computing e quindi verso la creazione di una società e un'economia più verde.

Gli scopi principali del Green Computing sono quelli di ridurre l'uso di materiali dannosi per l'ambiente nei componenti hardware, massimizzare l'efficienza energetica durante la vita del prodotto e promuovere il riciclo o la biodegradabilità dei prodotti in disuso o degli scarti di fabbricazione.

La storia

Una delle prime manifestazioni del movimento del Green Computing è stato il lancio del programma Energy Star nel 1992, da parte della US Environmental Protection Agency con fondi governativi: Energy Star è un programma di etichettatura volontaria che ha lo scopo di promuovere e classificare l'efficienza energetica di computer, monitor, dispositivi di climatizzazione, elettrodomestici, sistemi di illuminazione e altre tecnologie; il programma copre tutti i problemi legati ai consumi energetici. Energy Star è stato sviluppato da Jhon S- Hoffman, inventore dei Green Program alla US EPA, e implementato da Cathy Zoi e Brian Johnson. L'intenzione era quella di renderlo il primo di una serie di progetti volontari, tra cui la Green Light e il Methan Program, volti a rendere di immediata riconoscibilità ai consumatori quei prodotti energicamente efficienti.

Energy Star ha aiutato a raggiungere moltissimi risultati nel mondo della tecnologia eco-compatibile, tra cui la diffusione di semafori a led, di efficienti forme di illuminazione fluorescente e dell'uso della modalità Sleep nei monitor, che pone lo schermo in modalità stand-by se non viene rilevata un'attività dell'utente per un certo lasso di tempo. Inizialmente si occupava solo di prodotti IT, per poi allargarsi in componenti elettronici di tutti i generi, dall'illuminazione agli elettrodomestici e persino agli edifici pubblici e privati. Nel 2006 circa il 12% di abitazioni negli stati Uniti è stata etichettata col marchio Energy Star. Il programma si è diffuso anche oltre gli States ed è stato adottato nell'Unione Europea, in Australia, Nuova Zelanda, Canada, Giappone e Taiwan; ormai è divenuto

Illustrazione 1: etichetta Energy Star

uno standard internazionale e generalmente i prodotti marchiati consumano il 20-30% in meno dell'energia richiesta dagli standard federali. La decisione se premiare o meno un prodotto con l'etichetta Energy Star viene presa secondo i seguenti principi:

- il prodotto deve contribuire significativamente al risparmio energetico globale
- il prodotto deve rispondere esattamente alle richieste e alle necessità del consumatore
- se il prodotto ha un costo maggiore di una controparte convenzionale e meno efficiente, esso deve ripagare l'investimento in maggiore efficienza energetica in un lasso di tempo ragionevole
- sono privilegiate le tecnologie non-proprietarie, accessibili a più di un "artigiano"

Il termine "Green Computing" è stato probabilmente coniato poco l'inizio del programma Energy Star: vi sono parecchi post su USENET risalenti al 1992 in cui viene usato questo termine.[6] Nello stesso tempo, l'organizzazione svedese TCO Development ha rilasciato la certificazione TCO per promuovere il minor consumo elettrico e le minori emissioni elettromagnetiche dei monitor CRT.[7]

Aree tematiche

Per una completa ed efficace riduzione dell'impatto ambientale dell'IT, il Green Computing adotta un approccio olistico rendendo l'intero ciclo di vita della tecnologia più ecologico; questi sono i quattro passi da percorrere:

- **Uso Green**: ridurre il consumo di energia da parte dei computer e degli altri sistemi informatici e utilizzarli in modo ecologicamente corretto
- **Smaltimento Green**: revisionare e riutilizzare i vecchi computer, riciclare tutti i dispositivi elettronici non reimpiegabili
- **Design Green**: la progettazione a basso consumo energetico e componenti dell'ambiente, computer, server, apparati per il raffreddamento e data center
- **Produzione Green**: realizzare componenti elettronici, computer e altri sottosistemi con un minimo impatto ambientale

Questi quattro percorsi abbracciano diverse aree di interesse e attività che includono:

- design per la sostenibilità ambientale
- computer energeticamente efficienti
- power management
- progettazione dei data center, della configurazione e della posizione

Settore	Efficienza 1978	efficienza 2008	Miglioramento
Automobili	6,15 km/lt	8,50 km/lt	x 1.4
Aerei	6,0 revenue passenger mile/lt	13,3 revenue passenger mile/lt	x 2.2
Produzione acciaio	132 g/Kj	349 g/Kj	x 2.7
Illuminazione	13 lumen/watt (incandescenza)	57 lumen/watt (fluorescenza)	x 4.4
Sistemi di calcolo	1,4 mips/watt	40.000 mips/watt	x 28.571

Tabella 1: Miglioramento energetico del settore IT rispetto ad altri settori

- virtualizzazione dei server
- responsabilità di smaltimento e riciclaggio
- conformità normativa
- misurazioni verdi, strumenti di valutazione e metodologia
- attenuazione dei rischi connessi all'ambiente
- utilizzo di fonti energetiche rinnovabili
- eco-etichettatura dei prodotti IT

I moderni sistemi informatici si basano su un mix complesso di persone, reti e hardware; quindi in quanto tale, un'iniziativa green IT deve essere di natura sistemica, e affrontare dei problemi sempre più sofisticati. Gli elementi che contraddistinguono una certa soluzione rispetto ad un'altra potrebbero comprendere la soddisfazione dell'utente finale, la ristrutturazione gestionale, la conformità alle normative, l'eliminazione dei rifiuti elettronici, telelavoro, virtualizzazione di risorse dei server, risparmio di energia, soluzioni di tipo thin-client. L'imperativo per le aziende di assumere il controllo del loro consumo di energia, per la tecnologia e più in generale, rimane quindi pressante.

Gartner [6] sostiene che il processo di fabbricazione di PC sfrutti il 70% delle risorse utilizzate nell'intero ciclo di vita del PC stesso; per questo i maggiori sforzi nel ridurre l'impatto ambientale dei PC vengono fatti per prolungare la vita dell'apparecchio stesso. Oltre allo sfruttamento delle risorse energetiche, il processo di fabbricazione di materiale IT è responsabile anche della dispersione di sostanze tossiche nell'ambiente: basti pensare che secondo alcune recenti ricerche [7] il 70% dell'inquinamento del suolo da piombo, cadmio e mercurio deriva direttamente o indirettamente dall'IT.

L'adozione di tecnologia verde in azienda porta un vantaggio sia ai fornitori che ai loro clienti: infatti l'uso di tecnologie e iniziative di Green Computing aiutano ad

abbassare i costi. Per questo esse stanno prendendo sempre più piede all'interno di aziende IT: nel Febbraio 2010 il 46% delle aziende britanniche riteneva che rendere efficiente il loro comparto IT fosse la chiave per ridurre le loro emissioni di gas serra; inoltre le aziende si stanno rendendo conto che adottare tecnologie verdi può portare a un vantaggio rispetto ai loro competitors [8].

Secondo una previsione di Gartner [6], entro il 2008 il 50% dei data center avrà problemi ad approvvigionarsi di energia, visto che oltre al naturale consumo elettrico di un server c'è bisogno di energia anche per dissipare il calore che produce il server stesso. Inoltre, una ricerca della società IDC sostiene che ormai si spende più per mantenere attivi e accesi i data center che per acquistare l'hardware: un server oggi consuma in media quattro volte la corrente che richiedeva 10 anni fa. E se si considera che le aziende usano sempre più reti composte da tanti piccoli server piuttosto che grossi archivi centralizzati, si comprende perché i consumi complessivi dei server in tutto il mondo sono raddoppiati in soli cinque anni. Tra le diverse aziende che hanno adottato politiche di Green Computing, citiamo fra i casi di successo, quello di Radiator Express Warehouse, un distributore di componenti automobilistici, che è riuscito a rimuovere 31 server fisici riducendo del 25% il consumo di energia; oppure quello di Qualcomm, società che si occupa di comunicazione wireless, che ha ridotto l'utilizzo dei server dell'80%.

Ma anche in Italia le società tecnologicamente più evolute hanno cominciato ad affrontare concretamente il problema. CRIF - società bolognese specializzata nello sviluppo e nella gestione di sistemi di informazioni creditizie, di business information e di supporto decisionale - ha implementato un processo di virtualizzazione che le ha consentito di ridurre il consumo di energia di oltre un Megawatt e mezzo all'anno (1575974,4 kwh per l'esattezza). Per comprendere la

portata di questo intervento, in termini di emissione di anidride carbonica è come se si fossero piantati 7000 alberi, oppure eliminati dalle strade 427 autoveicoli.

Moda o necessità?

Inquinamento dovuto all'IT

La concentrazione di CO2 nell'atmosfera è la più alta negli ultimi 650,000 anni ed è cresciuta almeno del 35% dalla rivoluzione industriale e del 18% dal 1960. L'anidride carbonica è un gas serra che intrappola i raggi solari nella bassa atmosfera, cruciale per mantenere un clima stabile per la sopravvivenza della vita sulla terra. Più la concentrazione di CO2 aumenta, più calore viene intrappolato nell'atmosfera e meno ne può uscire. Questo calore intrappolato altera il clima e le condizioni meteorologiche, causando l'innalzamento del livello del mare, lo scioglimento dei poli e porta a più severe e frequenti precipitazioni. L'uso a Londra delle barriere sul Tamigi è cresciuto di una volta ogni due anni fino a una media di sei volte all'anno negli ultimi 5 anni [9]. Questi cambiamenti possono portare a una massiccia estinzione delle specie e danni irreparabili per essere umani.

I gas serra persistono nell'atmosfera per anni, decenni o addirittura millenni prima che decadano verso la superficie terrestre. Quindi non si ha ancora avuto esperienze dell'impatto totale dei gas prodotti dall'uomo già rilasciati [10].

Gli scienziati credono che dovremmo ridurre l'emissione di CO2 dal 70 all'80% per stabilizzare la concentrazione di CO2 nell'atmosfera — e quindi le temperature del pianeta [10]. Nonostante questo, alcune analisi riportano che la richiesta energetica mondiale aumenterà fino al 53%, portando quindi a un aumento del 55% di emissioni di CO2 dalla produzione di energia al 2030 [11].

Gli apparati IT consumano energia perché la stessa trasmissione di informazione richiede intrinsecamente energia. Un bit , cioè l'unità minima di informazione, è associato allo stato di un sistema fisico (per esempio, la carica di un insieme di elettroni o il campo magnetico su di una porzione di disco) e per poterlo commutare occorre cambiare lo stato del sistema stesso e quindi consumare energia. Recenti ricerche condotte al MIT (si veda il Teorema di Margolus Levitin [10]) hanno dimostrato che esiste un limite minimo al consumo di energia necessario per commutare un bit ad una data velocità, che è dettato dalle leggi della fisica quantistica. Questo limite minimo è raggiungibile quando ogni bit è associato allo spin quantistico di un elettrone. Questo tipo di commutazione è effettuabile solamente all'interno di computer quantistici, cioè di particolari elaboratori attualmente ancora in fase di studio che sfruttano le proprietà quantistiche della materia [11], e richiedono un'energia di circa 1025 J per 1 bit a 1 GHz. Il consumo dei più avanzati chip tradizionali su cui si sta attualmente facendo ricerca nei laboratori è arrivato alla soglia dei 1016 J. Questi valori di energia sono comunque ordini di grandezza più bassi rispetto ai valori di consumo discussi precedentemente. Questa apparente incoerenza è facilmente spiegata dal fatto che in realtà la commutazione dei bit è solamente il livello più basso di un sistema informatico:

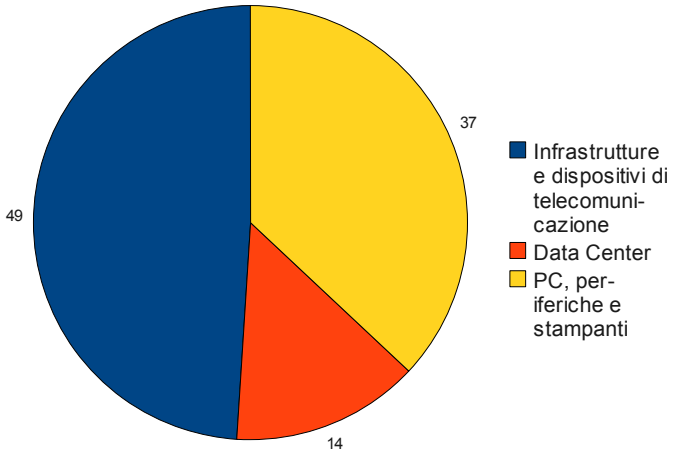


Illustrazione 2: percentuale sul totale di CO2 emessa

sopra ad esso esistono tanti altri livelli (layer) infrastrutturali che moltiplicano anche di un fattore 30 il consumo energetico della singola commutazione.

Il settore dell'IT può fare molto in questo contesto, iniziando a diminuire le proprie richieste energetiche — e quindi l'emissione indiretta di gas serra. Gartner, la compagnia di ricerca e consiglio per il settore dell'ICT, ha stimato che la produzione di prodotti ICT, il loro uso e il loro trasporto è stato l'artefice nel 2007 del 2% delle emissioni totali di CO2, che è equivalente all'apporto dato dal settore dell'aviazione [12]. La totalità della CO2 emessa dall'uomo è pari a circa 49 miliardi di tonnellate l'anno [13], quindi circa 1 miliardo di tonnellate proviene dal settore dell'ICT. Nel Regno Unito circa il 10% dell'energia totale richiesta è assorbita dal comparto dell'ICT [14], ovvero la potenza fornita da quattro centrali nucleari.

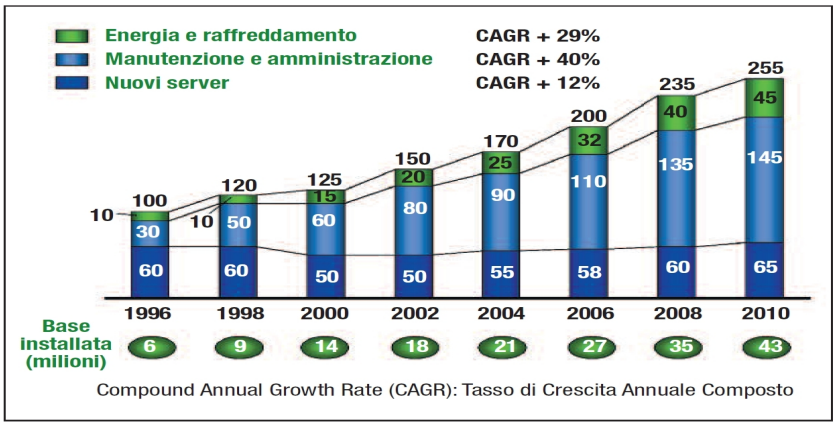
Ogni PC genera 1 tonnellata equivalente di CO2 all'anno e che un server consuma energia per la cui produzione viene emessa la stessa quantità di CO2 prodotta da un SUV che percorre 25 km [15]. La fortissima evoluzione degli ultimi decenni ha portato a processori sempre più piccoli e più veloci, ma ha anche indotto un forte aumento della potenza dissipata per il calcolo: mentre un 486 dissipava circa 10W, un Pentium IV ne dissipa 120, con un consumo energetico aumentato di un ordine di grandezza. Per avere un'idea dell'entità dell'energia consumata dai sistemi IT è sufficiente considerare che un moderno server blade consuma circa 1 kW, tanto quanto il frigorifero di casa. Conseguentemente, un rack di server blade, per esempio, formato da 5 scaffali con 8 unità ciascuno, consuma 40 kW, l'equivalente di una palazzina. Un data center di medie dimensioni consuma circa 250 kW, come un quartiere, mentre i grandi data center, che per esempio, servono grosse banche o internet service provider, possono arrivare a consumare 10 MW, l'equivalente di una cittadina. La considerevole crescita dei consumi energetici dell'IT sta sempre più attirando l'attenzione della comunità scientifica, dei

produttori di tecnologia e dei responsabili dei sistemi informativi delle aziende utenti.

Guardando l'altro lato della medaglia però, il settore dell'ICT può portare vantaggi all'ambiente: gli effetti benefici sull'ambiente si possono classificare come di primo, secondo e terzo ordine. I primi sono quelli diretti, derivanti dalla mera esistenza degli elaboratori e comprendono produzione, uso e smaltimento a fine vita. Gli effetti di secondo ordine discendono dalle applicazioni dell'ICT e includono l'ottimizzazione dei processi indotta in altri settori (per esempio, sul traffico), gli effetti di sostituzione (per esempio, la teleconferenza che elimina gli spostamenti) e gli effetti indotti, quando l'ICT crea più domanda in altri settori. Secondo Google, una query produce 0.2 g di CO2, ma l'uso di un laptop per un'ora ne produce 20 g, l'uso di un PC con monitor per un'ora 75 g . Una copia fisica di quotidiano invece ne produce 173 g, mentre andare da Parigi a Ginevra in TGV 13 kg e 56 kg in aereo (Economy Class) [16]. Un recente studio ha concluso che se il 20% dei viaggi di lavoro all'interno dell'Unione Europea fossero rimpiazzati da telecomunicazioni digitali dal 2010 potrebbe essere evitata la produzione di circa 25 milioni di tonnellate di CO2 all'anno.

La cattiva notizia è che il comparto dell'ICT sta crescendo molto più in fretta di altri settori (quali, riprendendo i dati precedenti, il settore aeronautico); però, a differenza del comparto dell'aviazione, piccole e semplici azioni possono essere intraprese per realizzare grandi risparmi. Per esempio, è stimato che il 35% di tutti i dati delle applicazioni sia duplicato [17]; nel 2006, sono state creati e copiati 161 exabytes (161 × 1012 bytes) di informazioni digitali. L'equivalente di tre milioni di volte le informazioni contenute in tutti i libri mai scritti. L'equivalente di 12 pile di libri, ognuna alta 93 milioni di miglia - la distanza dalla terra al sole. Nel 2010 tale quantità sarà sei volte più grande. L'idea, ampiamente diffusa negli ultimi decenni, che le risorse di storage dei dati abbiano costi trascurabili, ha

creato l'abitudine di duplicare e duplicare i dati, senza considerare le ridondanze inutili. Si indica come data de-duplication l'eliminazione dei dati inutilmente ridondanti. Nel processo di de-duplicazione i dati duplicati più volte sono eliminati in modo da lasciare solo una copia dei dati da memorizzare; l'operazione può arrivare a ridurre lo spazio disco per il back-up notevolmente. I livelli di ripristino di servizio sono più alti, gli errori da gestire diminuiscono in media e si rendono disponibili più punti di recovery sui media per il recovery veloce. La de-duplicazione dei dati riduce inoltre anche la mole di dati da inviare su WAN per back-up remoti, replicazione dati e disaster recovery. In uno studio del 2007 dell'EPA (Environmental Protection Agency) [18] sui data center si osserva che "le tecnologie esistenti e le strategie di progettazione hanno dimostrato come sia possibile ridurre del 25% il consumo energetico di un server"; senza sacrificare alcuna funzionalità è possibile programmare i PC in modo tale che quando non sono usati siano in stato energy-saving: la US Environmental Protection Agency (EPA) stima che lo sleep mode possa consentire risparmi nei consumi energetici del 60-70% [18, 19].



llustrazione 3: Spesa mondiale per i server (miliardi di dollari). Fonte: IDC (2006)

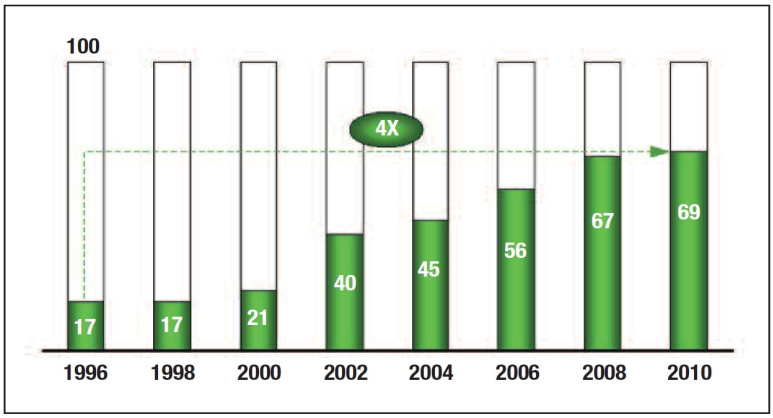


Illustrazione 4: Rapporto spesa per energia e raffreddamento - spese per l'acquisto di nuoviserver (Percentuale). Fonte: IDC (2006)

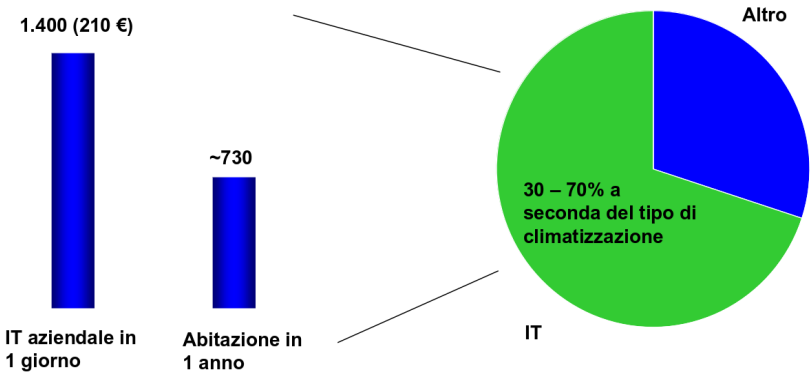


Illustrazione 5: Spesa energetica di un datacenter confrontata con quella di un'abitazione privata

I costi dell'energia elettrica

Il collegamento fra energia consumato dal settore dell'ICT e cambiamenti climatici è chiaro, come è chiaro il collegamento fra l'aumento della richiesta energetica e l'aumento dei costi.

Il costo dell'energia consumata dai sistemi IT copre una parte significativa del Total Cost of Ownership (TCO) dei sistemi ed è in continua crescita. L'illustrazione3 mostra i dati relativi alla spesa mondiale per i server negli ultimi anni e alcune stime per il futuro. Mentre il costo di acquisto dell'hardware negli ultimi dodici anni è cresciuto molto debolmente, il costo per alimentare e raffreddare i sistemi è quadruplicato. Per dare una stima tangibile dei costi affrontati a tutt'oggi, il consumo medio per un'azienda di servizi di medie dimensioni (200 dipendenti) è di circa 1400KWh (210,00 €), mentre un'abitazione privata in un anno consuma in media 730KWh [20]. Oggi il costo di energia e raffreddamento rappresenta circa il 60% della spesa in nuove infrastrutture, con un impatto più che significativo sul Total Cost of Ownership (si veda l'illustrazione 4): per 1,00 € speso per acquistare nuovi server, si spendono 0,60 € all'anno per alimentarli e raffreddarli.
Dall'illustrazione 5 si evince come la spesa energetica sia una parte consistente delle spese totali di un datacenter. Tale impatto è destinato a crescere ulteriormente come conseguenza del continuo aumento del costo unitario dell'energia. Se si tiene conto che le statistiche IDC da cui sono tratti i dati sono svolte a livello mondiale, dove il costo dell'energia è mediamente molto più basso che in Italia (a titolo di esempio, come si evince dalla tabella 2, il costo dell'energia per utenti industriali è di 6 $cent/kWh in USA e di 24 $cent/kWh in Italia), è facile rendersi conto di come nel nostro paese il peso del costo energetico possa essere ancora più significativo. Questi costi sono spesso nascosti e ignorati, in quanto da un lato mancano ancora gli strumenti e le metodologie per misurarli con esattezza,

Nazione	$cent / KWh
USA	6
Francia	5
Spagna	9
UK	13
Germania	8
Italia	24
Cina	6
Giappone	12
Sud Africa	2
Australia	6

Tabella 2: Costo medio dell'energia elettrica nelle nazioni più importanti

dall'altro molto spesso non vengono contabilizzati nel budget dei sistemi IT, ma vengono annegati nei consumi elettrici di tutta l'azienda, rendendo quindi difficile una chiara percezione del fenomeno.

L'intensa richiesta energetica per usare e raffreddare datacenter ora vale circa un quarto delle emissioni totali di CO2 di tutto il settore dell'ICT [21]; è stimato che un server di medie dimensioni emette indirettamente all'incirca la stessa quantità di CO2 prodotta da un SUV che ha consumi pari a 6,5 km/l [22]. La potenza richiesta per alimentare un rack di server blade ad alta densità può essere fino a 10-15 volte maggiore di un server tradizionale [23]: tale richiesta può arrivare fino a 40KW per rack (ovvero 5 server da 8 unità).

Un recente studio [24] negli USA ha scoperto che, nel 2006, 1.5% della totalità dell'energia nazionale richiesta proveniva dall'energia consumata dai datacenter; questa ricerca ha anche rilevato che il consumo di energia elettrica nei datacenter degli USA è raddoppiato negli ultimi cinque anni e che ci si aspetta che raddoppi nuovamente nei prossimi cinque anni fino a raggiungere un costo annuale di circa $7.4 Miliardi di dollari, poiché si stima che i datacenter dovranno ospitare il 50% dei server in più. Ma non è solo l'energia richiesta per far funzionare i server dei datacenter che contribuisce all'emissione di CO2: una simile quantità è richiesta per rimuovere il calore generato dai server, usando intense unità di condizionamento [23].

A questo proposito è opportuno notare che l'efficienza energetica dell'IT - ossia le prestazioni rapportate al consumo di energia - è di fatto cresciuta negli ultimi anni, dato che le prestazioni sono migliorate più di quanto non sia cresciuta la potenza richiesta. Per esempio, se si considera il benchmark TPC-C [25], comunemente utilizzato per valutare le prestazioni dei processori, l'efficienza energetica può essere misurata in migliaia di transazioni al minuto per Watt assorbito (Ktpm-

c/Watt). Il valore di tale indice nell'ultimo decennio è aumentato di un fattore 2,5, indicando quindi un miglioramento dell'efficienza energetica. Tuttavia la crescente domanda di capacità di calcolo e l'aumento del consumo energetico dell'IT in termini assoluti impongono di migliorarne ulteriormente e in modo più radicale l'efficienza energetica. D'altro canto la legge di Moore - per cui il numero di componenti per microchip raddoppia ogni 18-24 mesi non corrisponde alla legge economica del settore ICT: a fronte di una crescita esponenziale di prestazioni per chip abbiamo una crescita doppia delle prestazioni rispetto al costo [26, 27]. Sembra un paradosso, ma il costo dell'hardware decresce più in fretta di quanto aumenti la sua miniaturizzazione. Questo fa aumentare la domanda di servizi ICT e il risultato è che gli enormi miglioramenti nell'efficienza energetica non tengono testa all'aumento della richiesta di uso di computer, internet e cellulari; dunque il saldo netto dei consumi energetici del settore ICT è negativo: la domanda totale di energia dell'hardware installato è in crescita [25].

Per sensibilizzare i propri cittadini, alcune nazioni hanno creato una tassa sulla quantità di CO2 prodotta dalla richiesta energetica: prima fra tutte la Svezia, ha imposto una carbon tax di 0.25 SEK/kg ($100 per tonnellata) sull'uso di petrolio, carbone, gas naturale, ed altri combustibili inquinanti. Gli utenti industriali pagano la metà del tasso (tra il 1993 ed il 1997 il 25% del tasso). Nel 1997 il tasso fu incrementato a 0.365 SEK/kg ($150 per tonnellata) di CO2 emesso [28] . Finlandia, Paesi Bassi, e Norvegia introdussero anch'esse carbon taxes negli anni '90; in Italia la carbon tax è stata introdotta con l'art. 8 della legge n. 448 del 23 dicembre 1998, [28,29] secondo le conclusioni della Conferenza di Kyoto svoltasi dall'1 all'11 dicembre 1997.

L'energia elettrica come limite di scalabilità

Tra il 2000 e il 2006 l'energia consumata da prodotti IT non domestici è cresciuta di più del 70% [30] e ci si aspetta che cresca ancora del 40% per il 2020: l'alto

consumo di energia delle apparecchiature informatiche sta diventando un limite alla scalabilità dei data center di medie e grandi imprese dislocati in aree ad alta densità abitativa. La potenza elettrica richiesta sta crescendo dell'8-10% all'anno e i gestori della rete elettrica rischiano di non essere più in grado di convogliare così tanta energia in un'area ristretta di un centro urbano: in alcuni casi la densità di energia assorbita dai data center ha superato i 20 kW per metro quadro. La potenza assorbita per metro quadro dai nuovi server ad alta densità (blade) è spesso incompatibile con le caratteristiche elettriche degli attuali data center. Poiché le infrastrutture della rete elettrica sono difficilmente modificabili in aree urbane, per aumentare la capacità di calcolo degli attuali data center potrebbe quindi essere necessario edificare nuove strutture in aree a più bassa densità abitativa, con ulteriore impatto ambientale di costo. Secondo Forrester Research nei prossimi anni il 60% dei data center saranno limitati dal consumo di energia, dalle esigenze di raffreddamento e dallo spazio; inoltre, secondo studi di settore, nel breve futuro, molti datacenter non saranno in grado di avere l’energia necessaria al loro funzionamento.

3. Cloud Computing

Non esiste in letteratura una definizione univoca di Cloud Computing. Cercando di toccare tutti i punti presentati nelle varie definizioni, si presenta questa: il Cloud Computing è un paradigma di calcolo distribuito su larga scala, dove risorse condivise, software e informazioni sono gestiti ed erogati on-demand ai client esterni, quali computer e ad altri device, attraverso la rete Internet. Queste risorse, quali la potenza di calcolo, lo storage, sono astratti, virtualizzati e dinamicamente scalabili. Il termine «cloud» quindi è usato come una metafora per Internet: infatti in passato era usato un disegno di nuvola per rappresentare la rete telefonica, successivamente tale disegno è stato usato per rappresentare Internet nei diagrammi di reti di computer come un'astrazione delle infrastrutture sottostanti che rappresenta. In altre parole, il Cloud Computing è uno stile di computazione in cui il software è fornito come un servizio, consentendo all'utente di accedervi senza necessità di specifico know-how e soprattutto senza la necessità di avere un controllo diretto sulle infrastrutture di supporto.

Negli ultimi anni abbiamo assistito a molti sforzi volti a trasformare il calcolo in qualcosa di simile a un servizio di pubblica utilità (come quelli di energia elettrica e gas); già dall'inizio degli anni 80 abbiamo visto lo spostamento del calcolo dai mainframe a un paradigma di tipo client-server: i dettagli sono astratti agli utenti che non hanno più bisogno di essere esperti o di avere il controllo su infrastrutture tecnologiche che ora sono «in the cloud». Gli sviluppatori con idee innovative per nuovi servizi Internet non devono più investire grandi somme di denaro per comperare hardware per sviluppare il proprio progetto o per pagare operatori che lavorino sull'hardware stesso; non devono più preoccuparsi di predire esattamente la popolarità di un servizio che potrebbe essere sopravvalutato facendo quindi

perdere denaro e risorse, oppure sottovalutare un servizio che diventa velocemente popolare, mancando quindi dei potenziali clienti e dei potenziali guadagni; in più, compagnie con task di tipo batch possono ottenere velocemente risultati poiché i loro programmi possono scalare; infatti usare 1000 server per un'ora non costa di più che usarne uno per 1000 ore: questa elasticità di risorse, senza pagare un premio per la larga scala, è una novità nella storia dell'IT.

I primi significativi progressi sono stati fatti dal grid computing, che ha compiuto importanti passi avanti nell'area del High Performance Scientific Computing, nel tentativo di costruire utility di livello enterprise. Comunque, nessuno di questi tentativi si è materializzato in una utility general purpose di calcolo, accessibile da chiunque, in qualsiasi momento e da qualsiasi luogo. Quello che rende il cloud computing diverso può essere identificato nel fatto che trend come la vasta adozione di reti broadband, la veloce penetrazione di tecnologie di virtualizzazione per server x86, e l'adozione di Software as a Service, hanno finalmente creato l'opportunità e la necessità di una computing utility globale. La riluttanza a usare servizi online in sostituzione dei tradizionali software sta diminuendo: il successo di compagnie come salesforce.com prova che con il giusto insieme di garanzie di sicurezza e prezzi competitivi, le compagnie vorranno affidare anche i loro dati maggiormente di valore (le relazioni con i clienti) a un fornitore di servizi on-line. Allo stesso tempo, le tecnologie di virtualizzazione hanno reso possibile separare le funzionalità di un sistema eseguite dallo stack software (OS, middleware, application, data) dalle risorse fisiche computazionali che le eseguono. Il Cloud Computing descrive quindi un nuovo ulteriore modello per il consumo e la distribuzione di servizi basati su Internet e tipicamente involve la fornitura di risorse scalabili dinamicamente e spesso virtualizzate come un servizio su Internet; può essere visto come sottoprodotto di Internet con la conseguenza di essere di facile accesso attraverso il remote computing. Questo permette un nuovo modello di calcolo on-line:

invece di un software on-line costruito apposta, ora si può pensare in termini di macchine virtuali general purpose che possono fare qualsiasi cosa. Il termine Cloud Computing si riferisce sia alle applicazioni fornite come servizi su Internet che all'hardware e ai sistemi software nei data center che si occupano di quei servizi [31]; questi ultimi prendono il nome di Software as a Service (SaaS); l'hardware e il software di un data center è ciò che viene chiamata una cloud. Una cloud pubblica è quella che permette l'accesso verso chiunque (tipicamente reso disponibile attraverso forme di pagamento di tipo pay-as-you-go), e il servizio venduto è definito come utility computing; si usa il termine cloud privata per riferirsi a un datacenter interno a un'azienda o un'organizzazione, non pubblico all'esterno. Cloud Computing quindi è la somma di SaaS e Utility Computing, ma non include le cloud private.

In figura 6 vengono mostrati i ruoli degli utenti e dei provider nel cloud computing secondo questa visione. I vantaggi nell'utilizzo di Software as a Service per i service provider risiedono nella semplificazione dell'installazione del software, nella gestione e nel controllo centralizzato; gli utenti possono accedere al servizio in qualunque momento e da qualunque luogo, condividere dati e collaborare più facilmente. Il Cloud Computing non cambia questi vantaggi, ma permette a più application provider la possibilità di fornire i propri servizi come SaaS senza dover costruire un data center: come l'avvento di fonderie nel campo dei semiconduttori ha dato alle compagnie che progettavano chip la possibilità di venderli senza possedere un impianto di costruzione, il cloud computing permette il deployment di software as a service (e il relativo scaling

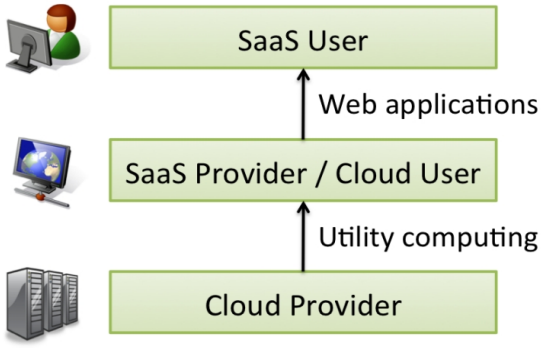


Illustrazione 6

on-demand) senza costruire o fornire un data center. Analogamente a come SaaS permette agli utenti di scaricare alcuni problemi al service provider, quest'ultimo, con il cloud computing, può scaricare alcuni dei suoi problemi sul cloud provider. I service provider gradiscono molto la semplificazione nell'installazione, nel mantenimento e in un controllo centralizzato del software, mentre gli utenti finali possono accedere al servizio «anytime, anywhere» (in ogni momento, in ogni luogo), condividere dati e collaborare molto più agevolmente e mantenere i propri dati in modo sicuro all'interno dell'infrastruttura. Analogamente a come il paradigma di SaaS permette agli utenti di scaricare problemi al SaaS provider, il SaaS provider ora può scaricare i suoi problemi al Cloud Computing provider.

Il cloud computing porta con se tre aspetti nuovi:

- L'illusione di infinite risorse di calcolo disponibili on-demand, eliminando quindi la necessità per gli utenti della cloud di pianificare sulle necessità di calcolo;

- L'eliminazione di un impegno a priori da parte degli utenti della cloud, permettendo alle compagnie di iniziare con poche risorse hardware e poi incrementarle quando vi è un cambiamento delle proprie necessità;

- La possibilità di pagare per l'utilizzo delle risorse di calcolo su periodi brevi (processori per ora o spazio storage al mese).

Nei passati tentativi falliti di Utility Computing mancava almeno uno di questi tre nuovi aspetti. Oggi creare un datacenter per il Cloud Computing può costare anche diverse centinaia di milioni di dollari, poiché include la struttura, la fornitura di hardware e la gestione di quest'ultimo; comunque, a causa della crescita esponenziale di servizi Web duranti i primi anni del ventunesimo secolo, diverse grandi compagnie Internet, quali Amazon, eBay, Google, Microsoft e altri,

si erano già diretti verso questa direzione; ugualmente importante queste aziende stavano già sviluppando infrastrutture per software scalabile (come per esempio MapReduce, Google FileSystem, BigTable e Dynamo [32,33,34,35]) e operatori esperti per gestire i propri datacenter contro potenziali attacchi fisici e elettronici.

Quindi, una condizione necessaria ma non sufficiente per diventare un provider di Cloud Computing è che deve avere investimenti già attivi non solo in datacenter, ma anche in infrastrutture software di larga scala e operatori esperti richiesti per tali software; date queste condizioni, diversi fattori possono influenzare queste compagnie per diventare provider di Cloud Computing

- Guadagnare molti soldi: benché 10 $cent per ora-server sembrino pochi, la tabella 3 riassume le stime [36] di grandi datacenter (decine di centinaia di server) che possono comprare hardware, bandwidth e elettricità a 1/5 o 1/7 del prezzo offerto a datacenter di media grandezza (centinaia o migliaia di server); perciò, il costo fisso dello sviluppo di software può essere ammortizzato su molte macchine.

Tecnologia	**Costo in un DataCenter medio**	**Costo in un DataCenter grande**	**Risparmio**
Network	$95 per Mbit/sec/month	$13 per Mbit/sec/month	7,1
Storage	$2.20 per GByte / month	$0.40 per GByte / month	5,7
Administration	140 Servers / Administrator	1000 Servers / Administrator	> 7,1

Tabella 3: Economie di scala nel 2006 per DataCenter medi (1000 server) e DataCenter grandi (50000 server)

- Far leva su investimenti già attivati: aggiungere servizi di Cloud Computing sulla base di già esistenti infrastrutture genera un nuovo flusso di guadagno con (idealmente) una spesa aggiunta molto bassa, aiutando ad ammortizzare i grandi investimenti per i datacenter. Per esempio, molti

servizi Amazon furono inizialmente sviluppati per operazioni interne ad Amazon stessa.

- Difendere una propria esclusività: poichè server convenzionali e imprese possono abbracciare il Cloud Computing, diversi produttori di software con una stabile clientela in queste applicazioni possono essere motivati a fornire una propria possibilità di Cloud Computing ai loro clienti. Per esempio, Microsoft Azure fornisce un'immediata modalità per migrare i propri clienti che usano soluzioni enterprise del proprio software verso una soluzione di tipo cloud.

- Prevenire la nascita di nuove aziende concorrenti: una compagnia con i requisiti adatti per essere un provider di soluzioni di Cloud Computing vorrebbe stabilire un monopolio per evitare la nascita di aziende concorrenti.

- Influenzare le relazioni coi clienti: organizzazioni di servizi IT come IBM Global Service hanno estese relazioni coi clienti per la loro offerta di servizi. Offrire una soluzione Cloud Computing propria da ai loro clienti la possibilità di continuare a essere seguiti da IBM, con un duplice vantaggi sia per i clienti che per l'azienda stessa.

- Diventare una piattaforma: l'iniziativa di facebook di abilitare applicazioni plug-in è molto adatta per una soluzione di tipo Cloud Computing: infatti il provider per l'infrastruttura di Facebook è Joyent, un Cloud Computing provider. A tutt'oggi la motivazione di Facebook è quella di rendere il loro social-network una piattaforma per lo sviluppo di nuove applicazioni.

Va aggiunto che le offerte commerciali generalmente vanno incontro all'esigenza di Quality of Service (QoS) da parte dei clienti e tipicamente offrono contratti di

tipo SLA. Come menzionato nei punti precedenti, i maggiori provider di servizi di Cloud Computing sono Goolge, IBM, Microsoft, HP, Amazon e VMware.

Classi di Utility Computing

Un'ulteriore diversità che si trova spesso in letteratura è sulla classificazione dei servizi di utility computing, spesso attraverso termini come Infrastructure as a Service (IaaS) o Platform as a Service (PaaS). Di seguito presentiamo brevemente le differenze principali delle classi di utility computing. Ogni applicazione ha bisogno di un modello computazionale, un modello di storage e un modello di comunicazione. Lo statistical multiplexing necessario per ottenere elasticità e l'illusione di una capacità infinita richiedono che le risorse siano virtualizzate, in modo tale che l'implementazione su come sono condivise può essere nascosta al programmatore. Diverse offerte di utility computing possono essere distinte in base al livello di astrazione presentato ai programmatori e il livello di astrazione della gestione delle risorse. Amazon EC2 rappresenta uno degli estremi: un'istanza EC2 assomiglia molto ad hardware fisico e gli utenti possono controllare praticamente l'intero stack software, dal kernel in su. La API esposta è composta da poche dozzine di chiamate, per richiedere e configurare l'hardware virtualizzato. Non ci sono limiti a priori sul tipo di applicazioni che possono essere ospitate. Questa tipologia di cloud computing, in letteratura, viene spesso denominata Infrastructure as a Service (IaaS). Altro estremo è rappresentato dalle piattaforme per applicazioni dal dominio ben specificato come Google AppEngine e Force.com, la piattaforma di sviluppo business di SalesForce. AppEngine è indirizzata esclusivamente alle tradizionali web application, dove è presente una chiara separazione tra uno strato di calcolo stateless e uno strato di storage stateful. Inoltre le applicazione AppEngine devono essere di tipo request-reply e il tempo di CPU necessario per servire una particolare richiesta è rigorosamente razionato. In questo modo AppEngine non è utilizzabile per calcolo general-

purpose. In modo analogo, Force.com è progettato solo per applicazioni business che sono eseguite sul database salesforce.com.

Microsoft Azure rappresenta il punto intermedio. Le applicazioni di Azure sono scritte usando le librerie .NET e il sistema permette calcolo general-purpose, piuttosto che una singola categoria di applicazioni. Gli utenti possono scegliere un linguaggio di programmazione, ma non possono controllare il sistema operativo sottostante o l'ambiente di esecuzione. Questa tipologia di cloud computing prende spesso il nome di Platform as a Service (PaaS).

I maggiori ostacoli al Cloud Computing

In [2] e in tabella 4 vengono elencati dieci ostacoli alla crescita del Cloud Computing. Ogni ostacolo viene collegato ad una soluzione (o ad una opportunità di ricerca) la quale, secondo gli autori, rappresenta il metodo per superare l'ostacolo e che può variare dallo sviluppo di semplici prodotti ai maggiori progetti di ricerca. La tabella sottostante riassume i dieci ostacoli e le relative proposte di soluzione. I primi tre sono ostacoli tecnici all'adozione del Cloud Computing, i successivi cinque sono ostacoli alla crescita del Cloud Computing e gli ultimi due sono ostacoli commerciali alla adozione del Cloud Computing. Di seguito analizzeremo alcuni di questi ostacoli, in particolare quelli maggiormente legati agli scopi di questa tesi.

- Availability of Service: la disponibilità del servizio (availability of service) è un punto cruciale per qualsiasi sistema distribuito e in particolare per le imprese. Recenti episodi [31] hanno dimostrato come, nonostante l'utilizzo di tecniche atte a diminuire l'incidenza dei guasti sulla disponibilità del servizio (es. sistemi di replicazione), un singolo cloud provider può rappresentare in un certo senso un "single point of failure". Come i grandi Internet service provider utilizzano più network provider cosicché il guasto

	Ostacolo	**Opportunità/Soluzione**
1	Availability of Service	Utilizzare più cloud provider per fornire continuità di servizio; scaling per difendersi da attacchi DDOS
2	Data Lock-In	API standard
3	Data Confidentiality and Auditability	Crittografia, VLAN e firewall; conciliare leggi nazionali con storage geografico
4	Data Transfer Bottleneck	Spedizione dischi, archivio dati; Abbattimento costi router WAN; switch LAN con bandwidth maggiore
5	Performance Unpredictability	Migliore gestione e supporto delle macchine virtuali;
6	Scalable Storage	Inventare Scalable storage
7	Bugs in Large-Scale Distributed Systems	Debugger basato su macchine virtuali
8	Scaling Quickly	Creare un sistema per scalare (sia in su che in giù) velocemente le applicazioni
9	Reputation Fate Sharing	Creare servizi di controllo sul comportamento degli utenti
10	Software Licensing	Licenze pay-for-use

Tabella 4

di uno di essi non pregiudichi il servizio, così la possibile soluzione alla ricerca di disponibilità dei servizi molto elevata risiede nell'utilizzo di più cloud provider. La filosofia alla base della high-availability è quella di evitare single point of failure. Anche se un cloud provider è dotato di diversi data center in varie regioni geograficamente distinte che utilizzano network provider diversi, può avere infrastrutture software e sistemi di contabilità comuni, oppure decidere un giorno di terminare il servizio. I clienti potrebbero essere riluttanti a migrare verso il Cloud Computing senza una strategia di continuità per queste situazioni. Quindi si ritiene che una possibile soluzione sia quella di essere forniti da diverse compagnie di utility computing.

- Performance unpredictability: la condivisione da parte di più macchine virtuali delle stesse risorse (in particolare CPU e I/O) può portare ad alcune problematiche sulle reali performance che si possono ottenere, e in particolare sulla loro variazione e non predicibilità nel tempo. Una soluzione a questo problema risiede nella ricerca e nel miglioramento delle architetture e dei sistemi operativi per gestire interrupt e canali di I/O in maniera più efficiente.

- Scaling quickly: il modello pay-as-you-go si applica bene a storage e larghezza di banda, perché in entrambi i casi si contano i byte. Il calcolo è leggermente differente, a seconda del livello di virtualizzazione. Google AppEngine scala automaticamente in risposta all'incremento o decremento del carico e agli utenti sono addebitati dei cicli di calcolo utilizzati. AWS addebita le ore per il numero di istanze richieste, anche se queste rimangono inutilizzate. Per ovviare servono meccanismi per aggiungere e togliere rapidamente risorse in risposta al carico al fine di risparmiare denaro, ma senza violare il contratto SLA.

QoS e SLA

L'interesse per applicazioni Internet è in costante crescita. Alcuni servizi esistenti e emergenti richiedono elevati livelli di qualità del servizio (quality of service o QoS) e hanno elevate esigenze di risorse (si pensa ad applicazioni real-time come videoconferenze). In letteratura la Qualità del Servizio (QoS) non trova una definizione univoca bensì differenti definizioni a seconda del campo a cui è applicata. Troviamo una definizione di qualità del servizio come "l'effetto collettivo delle performance di un servizio che determinano il grado di soddisfazione di un utente del servizio" [37]. Quando un'azienda o organizzazione si affida a servizi forniti da un altro ente o impresa (per esempio servizi web) per l'implementazione dei propri processi business, spesso vengono richieste garanzie

contrattuali sulla qualità del servizio, come allo stesso modo i fornitori del servizio richiedono garanzie affinché i clienti non abusino del servizio. Queste qualità e i vincoli di utilizzo sono spesso definiti in accordi bilaterali chiamati Service Level Agreement (SLA), che specificano la qualità del servizio richiesta e le penalità associate alle violazioni. Queste penalità sono tradotte con pagamenti pecuniari (o rimborso per il costo del servizio) e possono essere viste come un'assicurazione contro la fornitura di un servizio scadente e l'eccessivo utilizzo. Le applicazioni enterprise, a differenza delle normali applicazioni, hanno stringenti esigenze di qualità del servizio. Con ciò si intende una serie di caratteristiche, tipicamente requisiti non funzionali come disponibilità, scalabilità, affidabilità e tempistiche di risposta, che l'applicazione deve rispettare e che solitamente sono espresse attraverso l'uso di contratti SLA, che legano un fornitore di servizi ad un cliente che ne fa uso. I contratti SLA possono essere di diversi tipi e con diversi scopi. Attualmente la maggior parte dei cloud provider (pubblici) offre SLA limitati alla disponibilità del servizio e/o alle caratteristiche delle risorse offerte (es. numero e tipo di CPU [38]).

Testare la qualità dei servizi web utilizzando, per esempio, test su performance e affidabilità è necessario, ma non sufficiente. La qualità del servizio dipende fondamentalmente dalla fornitura di risorse computazionali che il service provider (o chi per esso) gestisce durante la vita del servizio. Per vigilare su una SLA, è necessario per l'utente del servizio monitorare costantemente, o almeno su intervalli statisticamente significativi, la qualità del servizio fornita a tempo di esecuzione. Inoltre il service provider dovrà anch'esso monitorare a run-time la qualità del servizio per controllare che l'utilizzo del servizio non ecceda i livelli concordati nella SLA, per proteggersi contro false affermazioni sulla scarsa qualità del servizio, ma soprattutto per determinare se aumentare le quantità di risorse utilizzate in caso la qualità del servizio si abbassa al di sotto di determinate soglie [39].

Molte applicazioni distribuite di classe enterprise vengono sviluppate per essere eseguite su piattaforme di Application Server, come J2EE, CORBA o .NET. Queste applicazioni possono richiedere requisiti di Qualità del Servizio (QoS), come scalabilità e disponibilità del servizio, attraverso un contratto di Service Level Agreement (SLA). Le SLA sono contratti che stabiliscono le garanzie di QoS che un ambiente di esecuzione deve fornire per le applicazioni che ospita. Per assicurare che la SLA di una applicazione non sia violata, una delle politiche adottabili è quella chiamata resource over-provisioning: vengono allocate staticamente un numero sufficiente di risorse per supportare un carico di lavoro in qualsiasi scenario, anche nel caso peggiore. Con questa politica però, una percentuale elevata di risorse può rimanere inutilizzata per gran parte del tempo. Al contrario una politica di utilizzo ottimale delle risorse può essere ottenuta fornendo a ogni applicazione ospitata il numero minimo di risorse richieste per onorare la SLA e continuando a gestire l'allocazione delle risorse anche a tempo di esecuzione.

I contratti di Service Level Agreement (SLA) sono l'attuale pratica del mondo economico per specificare requisiti di qualità del servizio nell'ambito IT. La SLA in rappresenta una collezione di clausole contrattuali che legano un QoS-aware cluster alle applicazioni che ospita. Questo particolare tipo di SLA (scritta utilizzando il linguaggio SLAng [40]) prende il nome di hosting SLA. Questa può comprendere due macro aree: diritti e obblighi dei client (Client Responsabilities) e diritti e obblighi della parte server (Server Responsabilities). Tra gli obblighi dei client troviamo il numero massimo di richieste che i client possono inviare all'applicazione entro un certo intervallo di tempo.

Il frammento sopra riportato mostra l'attributo requestRate che serve appunto per esprimere il numero massimo di richieste (100) che è possibile inviare alla applicazione in un secondo. Gli obblighi dell'applicazione possono includere

```
<ContainerServiceUsage name="HighPrority" requestRate="100/s">
            <Operations>
                <Operation path="catalog.jsp" />
                <Operation path="AddToCart" />
                <Operation path="checkout.jsp" />
                <Operation path="CheckoutCtl" />
            </Operations>
            ...
        </ContainerServiceUsage>
```

Testo 1: Frammento di SLA

garanzie su disponibilità del servizio, latenza delle risposte e percentuale delle violazioni sui termini della SLA che possono essere tollerate. Nel frammento sottostante si può vedere l'attributo serviceAvailability che specifica la probabilità che l'applicazione ospitata sia disponibile (ovvero la probabilità che l'applicazione risponda in tempo predicibile) su un determinato periodo: nell'esempio il valore indica che l'applicazione deve essere disponibile giornalmente per non meno del 99%.

Un ulteriore parametro che si può notare negli obblighi della parte server è il maxResponseTime che rappresenta il tempo massimo che deve trascorrere tra la ricezione di richiesta (per una delle operazioni specificate all'interno del tag Operations) e il completamento dell'invio della relativa risposta. Infine per specificare la percentuale di violazioni della SLA che possono essere tollerate prima che un application service provider incorra in penalità, sono stati utilizzati i parametri efficiency e efficiencyValidity. Nell'esempio questi attributi specificano che non meno del 95% delle richieste, in un intervallo di due ore, deve essere soddisfatto in conformità ai requisiti specificati nella parte di Server Responsabilities.

Architettura della Cloud Facility

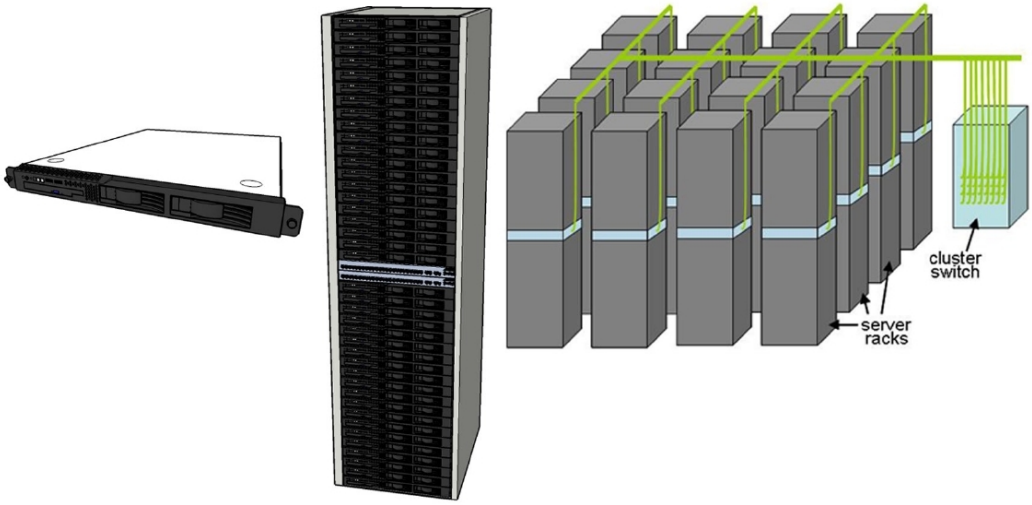


Illustrazione 7: Elementi tipici in un datacenter di larga scala: 1U server (sulla sinistra), rack di 7 piedi con Ethernet switch (nel mezzo) e un diagramma di piccoli cluster con un Ethernet switch/router al livello superiore (destra).

Le implementazioni hardware di soluzioni di Cloud Computing possono essere differenti da un'installazione ad un'altra: pure all'interno di una singola organizzazione come Google, sistemi sviluppati in anni diversi usano elementi di base differenti, riflettendo i miglioramenti hardware forniti dall'industria. Comunque, l'organizzazione architetturale di sistemi è stata relativamente stabile attraverso questi ultimi anni; la figura 7 illustra una delle configurazioni più popolari all'interno di un datacenter: un insieme di server, tipicamente in formato 1U o blade, sono montati in rack e interconnessi attraverso uno switch Ethernet locale; questi switch a livello rack possono usare collegamenti a 1 o 10 Gbps e hanno un numero di link attivi a uno o più Ethernet switch a livello cluster (o a livello datacenter); questo secondo livello di switching può potenzialmente coprire più di diecimila server individuali.

Storage

Generalmente, i disk drives sono connessi direttamente a ogni server individuale e gestiti da un file system distribuito (quale può essere Google GFS [41]) oppure possono essere parte di un Network Attached Storage (NAS) che sono connessi direttamente agli switch di livello cluster. Una soluzione NAS tende ad essere più semplice da sviluppare inizialmente poiché porta tutta la responsabilità per il data management e l'integrità dei dati al fornitore del NAS; dall'altra parte invece, una collezione di dischi direttamente attaccati a nodi server richiede un filesystem fault-tolerant a livello cluster: questo può essere difficile da implementare, ma può abbassare notevolmente i costi hardware e l'utilizzo di apparecchiatura network.

Apparecchiature Network

La scelta di apparecchiature network per un datacenter che ospita Cloud Computing involve un trade-off tra velocità, scalabilità e costi. Al tempo della scrittura di questa tesi, uno switch Ethernet da 1-Gbps con 48 porte, usato a livello rack, costa meno di $30/Gbps; dall'altra parte, uno switch Ethernet da 10-Gbps con un numero più elevato di porte, le quali sono necessarie per tenere assieme i cluster del datacenter, hanno un prezzo strutturale notevolmente più alto e costano dieci volte di più: in altre parole, uno switch che ha 10 volte la bandwidth costa circa 100 volte di più. Il risultato di questa discontinuità di costo è l'organizzazione del datacenter in una gerarchia a due livelli, come mostrato nell'illustrazione 7.

Power Usage

L'energia e la potenza richiesta sono importanti nel design di un datacenter che andrà ad ospitare Cloud Computing poichè i costi dell'energia elettrica sono diventati, come detto precedentemente, una parte fondamentale del TCO di questa classe di sistemi. L'illustrazione8 da una visione d'insieme di come l'energia sia

distribuita all'interno dei sistemi moderni. Benché questi dati possono variare sensibilmente rispetto a come si intende configurare il sistema in un certo dominio, questo grafo indica che non ci si può più focalizzare solo sull'efficienza energetica delle CPU poiché non sono esse che dominano l'intero profilo energetico.

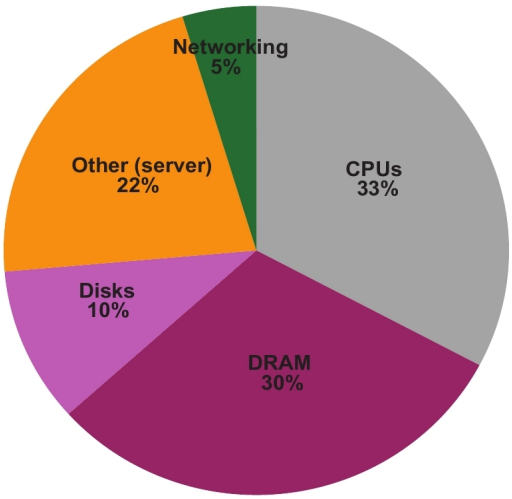


Illustrazione 8: Distribuzione approssimativa di picchi di potenza richiesta dall'hardware in uno dei datacenter Google (2007)

Classificazione dei datacenter

Il design complessivo di un datacenter è spesso classificato come appartenente a «Tier I-IV»:

- Tier I: datacenter che hanno una singola linea di potenza e di raffreddamento, senza alcun componente ridondante;
- Tier II: datacenter che hanno una componente di ridondanza aggiunta (N+1), aumentando così la stabilità;
- Tier III: datacenter che hanno multiple linee di potenza e di raffreddamento ma solo una attiva. Hanno anche componenti ridondanti e si può applicare manutenzione in modo concorrente, cioè hanno un sistema di ridondanza durante anche il ciclo di manutenzione, solitamente con un setup di N+2;

- Tier IV: datacenter che hanno due linee attive di energia e raffreddamento, componenti ridondanti su ogni linea e sono studiati per tollerare ogni singola failure dell'equipaggiamento senza avere impatti finali sul carico di lavoro.

Queste classificazioni tuttavia non sono accurate al 100%: alcuni datcenter commerciali ricadono fra Tier III e IV. Nel mondo reale la stabilità di un datacenter è fortemente influenzata dalla qualità dell'organizzazione che regola il datacenter stesso e non solamente dal design di quest'ultimo. Una disponibilità tipica stimata è di 99.7% per Tier II e da 99.98 a 99.995% rispettivamente per i Tier III e IV.

La grandezza di un datacenter può variare molto: due terzi dei server presenti negli USA sono presenti in datacenter più piccoli di 450 m² con meno di 1MW di potenza critica [42]. Datacenter più grandi sono costruiti per hostare server da compagnie differenti e possono supportare un carico massimo di 10-20MW; veramente pochi datacenter possono eccedere i 30MW di carico.

Il sistema energetico nei datacenter

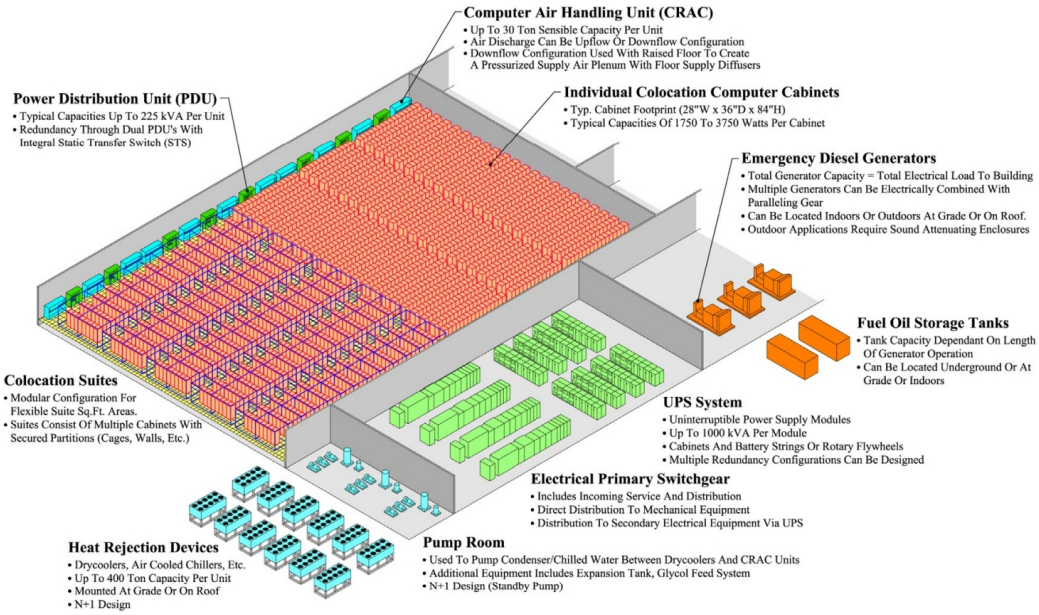


Illustrazione 9

L'illustrazione 9 mostra le componenti di un tipico datacenter. La potenza entra nell'edificio da un trasformatore esterno; questa parte di potenza è spesso chiamata «voltaggio medio» (tipicamente 10-20 kV) per distinguerla da quella di «alto voltaggio» delle linee di lunghe distanze e dal «basso voltaggio» per la distribuzione interna (110-600V). Le linee di voltaggio medio terminano al primo commutatore, che include interruttori automatici per proteggere il datacenter contro problemi elettrici e trasformatori per scalare il voltaggio fino a 400-600V. L'elettricità a basso voltaggio quindi fluisce fino ai gruppi di continuità (UPS — Uninterruptible Power Supply), i quali hanno in ingresso una seconda linea (di

uguale voltaggio) da un insieme di generatori diesel che entrano in azione nel caso che la linea primaria abbia dei problemi.

Sistemi UPS

Un sistema UPS tipicamente combina tre funzioni in un solo dispositivo:

- Primo, vi è all'interno uno switch che sceglie da quale linea attingere la potenza in ingresso (o l'elettricità del sistema elettrico o quella dei generatori). Se vi è stata una carenza di energia dalla linea principale, lo switch attiva la linea dei generatori quando sente che sono stati avviati; mediamente un generatore necessita di 10-15 secondi per avviarsi e raggiungere la potenza richiesta.

- Secondo, l'UPS contiene delle batterie o volani per far da ponte fra il tempo senza la linea principale e l'avvio dei generatori. Un tipico UPS soddisfa questa richiesta attraverso una conversione elettrica AC-DC-AC (alternata-continua-alternata): cioè prende in ingresso corrente alternata, la trasforma in corrente continua per alimentare le batterie in serie all'interno dell'UPS stesso; in seguito, la corrente DC in uscita dalla linea di batterie viene riconvertita in alternata per alimentare l'equipaggiamento dei datacenter. Quindi, quando la linea elettrica principale si guasta, l'UPS perde il suo input ma ha internamente potenza derivata dalle batterie, e quindi continua a inviare corrente alternata dopo il secondo step di conversione; in seguito, il generatore si avvia ridando all'UPS la linea elettrica in ingresso.

- Terzo, gli UPS condizionano la potenza in ingresso, rimuovono picchi o mancanze di voltaggio, oppure distorsione di armoniche nella corrente.

Questo condizionamento è naturalmente compiuto attraverso la doppia conversione AC-DC-AC.

Poiché gli UPS necessitano di diverso spazio, sono tipicamente messi in una stanza separata e non nel piano del datacenter. Un tipico UPS ha un range che varia dalle centinaia di KW ai 2MW.

Power Distribution Units

La corrente elettrica in uscita dagli UPS è quindi l'ingresso delle unità di distribuzione della potenza (PDU — Power Distribution Unit) che sono presenti nel piano del datacenter. I PDU assomigliano ai quadri elettrici presenti nelle abitazioni private: prendono in ingresso una grossa linea ad alto voltaggio (tipicamente 200-480V) e la spezzano in diverse linee da 110-220 V che alimentano i server sul piano. Ogni circuito è protetto da un suo interruttore automatico così che un cortocircuito in un server o un alimentatore possa dare problemi solo a quel circuito stesso, e non all'intero PDU e addirittura agli UPS. Un tipico PDU lavora con carichi dell'ordine dei 75-225 kW, poichè un singolo circuito può gestire dai 20 ai 30A a 110-220V, ovvero un massimo di 6kW. I PDU spesso provvedono una ridondanza addizionale accettando due linee di potenza indipendenti in ingresso (tipicamente chiamate «A side» e «B side») e sono capaci di fare switch tra di loro con un delay davvero minimale, in modo che la perdita di tensione non influenzi minimamente la potenza dei server. In un tale scenario, gli UPS sono duplicati nella parte «A side» e «B side», così pure nel caso di fallimento di una linea di UPS, la corrente non verrà interrotta ai server.

I datacenter del mondo reale contengono diverse varianti della versione semplificata mostrata in questa tesi: tipiche varianti includono il «parallelismo» dei generatori o di unità UPS, un'organizzazione dove multipli device condividono una stessa linea di ingresso, così che il carico di un device guasto può

essere preso in carico dai rimanenti device, similarmente a come succede per i sistemi RAID. Un parallelismo comune include una configurazione N+1 (permettendo un fallimento o manutenzione), N+2 (permettendo un fallimento perfino quando un'unità è in fase di manutenzione) e 2N (coppie di ridondanza).

Sistema di raffreddamento nei datacenter

Il sistema di raffreddamento è in qualche modo più semplice del sistema elettrico. Tipicamente, il piano del datacenter è rialzato — una griglia di metallo è installata 0.5-1.0 metri da terra (come mostrato nell'illustrazione 10). La parte sotto la griglia è spesso usata per i cavi di alimentazione, ma è principalmente usata per distribuire aria fresca ai rack di server.

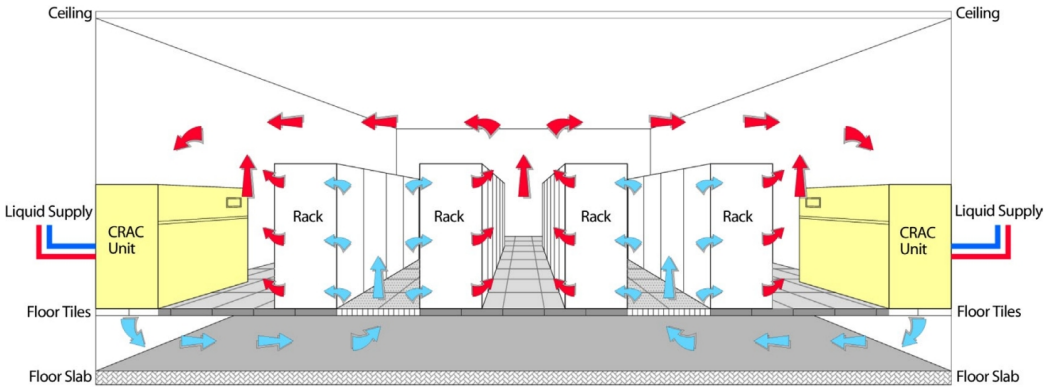


Illustrazione 10

Unità CRAC

Le unità CRAC (termine che deriva dagli anni '60 che sta per Computer Air Room Conditioning) aumentano la pressione del piano sollevato introducendo aria fredda. Questa aria fredda esce attraverso delle piastrelle perforate che sono

piazzate davanti ai rack di server; quindi essa fluisce attraverso i server, che espellono aria calda dalle bocchette posteriori. I rack sono posizionati in lunghe corsie in modo da alternare le corsie con aria calda da quelle con aria fredda per non mischiare i flussi con le due diverse temperature. Infatti mischiare aria calda con aria fredda riduce l'efficienza del sistema di raffreddamento; alcuni nuovi datacenter addirittura dividono le corsie con dei muri in modo da non lasciare che l'aria calda rientri nella stanza fresca [43]. Infine l'aria calda prodotta dai server ricircola indietro nelle prese d'aria delle unità CRAC che la raffreddano nuovamente e la soffiano una volta fredda nel pavimento rialzato.

Le unità CRAC consistono di spirali nelle quali è pompato un liquido refrigerante; delle ventole spostano l'aria attraverso queste spirali, raffreddandola. Un insieme di pompe ridondanti fanno circolare il liquido refrigerante freddo verso l'unità CRAC e il liquido refrigerante caldo verso un refrigeratore o una torre di raffreddamento che espelle il calore fuori dall'edificio. Tipicamente, il liquido refrigerante in ingresso ha una temperatura di 12-14°C e l'aria che esce dalla unità CRAC è attorno ai 16-20°C, fino all'ingresso dei server in cui sarà di 18-22°C: questo perché la temperatura aumenta man mano che ci si allontana dalle unità e CRAC e perché il ricircolo non può essere del tutto eliminato.

Raffreddamento libero

I nuovi datacenter spesso inseriscono una torre di raffreddamento per preraffreddare il fluido refrigerante usando un «raffreddamento libero» prima che raggiunga il refrigeratore. Il raffreddamento libero non è davvero libero da consumo di energia, ma comunque è molto più efficiente del raffreddamento con il refrigeratore.

Le soluzioni di raffreddamento libero basate su acqua usano delle torri di raffreddamento per dissipare calore. Queste torri usano un circolo di

raffreddamento separato nel quale l'acqua assorbe il calore del liquido refrigerante. Nella torre di raffreddamento, l'acqua calda quindi fluisce verso una struttura che ha una superficie molto larga nella quale può trasferire calore verso l'esterno attraverso l'evaporazione, quindi si raffredda. Si precisa che se l'aria è molto secca, l'acqua può essere raffreddata a una temperatura più bassa di quella dell'ambiente: poiché l'aria dell'ambiente è composta da una parte di acqua (umidità), l'evaporazione può abbassare la temperatura dell'acqua fino quasi al punto di rugiada, quindi più bassa della temperatura stessa dell'acqua. Le torri di raffreddamento funzionano bene nei climi temperati con bassa umidità; ironicamente, funzionano male nei climi molto freddi poiché bisogna inserire ulteriori meccanismi per prevenire la formazione di ghiaccio sulle torri.

Alternativamente si può usare un sistema di raffreddamento libero che faccia uso di un radiatore con glicol etilenico (liquido di raffreddamento [44]) montato al di fuori dell'edificio per dissipare il calore. Questo sistema lavora bene nei climi freddi, ma lavora meno bene in un clima moderato o caldo poichè il calore scambiato via aria è meno efficiente del calore scambiato via evaporazione. Altri sistemi eliminano del tutto la componente CRAC, portando dentro l'aria esterna e soffiando fuori l'aria all'interno, ovviamente se le temperature esterne lo permettono.

La maggior parte dei sistemi di raffreddamento viene messa in sicurezza con dei generatori (e a volte con unità UPS) poiché il datacenter non può operare senza un sistema di raffreddamento per più di pochi minuti prima di surriscaldarsi. In un datacenter tipico, i refrigeratori e le pompe possono aggiungere il 40% o più al carico critico totale.

Alcune considerazioni sul flusso d'aria

La maggior parte dei datacenter usano una configurazione con un pavimento rialzato come discusso precedentemente. Per cambiare l'ammontare della quantità d'aria inviata a un certo rack o una certa colonna di rack, possiamo semplicemente cambiare il numero di mattonelle perforate in una certa corsia sostituendole con mattonelle perforate o viceversa. Per far funzionare meglio il raffreddamento, il flusso d'aria fredda che entra attraverso le mattonelle deve essere proporzionato rispetto al numero di server presenti in quel rack; per esempio, se un rack ha 10 server e ognuno assorbe 28 m²/minuto d'aria, allora il flusso d'aria uscente dal pavimento dovrebbe essere 280 m²/minuto. Se è meno, tutta l'aria fredda verrà presa dai server che stanno in basso, mentre i server in alto prenderanno l'aria calda uscente dagli altri server; questo effetto non desiderabile viene spesso chiamato «ricircolo» poiché l'aria calda ricircola dall'uscita di un server verso l'entrata di un server vicino.

Questa necessità di una certa quantità d'aria limita la densità di potenza dei datacenter. Per una temperatura fissa per ogni server, il flusso d'aria di un rack deve crescere con la potenza consumata dal rack, e quindi il flusso d'aria uscente dal pavimento deve aumentare linearmente con la potenza. Questo a sua volta aumenta la pressione che dobbiamo avere nel pavimento rialzato, che aumenta la potenza delle ventole richiesta all'unità CRAC per inserire aria fredda nel pavimento rialzato. A basse densità, questo è facile da farsi, ma se si aumenta la pressione a un certo punto le leggi della fisica rendono il sistema economicamente impraticabile. Tipicamente, queste limitazioni rendono difficile aumentare la densità di potenza oltre ai 150-200W al m² senza aumentare enormemente i costi.

Come menzionato precedentemente, i nuovi datacenter hanno cominciato a separare fisicamente le corsie calde da quelle fredde per eliminare il ricircolo: in

questa configurazione l'intera stanza è piena di aria fresca e quindi tutti i server in un rack prenderanno l'aria con la stessa temperatura [43].

Raffreddamento in-rack

Il raffreddamento in-rack è una variante dell'idea di riempire l'intera stanza con aria fresca e può anche aumentare la densità e l'effetto del raffreddamento oltre i limiti convenzionali del piano rialzato. Tipicamente, un raffreddatore in-rack aggiunge un raffreddatore aria-acqua dietro al rack così che l'aria che esce dai server immediatamente fluisce attraverso spirali raffreddata dall'acqua, essenzialmente corto-circuitando il percorso fra l'uscita del server e l'ingresso delle unità CRAC. In alcune soluzioni, questo metodo diminuisce solo una parte del calore, in modo da ridurre il carico per le unità CRAC, mentre in altre soluzioni rimpiazza efficacemente le unità CRAC. Il maggior effetto negativo in questa tecnica è il fatto che necessita di portare a ogni rack acqua refrigerata, incrementando molto il costo delle tubature.

Datacenter basati sui container

I datacenter basati su container fanno un passo avanti rispetto al raffreddamento in-rack piazzando i server all'interno di container standard e integrando all'interno del container stesso un raffreddatore di aria e un distributore di energia elettrica. Similmente al raffreddamento in-rack, il container ha bisogno di una riserva di acqua refrigerata e usa eliche per rimuovere tutto il calore dall'aria che ci passa attorno. La gestione dell'aria è simile al raffreddamento in-rack e tipicamente permette grandi densità di potenza rispetto ai normali datacenter con pavimento rialzato. Quindi, un datacenter formato da container da tutte le funzionalità di un tipico datacenter (rack, CRAC, PDU, cablaggio, luce) in un piccolo pacchetto; come un datacenter regolare, devono essere integrati da infrastrutture esterne quali raffreddatori, generatori, unità UPS.

Efficienza energetica

Come è accaduto per la virtualizzazione anni or sono, il cloud computing sta manifestando delle intrinseche potenzialità green. Avere allocato su server remoti e virtualizzati i servizi, diminuisce la necessità di gestire infrastrutture fisiche e dunque anche l'impatto ambientale delle aziende o organizzazioni che scelgono tale modello. Analogamente l'uso di thin client nelle organizzazioni riduce la necessità di desktop di front end fornendo agli utenti la possibilità, a basso impatto ambientale, di accedere al proprio virtual desktop. "Thin" è un aggettivo della lingua inglese che significa magro, sottile; il "thin client" realizza il disaccoppiamento tra l'interfaccia utente e le risorse di computazione; separando le funzioni di Input/Output dal calcolo vero e proprio e "remotizzandole" si rende il sistema a disposizione dell'utente estremamente leggero, essenziale, indipendente dalla piattaforma di calcolo e senza necessità di manutenzione. Il calcolo e le risorse necessarie ad esso invece sono posizionate sul terminal server, quindi saranno centralizzate e facilmente manutenibili dall'amministratore di sistema [46]. Il thin client è un computer che dipende per tutte le attività di elaborazione da un terminal server; come thin client si possono utilizzare dispositivi ad hoc (progettati dunque per fungere da terminali), oppure dei PC con hardware poco potente. I benefici ambientali derivanti dalla sostituzione di PC desktop con thin client sono molteplici. I dispositivi hardware thin client consentono un risparmio di energia elettrica almeno del 50%. Un dispositivo [46] di nuova generazione ha una potenza di 5-30 W, contro i 150 W di un desktop tower. Il thin client consente all'utente di effettuare le operazioni di Input/Output con il server (e ridotte elaborazioni locali), inoltre consente all'utente di mantenere il necessario e voluto controllo sul suo "virtual desktop" e sul suo accesso personale. Rappresentano l'alternativa per compagnie, organizzazioni o istituzioni che abbiano un consistente parco installato di desktop. Non è, come evidente, valida per le organizzazioni che prevedono uso di laptop o altri dispositivi mobili,

ma l'universo di situazioni in cui si opera su postazioni fissa resta enorme. Oltre ai vantaggi ambientali, presentano notevoli vantaggi organizzativi, per esempio nelle scuole, come quello di alleggerire enormemente la gestione delle aule informatiche e del parco desktop poiché la manutenzione avviene solo sul terminal server [47].

Possiamo dire quindi che il cloud computing contribuisce a ridurre in vari modi le emissioni di CO2. Va detto tuttavia che la remotizzazione diminuisce, ma non azzera le emissioni di CO2: evidentemente bisogna calcolare l'impatto ambientale e i costi energetici che hanno i datacenter che ospitano soluzioni di Cloud Computing che ad oggi in media sono molto lungi dal rappresentare lo stato dell'arte. É per questo che l'efficienza energetica nei datacenter è alla base per risparmiare energia, quindi ridurre le emissioni di CO2 e i costi totali del datacenter.

L'efficienza energetica è stata uno dei maggiori impulsi per lo sviluppo di device mobili ed embedded, ma ora ha preso un focus per il computing più generale. I primi studi enfatizzavano la maggior durata della batteria ma si sono presto espansi allo studio per la riduzione di picchi di energia richiesti dalla CPU, poiché questi producevano calore che cominciava a ridurre le performance della CPU stessa. L'efficienza energetica è ora un punto chiave per le operazioni di server e datacenter, focalizzandosi sulla riduzione di tutti i costi relativi all'energia che includono capitali, spese operazionali e impatti ambientali. Molte tecniche per risparmiare energia che sono state sviluppate per i device mobili sono i primi candidati per risolvere questo nuovo dominio di problemi, ma un datacenter che ospita soluzioni di Cloud Computing è sicuramente molto diverso da un device mobile.

Efficienza in un datacenter

L'efficienza di un datacenter che ospita Cloud Computing è generalmente definita come l'ammontare del lavoro computazionale effettuato diviso per l'energia totale usata in tale processo.

Per misurare l'efficienza energetica all'interno dei datacenter vi sono diverse metriche di valutazione: le più famose e le più usate sono PUE e DCiE, altre sono state definite da Greenpeace e dall'Uptime Institue (per esempio SI-POM).

Il Power Usage Effectiveness riflette bene la qualità dell'infrastruttura del datacenter [35 Google] e cattura il rapporto fra la potenza richiesta totale e la potenza richiesta per l'IT, cioè l'energia consumata dall'apparecchiatura per il computing (server, apparecchiatura di rete, etc.):

PUE = Energia totale richiesta / Energia richiesta dal comparto IT

In letteratura per potenza richiesta dall'IT si fa riferimento alla potenza critica. Poiché i fattori PUE sono ignari della natura della computazione che viene fatta, possono essere misurati oggettivamente e continuatamente da un set di hardware per il monitoraggio senza alcuna disfunzione delle normali operazioni. Il PUE medio dei datacenter, secondo una ricerca del 2006 [48], è stato stimato essere, nell'85% dei datacenter, maggiore di 3.0, ovvero i sistemi elettrici e meccanici dell'edificio consumano il doppio di energia dell'attuale consumo per il computing; solo il 5% ha un PUE di 2.0.

Il Data Center Infrastructure Efficency (DciE) è il reciproco del PUE, ovvero:

DciE = (Energia richiesta dal comparto IT / Energia totale richiesta) x 100%

Esso mostra più rapidamente l'efficienza del datacenter; quindi, per esempio, un valore DciE del 33% (equivalente a un PUE di 3.0) indica che il comparto IT consuma il 33% dell'energia totale richiesta dal datacenter; quindi per 100$ spesi in energia, solo 33$ sono realmente usati dalle apparecchiature IT. Quindi il valore ideale per un datacenter è un DciE del 100%.

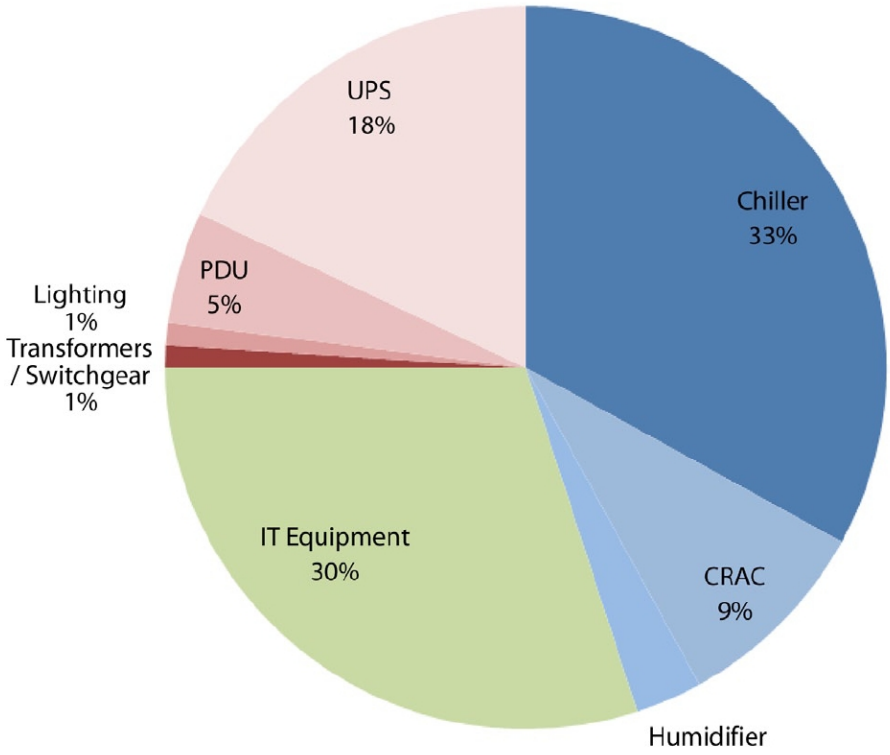


Illustrazione 11: Rapporto del consumo di energia all'interno di un datacenter

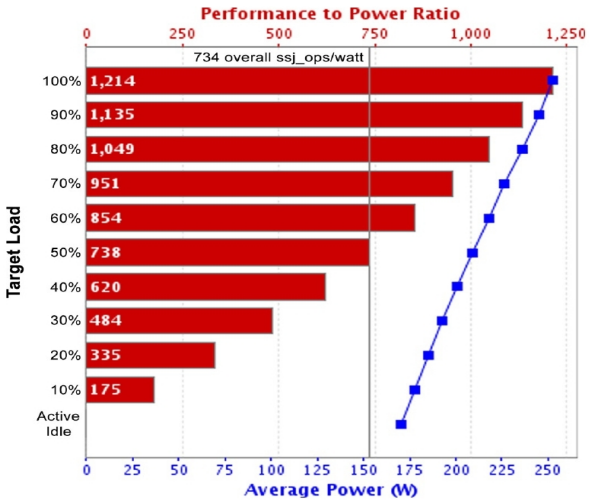
Alti valori per il PUE sono dovuti a molteplici fonti di overhead (si veda illustrazione 7); in un tipico datacenter con pavimento rialzato, i raffreddatori consumano la maggior parte dell'energia quando sono sovrasfruttati, tipicamente 30-50% del carico delle apparecchiature IT. Le unità di condizionamento (CRAC) consumano 10-30% del carico IT (la maggior parte per le ventole), seguiti dai sistemi UPS che consumano il 7-12% della potenza critica attraverso la conversione AC-DC-AC (le perdite sono maggiori quando il carico degli UPS è

basso). Altri elementi dell'infrastruttura (umidificatori, PDU, luci) contribuiscono ulteriormente a raggiungere livelli PUE alti. La poca efficienza di questi sistemi è causata da una storica mancanza di attenzione verso l'efficienza e non da limitazioni imposte dalla fisica. È comunemente accettato che un datacenter ben disegnato e ben operante dovrebbe avere un PUE minore di 2. I maggiori miglioramenti possono essere raggiunti con l'uso di torri di raffreddamento, un migliore e più efficiente circolo dell'aria e l'eliminazione delle perdite nella conversione dell'energia elettrica.

Le sorgenti della perdita di efficienza nei datacenter

Percorrendo la sttada che effettua la corrente elettrica nel datacenter, si parte dai trasformatori che trasformano la corrente da alto voltaggio (115kV) a un voltaggio medio (negli USA 13.2kV): questo processo è ragionevolmente efficiente e in ugual modo lo sono i trasformatori che trasformano la corrente a un voltaggio basso (480V). In ogni caso, le perdite durante la trasformazione sono meno dello 0.5%. I gruppi di continuità sono la fonte delle maggiori perdite di conversione, tipicamente lavorano con un'efficienza del 88-94% nei migliori casi (meno, se il carico è basso). I gruppi di continuità col volano e gli UPS ad alta efficienza che eliminano la necessità della doppia conversione (bypassando l'UPS durante le normali operazioni) possono raggiungere efficienze dell'ordine del 97%. Infine, una piccola quantità di potenza può essere persa portando la corrente elettrica a basso voltaggio (110 o 220V) ai rack con cavi molto lunghi: poiché un'infrastruttura di un datacenter può avere un piano più lungo o largo di 100 metri, 1-3% della potenza può essere persa nei cavi.

La maggior inefficienza, inoltre, avviene anche durante il raffreddamento. Le ventole per spostare aria fredda per una lunga distanza dalle unità CRAC ai rack consumano molta energia, ugualmente per spostare l'aria calda indietro alle unità CRAC; quel che è peggio è che durante questi lunghi tragitti, aria calda e aria



fredda possono mischiarsi, riducendo di molto l'efficienza delle unità CRAC [49]. Similmente, la comune pratica di tenere i datacenter molto freddi richiede che la temperatura dell'acqua all'interno dei raffreddatori sia di 10°C, accrescendo così il carico dei raffreddatori stessi; temperature così basse inoltre portano al formarsi di condensa all'interno delle eliche delle unità CRAC, riducendo così l'efficienza degli stessi e, ironicamente, richiedendo quindi più energia per deumidificare.

Benché la misurazione PUE cattura gli overhead dell'infrastruttura, non tiene d'acconto delle inefficienze degli apparecchi IT; server e altri apparecchi di computing usano meno del 100% della loro potenza in ingresso per la computazione. In particolare, molta parte dell'energia può essere persa nell'alimentatore del server, nei regolatori di voltaggio e nelle ventole di raffreddamento; molti alimentatori sono per l'80% efficienti e molte schede madri usano i regolatori di voltaggio che sono ugualmente efficienti, perdendo quindi circa il 30% dell'energia in ingresso per perdite di conversione.

Da aggiungere che sotto bassi livelli di utilizzo, i sistemi di calcolo tendono ad essere molto inefficienti rispetto a quando sono usati al massimo livello di utilizzo. Come si evince dall'illustrazione 5.3, il rapporto potenza/performance cala molto se il carico del sistema decresce poiché la potenza necessaria decresce molto più lentamente della performance: per esempio, l'efficienza energetica al

30% di carico è meno della metà dell'efficienza al 100%; in più, quando il sistema è in idle, consuma solo 60W in meno, che è più della metà del picco di consumo del server.

Carico vs. Efficienza

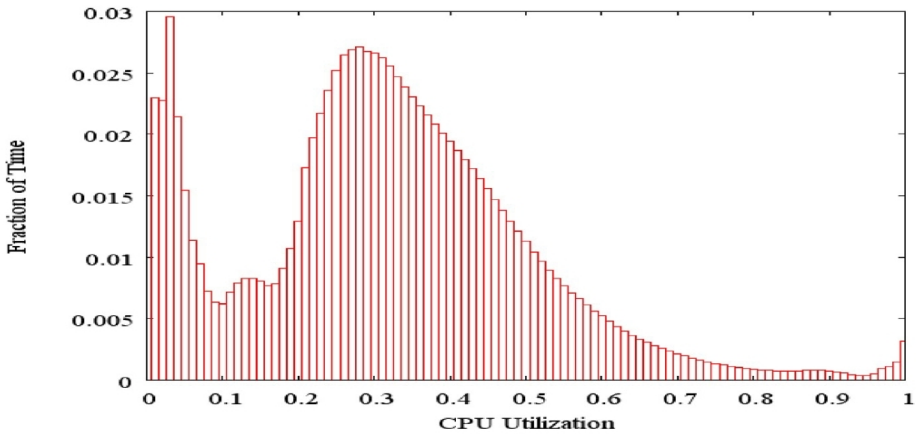


Illustrazione 12

L'illustrazione12 mostra un esempio di profilo di carico per un datacenter che ospita Cloud Computing: esso mostra l'utilizzo medio delle CPU di 5000 server Google in un lasso di tempo di 6 mesi; la linea di tendenza che si può vedere in media è che i server passano poco tempo a livelli di carico elevati. Invece, quel che si può notare è che spendono la maggior parte del tempo fra il 10% e il 50%; questo profilo di attività fa vedere che vi è un serio problema energetico nei server moderni poiché passano la maggior parte del tempo in regioni di carico dove sono maggiormente inefficienti. Si può inoltre notare che per diverso tempo i server sono idle.

In datacenter che offrono servizi Internet il bilanciamento del carico è fatto in modo da bilanciarlo fra tutti i server disponibili, creando quindi situazioni in cui,

quando il carico è basso, abbiamo un carico basso distribuito su più server invece che concentrato su poche macchine e lasciare idle le altre. Come si è discusso precedentemente, la discrepanza fra il carico dei server e l'efficienza energetica deve essere ricercata principalmente a livello hardware; il software da solo non può efficientemente risolvere queste problematiche.

Cause della cattiva proporzionalità energetica

Negli anni recenti, sono stati fatti molti sforzi per migliorare l'efficienza energetica delle CPU; diversamente non è stato fatto per altri componenti interni; lo switch verso sistemi multicore piuttosto che tentare di aumentare sempre di più la frequenza del core singolo ha portato all'aumento dell'efficienza energetica nelle CPU.

L'illustrazione13 mostra la potenza usata dai sottosistemi di un server Google da uno stato idle a uno stato di pieno carico; il contributo della CPU alla potenza del sistema è quasi 50% quando è al massimo carico, ma cade fino al 30% a bassi livelli di carico, rendendo questo componente il più proporzionale di tutti gli altri.

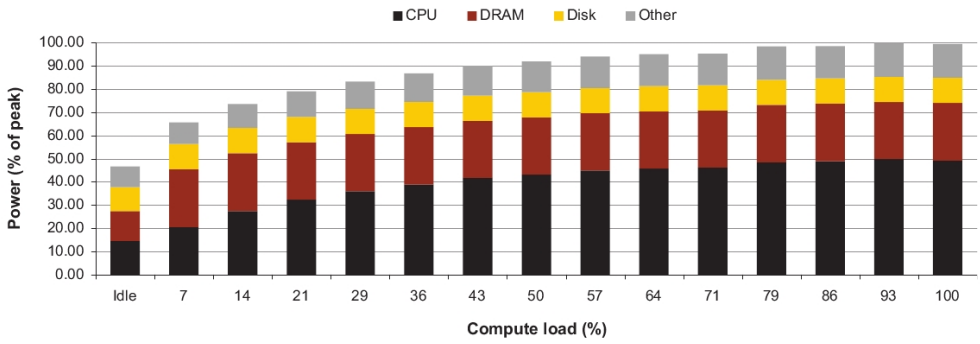


Illustrazione 13: Utilizzo energetico nei sottosistemi di un x86 mentre il carico cambia da idle a full

Fornitura dell'energia elettrica nel datacenter

Le ottimizzazioni di efficienza energetica sono naturalmente associate a costi minori dell'energia elettrica. Comunque, un altro fattore relativo all'energia è qualche volta più significativo del costo energetico stesso: il costo dell'infrastruttura del datacenter che ha la capacità di fornire un certo livello di potenza a un gruppo di server. Usando i costi definiti dall'Uptime Institute, fornire 1 watt per l'IT costa circa $10-22. Un'implicazione di questa struttura è che ridurre i consumi energetici a livello server può risultare in un vantaggio doppio poiché si possono diminuire anche le unità dell'infrastruttura che stanno attorno ai server stessi (PDU, CRAC, etc) e quindi diminuire il costo totale dell'infrastruttura.

4. Politiche e metodologie di miglioramento energetico

Un design attento all'efficienza può migliorare molto il PUE [50,51,52]. Benché molti datacenter hanno un PUE di 2 o maggiore, è possibile costruire datacenter molto più efficienti; oltre a soluzioni hardware, quali per esempio l'aumento dell'efficienza energetica dei singoli componenti dei server, sono percorribili soluzioni software e soluzioni fisiche. In questo capitolo verranno analizzate e discusse alcune di queste soluzioni.

Soluzioni fisiche

Vi sono diversi fattori per rendere efficiente il sistema energetico e di raffreddamento all'interno di un datacenter: miglioramenti nella gestione e configurazione delle sale server e dei rack stessi può aumentare l'efficienza energetica con un investimento iniziale molto basso.

Una delle soluzioni più facili è quella di alzare la temperatura in ingresso ai server fino a 25-27°C invece che della tradizionale 20°C. Virtualmente nessun attrezzatura server o di rete ha necessità di ricevere in ingresso aria a 20°C. L'aumento della temperatura in ingresso ai server permette temperature più alte nel liquido dei refrigeratori, aumentando così l'efficienza e riducendo il loro tempo di utilizzo quando combinati con raffreddamento libero. Similmente, una gestione efficiente del calore in uscita può aumentare di molto l'efficienza totale del raffreddamento: questo è uno dei motivi principali per cui datacenter basati su container sono molto efficienti. Un datacenter di 3000 m² con 1000 rack di server

che richiedono 10 KW ognuno di potenza, il costo iniziale delle unità CRAC è circa di 2-5 milioni di dollari; con un costo medio per l'elettricità di $100/MWh, il costo annuale è di circa 4-8 milioni di dollari [53]. Un datacenter che esegue la stessa configurazione e lo stesso carico di lavoro, ma mantiene la temperatura della sala server più alta di 5°C attraverso una gestione efficiente del sistema di raffreddamento, può risparmia il 20-40% dei costi, ovvero 1-3 milioni di dollari l'anno [54]. L'efficienza quindi di un sistema di raffreddamento (misurata con il Coefficent of Performance — COP) non è costante, ma aumenta con l'aumentare della temperatura che l'unità CRAC deve immettere nel pavimento rialzato. L'illustrazione2 mostra come il COP aumenta con più alte temperature per una tipica unità CRAC con evaporazione di acqua. Per esempio, se l'aria ritornata all'unità CRAC è di 20°C e rimuoviamo 10KW di calore, raffreddandola fino a 15°C, spendiamo circa 5.26KW; se invece alziamo la temperatura dell'aria in uscita dall'unità CRAC fino a 20°C, ogni cosa all'interno della sala server avrà una temperatura maggiore di 5°C. Raffreddare ora di 10KW lo stesso volume d'aria ma partendo da una temperatura di 25°C richiede solo 3.23KW. Questo processo fa risparmiare quasi il 40% dell'energia richiesta.

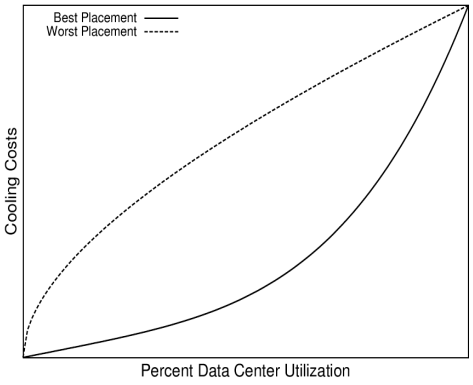


Illustrazione 14

Questo ovviamente deve essere messo a confronto con l'affidabilità dell'hardware e con possibili guasti di quest'ultimo: infatti un server tipico ha bisogno di temperature dell'aria in ingresso di circa 20-30°C; ogni 10°C in più rispetto alla temperatura di 21°C fa decrescere la stabilità e l'affidabilità a lunga durata dei componenti elettronici di circa il 50% [55]; altri studi hanno dimostrato che

l'incremento di 15°C di un hard disk fa aumentare il rischio di guasto dello stesso di un fattore 2 [56,57].

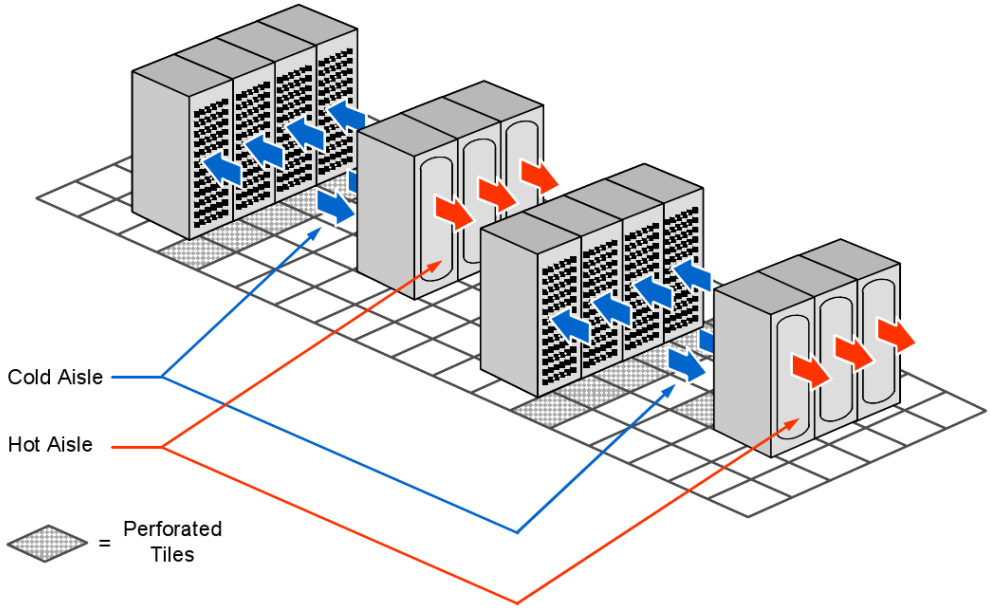


Illustrazione 15: Configurazione di corsie calde e fredde

Per aumentare l'efficienza del sistema di raffreddamento, è possibile scegliere di raffreddare i rack di server con raffreddamento a liquido, infatti l'aria è un mezzo di raffreddamento davvero inefficiente: un litro d'acqua può assorbire circa 4000 volte di più il calore assorbito dallo stesso volume di aria. Purtroppo però i sistemi di raffreddamento a liquido hanno dei costi infrastrutturali molto più alti di un classico raffreddamento ad aria.

Molto importante è organizzare la sala server e più in particolare i rack di server in modo da creare delle corsie fredde e calde, fondamentale per limitare il ricircolo dell'aria e per non fare mischiare l'aria fredda con quella calda; ciò infatti limita di molto l'efficienza totale del sistema di raffreddamento. L'illustrazione15 mostra come dovrebbero essere disposti i rack di server, come raccomandato

dall'American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). Da notare come le mattonelle perforate siano posizionate solo sotto i rack di server, in modo da massimizzare l'aria fredda in ingresso ai server stessi.

L'illustrazione4.7 mostra quindi i benefici che si hanno da una configurazione a corsie calde e fredde: l'aria calda ritorna all'unità CRAC non mischiandosi con l'aria fredda che viene presa in ingresso dai server, aumentando quindi l'efficienza dell'unità CRAC stessa.

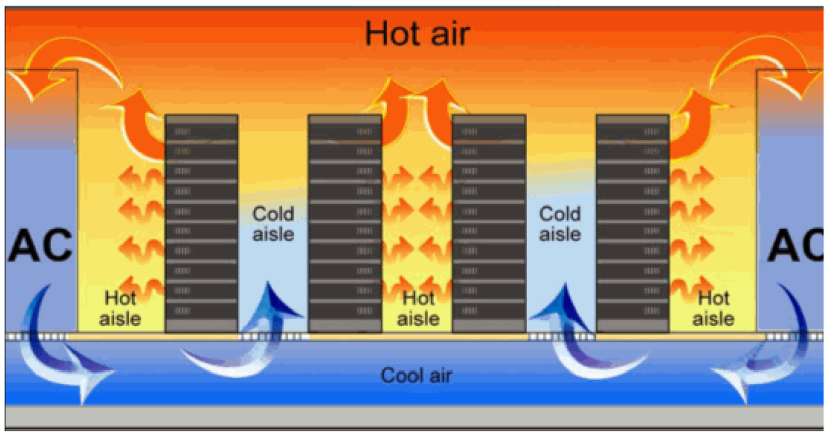


Illustrazione 16: Circolo ottimale di aria calda e aria fredda all'interno del datacenter

Un altro problema che limita fortemente l'efficienza dei datacenter è la presenza di hotspot (ovvero punti caldi all'interno della sala server): questi punti caldi aumentano la richiesta di raffreddamento dell'intera sala server, nonostante solo alcuni punti ne abbiano realmente bisogno. Per sopperire a questa problematica, vi sono sia alcune soluzioni software che verranno discusse in seguito, sia alcune configurazioni possibili del sistema di raffreddamento: la più semplice è quella di dividere la sala server in più spazi gestiti da sistemi di raffreddamento differenti; in questo modo la presenza di hotspot limiterà l'inefficienza solo a un range di macchine più piccolo. Un'altra soluzione è quella di usare un raffreddamento a liquido localizzato, il quale possa raffreddare in modo più specifico ogni server.

Un'ulteriore possibile soluzione per diminuire i costi del raffreddamento è quella di posizionare i datacenter in luoghi geografici con temperature più basse: questo permetterebbe di usare più massicciamente il raffreddamento libero e quindi risparmiare sui costi. Vi è però da fare un tradeoff fra il costo dell'energia elettrica e gli eventuali risparmi sui costi di raffreddamento: a titolo d'esempio, in Alaska (Anchorage) l'elettricità costa 0.1389 $/KWh, mentre in Ohio costa 0.067$/KWh; in più, come detto precedentemente, un clima troppo rigido provoca il congelamento dell'acqua se si usa il raffreddamento libero, quindi vi servirà ulteriore energia per prevenire la formazione di ghiaccio. Se si hanno a disposizione due datacenter posizionati in due parti opposte del globo (per esempio USA e Cina), per fare delle computazioni si potrebbe usare il datacenter negli USA quando è notte e il datacenter della Cina quando è giorno, risparmiando quindi sui costi di raffreddamento poichè si usa quel datacenter in cui è notte, ovvero quando l'atmosfera ha temperature significativamente più basse.

Soluzioni software

Virtualizzazione

La virtualizzazione probabilmente è la soluzione software che offre maggiori risparmi energetici a fronte di una spesa minima. Poiché, come detto precedentemente, i server a basso carico sono estremamente inefficienti tramite la virtualizzazione è possibile rendere pienamente utilizzata una macchina fisica eseguendo su di essa più istanze di macchine virtuali, riducendo quindi l'inefficienza energetica e in secondo luogo il numero di macchine fisiche necessarie.

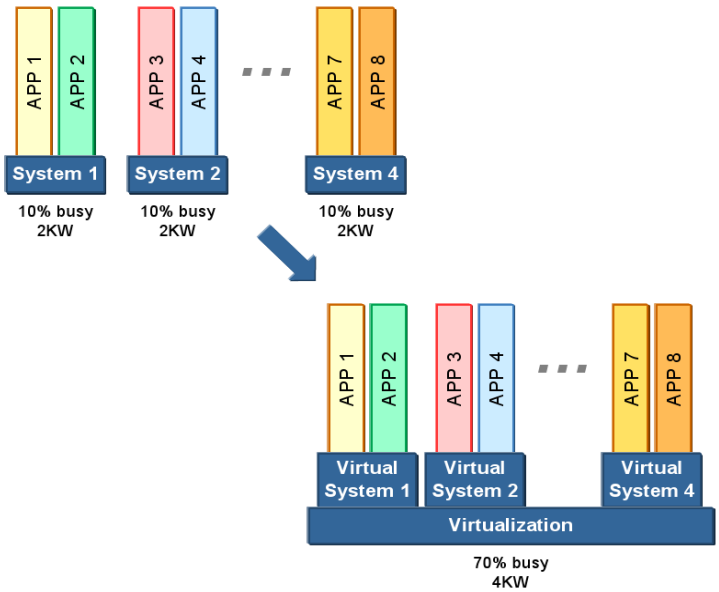


Illustrazione 17: La virtualizzazione consente il consolidamento di sistemi

Eseguire più macchine virtuali su uno stesso server fisico però presenta delle sfide architetturali: dobbiamo decidere quale combinazione applicazione/macchina virtuale sia buona e quale no; per esempio, eseguire due applicazioni che hanno computazioni intensive sulla stessa macchina fisica probabilmente non darà buoni risultati rispetto a eseguire una combinazione di applicazioni CPU bound e I/O bound su una stessa macchina fisica. Si possono dividere le decisioni da prendere in due categorie: decisioni statiche e decisioni dinamiche.

Le decisioni architetturali che comprendono la scelta di quali applicazioni eseguire su quali macchine virtuali sono chiamate decisioni statiche; queste decisioni richiedono uno studio del profilo dell'applicazione stessa e una conoscenza di quali risorse l'applicazione richiederà. Queste decisioni sono chiamate statiche poiché sono da effettuarsi prima del dispiegamento delle macchine virtuali, infatti una volta che sono state avviate è difficile cambiare il mix di applicazioni sulle macchine virtuali a meno che esse non siano state

progettate per questo scopo; al di fuori delle applicazioni, in ogni caso l'intera macchina virtuale può essere spostata da un sistema fisico ad un altro.

Le decisioni dinamiche invece permettono la decisione, a un particolare istante, su quale macchina fisica una macchina virtuale dovrebbe essere eseguita. Questa scelta è da farsi in real-time, dopo aver studiato il carico di lavoro corrente delle diverse macchine virtuali. Diversi hypervisor commerciali (per esempio Vmotion di VMware) sono in grado di salvare lo stato delle macchine virtuali, metterle in pausa, spostarle verso una diversa macchina fisica e farne il resume dallo stato in cui sono state messe in pausa.

Per effettuare queste scelte, l'infrastruttura ha bisogno di dati sul carico corrente, sul consumo energetico istantaneo, e di un sistema che fornisca tracciabilità delle informazioni. Avere questa tracciabilità permette di effettuare scelte migliori quando si allocano macchine virtuali su macchine fisiche. Per esempio, se una macchina fisica ha un carico eccessivo, si può spostare una macchina virtuale verso una macchina che ha un carico più basso; ugualmente questo può essere fatto per consolidare un sistema, ovvero se una macchina ha un carico di lavoro troppo basso, sotto la soglia minima di efficienza.

In aggiunta a ciò, la virtualizzazione può essere usata anche per migliorare l'efficienza del sistema di raffreddamento. Se il sistema di raffreddamento e il sistema di gestione delle macchine virtuali comunicano, può essere configurato un sistema che eviti hotspot, allocando per esempio macchine virtuali CPU bound in punti della sala server più freddi.

Load balancing

Come discusso precedentemente, alzare la temperatura globale all'interno della sala server porta a un facile risparmio energetico e quindi di costi. I problemi che possono insorgere tuttavia sono principalmente due:

- la temperatura di un server può andare oltre la temperatura limite, spegnendosi automaticamente; questo può portare a un pericoloso denial of service e un aumento di carico verso gli altri server, creando un effetto domino difficile da riparare velocemente;

- l'affidabilità dei componenti diminuisce, come descritto precedentemente

In quest'ottica sono stati sviluppati software di bilanciamento del carico, in particolare si cita l'algoritmo OnePassAnalog di Moore et al. [58]: esso assegna carichi di lavoro ai server in modo da bilanciare la temperatura dell'aria in uscita dai bocchettoni esterni dei server stessi, evitando così la formazioni di hotspots difficili da raffreddare. Questo viene fatto mantenendo un dispatcher iniziale che smista le richieste verso alcune macchine.

Inoltre, si cita l'algoritmo Freon e Freon-EC, studiati da Heath et al. [59]. Freon è un sistema che gestisce le temperature dei componenti di un datacenter tramite un load balancer a monte; l'obbiettivo principale di Freon è quello di gestire le emergenze termiche senza usare l'approccio tradizionale di spegnere i server. Inoltre, Heath et al. ha sviluppato Freon-EC, che combina politiche di conservazione dell'energia e gestione termica: la politica così ottenuta spegne i server quando questo non degrada il throughput del datacenter.

Freon usa un balancer che usa una politica di distribuzione delle richieste basate sul carico. L'uso tradizionale delle distribuzioni segue una politica eterogenea: un

server che ha il doppio della velocità avrà il doppio del carico. Il load balancer usato da Freon è LVS, un modulo kernel di Linux [60]: esso è formato da LVS e una coppia di demoni che comunicano fra di loro. Più precisamente, vi è un processo chiamato tempd (Temperature Daemon) su ogni server che controlla la temperatura dei dischi e della CPU sul server stesso. Tempd si sveglia periodicamente (negli esperimenti di Heath et al. una volta al minuto) per controllare le temperature dei componenti. Quando la temperatura supera una soglia prefissata, viene inviato un pacchetto UDP al processo Freon sul nodo load balancer, chiamando quindi admd (Admission Control Daeomon) per aggiustare il carico inviato al server troppo caldo. La comunicazione verso il demone e il conseguente aggiustamento del carico sono ripetute periodicamente fino a che la temperatura del server non sia ritornata sotto il livello di soglia. Inoltre, se la temperatura si abbassa sotto una soglia minima prefissata, tempd ordina ad admd di eliminare ogni restrizione su quel server poiché ora è freddo e di mandare richieste al server. Se la temperatura invece rimane stabile fra la soglia minima e massima, Freon non aggiusta la distribuzione del carico e non vi sono comunicazioni fra i due demoni.

Freon spegne il server solo se la temperatura di quest'ultimo eccede una soglia massima: questa soglia determina la temperatura massima che possono raggiungere i componenti senza che essi abbiano seri danni o problemi di affidabilità.

Freon-EC invece oltre a gestire le emergenze termiche, conserva anche energia combinando le politiche di risparmio energetico e di gestione delle temperature. Per conservare energia, Freon-EC spegne più server che può senza degradare le performance; Freon-EC sceglie quali server spegnere a seconda della loro temperatura e della locazione fisica nella stanza server. In particolare, Freon-EC associa a ogni server una particolare regione; per esempio, uno schema intuitivo

può essere quello di definire due regioni se la sala server è refrigerata da due condizionatori: il guasto di un sistema di raffreddamento condizionerà maggiormente solo una regione. Rispettando la politica di gestione delle emergenze termiche, Freon-EC spegne i server caldi e li rimpiazza con server potenzialmente non affetti dall'emergenza, per esempio di un'altra regione.

Spegnere un server comporta istruire LVS di non inviare più richieste al server, aspettando che le connessioni correnti terminino e quindi spegnerlo. Accendere un server comporta fargli fare il boot, aspettare che sia pronto ad accettare connessioni e istruire LVS a cominciare a inviargli richieste. La decisione di aggiungere o rimuovere un server dal cluster attivo è basata sulle proiezioni di utilizzo dei componenti: si aggiunge un server quando quando le proiezioni di utilizzo di tutti i componenti sono maggiori di una certa soglia. Ugualmente, Freon-EC spegne un server quando quest'azione fa rimanere l'utilizzo medio delle componenti sotto una certa soglia.

Da sottolineare che Freon implementa una certa forma di riduzione della frequenza della CPU (throttling) riducendo il carico della macchina stessa; poichè questo throttling è eseguito non dalla macchina in se ma dal load balancer, esso viene chiamato «remote throttling».

Sleep di server

È possibile raggiungere un'alta proporzionalità energetica usando diversi tipi di sleep mode; questi modi sono detti «inattivi» poiché la macchina non può essere usata durante questi stati e tipicamente vi è una latenza e un uso di energia quando la macchina è riattivata da uno stato di inattività. Modi inattivi furono originariamente sviluppati per device embedded o mobile ed ebbero molto successo in questo dominio.

Il software MUSE di Chase et al. [61] si prefigge come scopo quello di ridurre l'insieme di server disponibili dinamicamente, in modo da adattarsi alle richieste in ingresso e in modo da non infrangere la SLA. I risultati degli esperimenti effettuati da Chase et al. Mostrano che si può ridurre fino al 29% l'energia usata da un tipico datacenter che ospita servizi Web. Grazie a MUSE inoltre si può eliminare il problema dell'effetto domino, come detto precedentemente, dovuto a guasti delle unità CRAC o a spegnimenti improvvisi di server (e quindi redistribuzione del carico su altri server): infatti verrà ridotta la richiesta energetica con lo spegnimento di alcuni server, in modo da rimanere sotto a un limite massimo di temperatura.

MUSE è un sistema operativo per i datacenter; esso mantiene un set attivo di server per ogni servizio che possono rispondere a qualsiasi richiesta in ingresso; poiché i server possono essere condivisi da più servizi, il set attivo può avere delle sovrapposizioni. Il sistema redirige le richieste in modo da avere un set di server attivi minimale; questo set lavora con un carico di lavoro prefissato, mentre i server che hanno un carico di lavoro inferiore a quella data soglia vengono messi in stand-by. I server spenti oltre a risparmiare energia di per sè, fanno risparmiare energia anche al sistema di raffreddamento poiché non producono calore. Nel loro studio, Choose et al. hanno scelto di mettere i server in stand-by poiché il costo fisso di tenere acceso un server in idle era troppo elevato: questo perché l'alimentatore impone un costo fisso energetico poiché deve mantenere la capacità di rispondere velocemente a possibili richieste.

Un effetto collaterale di questo algoritmo è la minor durata dell'hardware nel tempo di alcuni componenti: per esempio, il cambiamento di stato da attivo a spento di un elemento quale l'hard-disk può ridurre il tempo di vita di quest'ultimo, mentre può aumentare la vita di altri componenti (per esempio la motherboard); la vita di un tipico hard-disk moderno è stimata essere dai 30,000 ai

40,000 cicli start/stop. Una possibile soluzione è quella di mantenere i server stateless (senza hard disk) spenti mentre gli storage network accesi anche se idle e in quest'ultimi, se possibile, usare disk drives più performanti. Infatti i disk drive spendono la maggior parte della loro energia mantenendo la rotazione dei piatti, circa il 70% dell'energia totale per dischi ad alti RPM: Sankar et al. [62] ha esplorato differenti architetture differenti per i disk drive, osservando che i movimenti delle testine sono proporzionali all'energia spesa per muoverle, e un disco con più basse velocità rotazionali e testine multiple può raggiungere performance simili e meno energia spesa se comparato con un disco a testina singola con alti RPM.

Un secondo effetto collaterale è che la transizione di stato impone un ritardo nella risposta che varia da diversi secondo a minuti. Tuttavia, vi sono alcune tecniche con bassi tempi di latenza, com'è il caso dello stato della CPU a basso consumo energetico (come lo stato x86 C1E); sfortunatamente, queste tendono ad essere i modi col minor grado di risparmio energetico. Una grande quantità di energia può essere risparmiata da stati di inattività come fermare la rotazione dei dischi: un disk drive con dischi fermi può anche non usare alcuna energia, ma una transizione verso un modo attivo ha una penalità 1000 volte più alta rispetto a un accesso regolare. Queste grandi penalità riducono moltissimo l'uso di modalità inattive per i server: sarebbero di guadagno solo in quelle situazioni in cui il server non è usato per diversi minuti, cosa che accade raramente nei server.

Tuttavia vi sono dei modi di risparmio energetico «attivi», ovvero quelli che salvano energia perdendo in performance ma non richiedendo l'inattività: un esempio è lo scaling del voltaggio della CPU poiché può sempre eseguire istruzioni anche se a un basso rate; un altro esempio di classe di questi modi è leggere e scrivere dati da disk drive a più basse velocità rotazionali. Diversamente dai modi inattivi, i modi di risparmio energetico attivi sono utili anche quando la

latenza e la penalità energetica per passare a modi di completo carico sono significativi: poiché le macchine nei modi attivi sono operative, i sistemi possono rimanere in stati di basso consumo energetico fino a che essi rimangono sotto un certo livello di carico. Questi livelli di basso carico sono molto più comuni e lunghi rispetto a momenti in cui la macchina è completamente idle e quindi l'overhead del passaggio fra stato di risparmio energetico e stato totalmente attivo può essere ammortizzata più efficacemente.

5. Comparazione di casi di studio

Nel mondo il Cloud Computing è ormai una realtà concreta. Attualmente sul mercato sono presenti diverse aziende che offrono servizi Cloud Computing: da colossi come Amazon, IBM, Google, Microsoft, Apple, fino a realtà cloud offerte da società minori come Rackspace, GoGrid, Joyent.

In questo capitolo si andranno ad illustrare i servizi offerti da Amazon, Microsoft e Google; tali soluzioni inoltre verranno confrontate fra di loro sulle basi di alcune caratteristiche importanti per l'adozione di politiche Green volte al risparmio energetico, nell'ottica di un Cloud Computing "verde". Tuttavia, poichè non è stato possibile reperire informazioni di questo tipo sui servizi di Amazon, esso verrà solamente introdotto e non inserito nella comparazione.

Amazon EC 2

Amazon Elastic Compute Cloud (noto anche come «EC2») è un servizio web che fornisce capacità computazionale ridimensionabile all'interno di un'infrastruttura di Cloud Computing. L'obbiettivo del cloud di Amazon è quello di fornire applicazioni software sempre più semplici ed intuitive, sia per l'utente che per lo sviluppatore. Esso, infatti, fornisce il controllo completo delle risorse di elaborazione e consente di eseguire le istanze direttamente su Amazon, ambiente collaudato e affidabile.

L'EC2 riduce il tempo richiesto per ottenere le istanze; tale proprietà permette una rapida scalabilità delle capacità, sia verso l'alto che verso il basso, in funzione del

comportamento dell'utilizzatore. Attualmente gli utenti che si servono di tale strumento possono creare, avviare e chiudere le istanze del server in maniera del tutto autonoma, da qui il termine «elastico».

Amazon EC2 cambia l'economia del Cloud Computing consentendo di pagare solo per le capacità che effettivamente vengono utilizzate.

Funzionalità di Amazon EC2

Amazon EC2 presenta un vero e proprio ambiente virtuale che consente di utilizzare le interfacce di servizi web per lanciare le istanze sotto una varietà di sistemi operativi, impiegando per ogni ambiente applicazioni specifiche. Per utilizzare Amazon EC2 occorre semplicemente:

- Creare un Amazon Machine Image (AMI), che contenga le applicazioni, le librerie, i dati e le impostazioni di configurazione, o in alternativa, caricare una macchina virtuale già pre-configurata per ottenere una AMI che sia immediatamente funzionante.

- Caricare l'AMI in Amazon S3. EC2 fornisce gli strumenti che rendono più semplice la memorizzazione della stessa. Amazon S3 fornisce un sicuro, affidabile e veloce «magazzino» per memorizzare le «immagini».

- Utilizzare il servizio web di Amazon EC2 per la configurazione della sicurezza e l'accesso alla rete.

- Scegliere il tipo di istanze da utilizzare, il sistema operativo che si desidera e successivamente farle eseguire (lancio istanza — termino istanza); controllare il numero di istanze dell'AMI utilizzando, se necessario, il servizio Web API o altri strumenti di gestione.

- Stabilire se si desidera eseguire le istanze in locazioni multiple (IP statico endpoints, o l'attach persistent block storage).

- Pagare solo le risorse che si consumano effettivamente, come ad esempio il tempo o il trasferimento di dati.

Servizi hilights

- Elasticità — Amazon EC2 consente di aumentare o diminuire la capacità di calcolo nel giro di pochi minuti, non in ore o in giorni. È possibile utilizzare centinaia o addirittura migliaia di server contemporaneamente, tutto questo grazie al controllo ottenuto mediante le API del servizio Web. Le applicazioni degli utenti si possono scalare in maniera del tutto automatica, sia verso l'alto che verso il basso, a seconda delle esigenze.

- Completamente controllata — L'utente ha il completo controllo delle sue istanze. Ha accesso come root in ciascuna di esse e può interagire con loro come si farebbe con qualsiasi macchina. Le istanze possono essere riavviate in remoto usando le API del servizio web. Si può anche ottenere un accesso alla console di uscite delle proprie istanze.

- Flessibilità — L'utente può scegliere tra varie tipologie di istanze, ha a disposizione vari sistemi operativi e pacchetti software. Amazon EC2 consente di selezionare una configurazione di memoria, di CPU, e della migliore istanza di archiviazione possibile, in funzione della scelta del sistema operativo e delle applicazioni. Ad esempio, la scelta di sistemi operativi comprende numerose distribuzioni di Linux, di Microsoft Windows Server e di OpenSolaris.

- Integrazione — Amazon EC2 è progettato anche per l'utilizzo congiunto con altri Amazon Web Services. Esso, infatti, lavora in collaborazione con Amazon Simple Storage Service (Amazon S3), Amazon SimpleDB e Amazon Simple Queue Service (Amazon SQS) per fornire una soluzione completa di calcolo, l'elaborazione e l'archiviazione di ricerca in una vasta gamma di applicazioni.

- Affidabilità — Amazon EC2 è altamente affidabile ed offre un ambiente in cui le istanze possono essere comodamente previste, quindi, facilmente commissionate, ma anche rapidamente sostituite (database di istanze pre-configurate). I vari servizi vengono eseguiti all'interno dell'infrastruttura di rete e dei data center di Amazon.

- Personalizzazione — Amazon EC2 fornisce agli utenti particolari strumenti per creare applicazioni tra i quali: Amazon Elastic Block Store, Multiple Locations, Elastic IP Addresses.

- Sicurezza — Amazon EC2 fornisce delle interfacce web service per configurare le impostazioni del firewall che controllano l'accesso alla rete e tra gruppi di istanze.

- Economicità — Amazon EC2 fornisce a tutti i suoi utenti i vantaggi finanziari derivanti dalla scalabilità. Le spese sostenute da ogni utente per il pay-for-use risultano essere molto più convenienti se messe a confronto col tradizionale e significativo investimento iniziale dovuto all'acquisto e alla manutenzione dell'hardware.

L'impiego dell'EC2 svincola da molte complessità dovute alla pianificazione delle capacità, trasformando quello che comunemente è rappresentato da grandi costi fissi in costi variabili molto più piccoli ed eliminando la necessità di acquistare

più di una rete di sicurezza volta ad essere utilizzata ogni qualvolta occorre gestire i saltuari picchi di carico.

Microsoft Azure

Microsoft Azure Service Platform è la nuova piattaforma di servizi che Microsoft offre per il cloud computing.

La sua prima apparizione risale alla PDC 2008 [63] dove è stata presentata alla comunità informatica in versione per sviluppatori. La nuova architettura uscita in versione definitiva il 1 gennaio 2010, è disponibile dal 1 febbraio 2010.

La piattaforma dei servizi Azure sfrutta un sistema operativo specializzato, Windows Azure, «pensato» per operare su internet, che viene eseguito nei cluster dei datacenters di Microsoft, posizionati in diversi punti del mondo (attualmente solo in USA e presto in Europa).

La gestione hardware è quindi interamente demandata a Microsoft che si occupa delle macchine, della connettività e dispone di sistema automatici di monitoraggio e di gestione delle macchine virtuali.

È un sistema operativo «on the cloud» che offre servizi per lo sviluppo, per l'hosting e la gestione delle applicazioni che vi «gireranno».

Il mondo cloud fornisce ai proprio utilizzatori soluzioni software differenti. Esiste, infatti, la possibilità di affidarsi completamente alla cloud reperendo nella stessa sia i dati da elaborare che gli applicativi in grado di manipolarli.

Per far fronte a tale pluralità di servizi, Azure supporta un gran numero di tecnologie di cloud, ognuna delle quali adatta ad uno specifico set di servizi destinati agli sviluppatori degli applicativi.

Microsoft garantisce la completa compatibilità di Azure con le versione di Windows più diffuse (Server, Vista, 7, XP, Mobile) ed alcuni suoi componenti possono essere impiegati anche in altri sistemi operativi.

Azure Services Platform è caratterizzato da una pluralità di applicativi e di servizi realizzati per eseguire specifici processi. Prevalentemente Azure è costituita dalle seguenti componenti:

- Windows Azure: fornisce un ambiente di sviluppo basato su Windows per l'esecuzione delle applicazioni e la memorizzazione dei dati sui server di Microsoft datacenter.

- Microsoft .NET Services: offre infrastrutture distribuite sia per i servizi di tipo cloud-based e sia per le applicazioni locali.

- Microsoft SQL Services: fornisce servizi di memorizzazione dati nella cloud basati su SQL Server, web-based distributed relational databases.

- Live Services: attraverso il Live Framewrok, fornisce l'accesso ai dati e a varie applicazioni, tra cui quelle di Microsoft. Il Live Framework permette anche la sincronizzazione dei dati tra computer desktop e dispositivi, la ricerca, il download di applicazioni e molto altro ancora.

Anche se non presenti nello schema di Azure, bisogna citare anche due applicazioni di tipo aziendale che in ambito business sono molto utilizzate e che

impiegano attivamente l'architettura cloud: Microsoft SharePoint Online e Microsoft Dynamics CRM Online Services.

Microsoft SharePoint Online è un software che permette la gestione centralizzata delle risorse e del flusso di lavoro. Esso, infatti, offre la possibilità di fornire ai propri utilizzatori un metodo sicuro ed efficiente per creare collaborazioni fra dipendenti, trovare le risorse organizzative, gestire i contenuti ed il flusso di lavoro e ricavare i dati necessari per prendere decisioni basate su informazioni aggiornate. I dipendenti possono creare e gestire siti intranet personalizzati appositamente per un team o un progetto, allo scopo di favorire la collaborazione e la condivisione dei documenti.

Microsoft Dynamics CRM Online Services è un software dedicato alle aziende che ha come caratteristica principale la drastica riduzione dei tempi di messa in esercizio (instant deployment over the Internet) e l'assenza di una struttura HW dedicata. Il CRM si occupa prevalentemente delle relazioni fra l'azienda che lo impiega ed i suoi clienti. Esso è caratterizzato principalmente da tre moduli: acquisizione (marketing), vendita (sales), postvendita (customer service).

Questi due prodotti sono stati progettati in origine per fornire supporto alle aziende. La possibilità che tali soluzioni software possano interfacciarsi con la piattaforma Azure le renderà ancora più performanti. Questo incremento dell'efficienza è dovuto all'indipendenza hardware ed all'elasticità propria di Azure, in unione con i servizi forniti da Dynamics CRM per la gestione dei rapporti cliente-azienda e da SharePoint per la coordinazione e la condivisione dei flussi di lavoro tra i vari dipendenti della medesima impresa.

Questa tipologia di prodotti offre un sistema operativo ed un set di servizi per gli sviluppatori, che possono essere impiegati in maniera del tutto indipendente. Godendo di flessibilità ed interoperabilità, la piattaforma Azure può essere usata

per progettare nuove applicazioni cloud o migliorare quelle «vecchie» equipaggiandole con capacità tipiche della rete a cloud.

La sua è un'architettura aperta (possibilità di personalizzazione della rete) che fornisce agli sviluppatori la capacità di scegliere il tipo di sistema su cui far girare le proprie applicazioni web, come ad esempio, sulla propria macchina, su più PC, su server o su soluzioni ibride realizzate per sfruttare il meglio delle varie realtà.

Azure, come ogni architettura di questo tipo, evita la forte dipendenza dai vincoli hardware consentendo agli sviluppatori di creare rapidamente e facilmente nuove applicazioni di tipo cloud usando le loro abilità con i prodotti Microsoft, impiegando gli ambienti di sviluppo di Visual Studio o .NET Framework.

Oltre al codice di tipo .NET, Azure in futuro supporterà diversi linguaggi di programmazione e di ambienti di sviluppo. La gestione dell'infrastruttura è automatizzata mediante una piattaforma progettata per gestire un gran numero di richieste, scalabili dinamicamente e per soddisfare le necessità d'impiego mediante il modello di pagamento pay-as-you-go (del tutto simile al pay-for-use).

Il provider Azure è impostato su uno standard aperto (predisposizione all'implementazione di vari standard applicativi web-oriented) basato su ambienti di interoperabilità, i quali sfruttano protocolli Internet multipli, incluso l'HTTP, il PHP, il Representation State Transfer (REST), il Simple Object Access Protocol (SOAP) e il Extensible Markup Language (XML).

Microsoft, inoltre, offre anche applicazioni di tipo cloud già pronte per i clienti, come Windows Live, Microsoft Dynamics e molti servizi on-line di tipo business come Exchange Online e Office SharePoint Online.

L'Azure Service Platform lascia ai suoi sviluppatori la più completa libertà nella creazione delle applicazioni, limitandosi ad offrire solamente i componenti fondamentali, come potenza di calcolo, di memorizzazione e l'assistenza tecnica nelle applicazioni a cloud, con personale competente.

Sono disponibili, inoltre, per gli utenti di Visual Studio le librerie, esempi di codice, ed alcuni tool finalizzati alla creazione di applicativi destinati alla cloud.

La piattaforma dei servizi può gestire attualmente applicazioni di .NET Framework scirtte in C# che devono supportare ASP.NET e tutti i metodi ed i servizi necessari alla pubblicazione dell'applicazione stessa nel cloud.

Sono stati realizzati anche due particolari SDK per garantire l'interoperabilità con Azure: il Java SDK e il Ruby SDK entrambi disponibili per .NET Services.

Google App Engine

Visto lo sviluppo della rete a nube, anche Google offre, ormai da diversi mesi il suo nuovo prodotto cloud: Google Appe Engine (GAE).

GAE è stato studiato come metodo da seguire per creare applicazioni web (HTTP-driven) da ospitare sui web server di Google. A differenza di EC2, GAE scende nel campo delle piattaforme cloud fornendo un framework di tecnlogie rigide, con le quali creare applicazioni senza preoccuparsi della loro architettura e dei picchi di traffico e carico. Google App Engine permette di eseguire le applicazioni web all'interno delle infrastrutture di Google. Il software è realmente open source, infatti nelle pagine dedicate alla cloud di Google si trova tutto, compresi i listati dei sorgenti (Python).

Con App Engine non ci sono server da mantenere: è sufficiente caricare l'applicazione ed è immediatamente pronta per essere utilizzata. È possibile condividere la propria applicazione con il mondo, o limitare l'accesso ai membri della propria organizzazione. Google App Engine supporta software scritti in diversi linguaggi di programmazione.

Con l'ambiente App Engine runtime Java, è possibile creare la propria applicazione utilizzando le tecnologie standard di Java, tra cui la JVM, Java Servlet e il linguaggio di programmazione Java o qualsiasi altro linguaggio utilizzando una JVM-based interprete o compilatore, come JavaScript o Ruby.

App Engine offre inoltre un ambiente di sviluppo runtime Python dedicato, che include un interprete Python veloce e la libreria standard di Python. Gli ambienti runtime Java e Python sono costruiti per garantire che l'applicazione venga eseguita in modo rapido, sicuro e senza interferenze con altre applicazioni sul sistema.

Con App Engine, si paga solo per quello che si utilizza. Non ci sono costi di set-up e spese ricorrenti. Le risorse che l'applicazione utilizza sono tutte monitorate, come ad esempio «storage» e «bandwidth», misurati in gigabyte e vengono fatturati a prezzi competitivi.

Le risorse e i limiti

La prima caratteristica della rete che si riscontra è la mancata scalabilità hardware della rete stessa. Per ogni account, infatti, il Google App Engine associa solo una CPU di tipo mono-core. Attualmente non è prevista la possibilità dell'impiego di architetture multi-core tipiche del cloud.

Un'ulteriore caratteristica che discosta GAE dalle comuni reti cloud è la completa mancanza di virtualizzazione di sistemi operativi di ogni tipo; esso, infatti, permette la sola esecuzione di codice Python, eventualmente associato ad un framework proprietario o ad una versione alleggerita di Django.

Per usufruire della rete GAE e, quindi, iniziare a sviluppare le applicazioni, è necessario scaricare l'ambiente di sviluppo, che ricreerà le stesse condizioni tecniche di GAE sulla propria macchina locale; l'SDK contiene, infatti, un web server, un database, strutture per recuperare indirizzi HTTP(s) e quelle per l'invio di e-mail o per la manipolazione d'immagini.

Un sistema rigido come la Rete GAE comporta anche diversi limiti. Non si tratta, infatti, di un prodotto completamente nuovo: la piattaforma è la stessa alla base di molti attuali servizi di casa Google, come Google Earth, Google Sites o Google Finances e al suo interno integra una serie di applicazioni e funzionalità che Google utilizza già da tempo in tutto il mondo.

I componenti fondamentali

Per poter controllare il consumo effettivo di risorse, il GAE mette a disposizione tutta una serie di servizi che le applicazioni possono sfruttare:

- Il Datastore: database un po' particolare denominato BigTable, formato da una piattaforma distribuita operante sul file system proprietario, GFS. Ha un linguaggio simile a SQL con delle limitazioni a carattere operativo chiamato GQL. Nonostante le limitazioni, comunque, è consentito un uso standard abbastanza vasto del database da parte delle applicazioni.

- Google Accounts: è una API che permette di avere automaticamente un sistema di login per le applicazioni, basato sugli accounts Google.

Potrebbe essere un problema di sicurezza basare l'identificazione dell'utente solo sull'impiego dell'account, ma considerato che non esiste una versione business di GAE, il problema passa in secondo piano.

- URL Fetch: le applicazioni possono accedere all'esterno, recuperando il contenuto di URLs remoti sfruttando API basate sulla stessa infrastruttura che Google usa per altri suoi prodotti. Resta comunque possibile usare le librerie standard di Python se non si vuole usufruire dell'infrastruttura Google (e dei suoi limiti).

- Mail: usata per inviare email con o senza allegati anche verso l'«admins» delle applicazioni.

- Memcache: è uno storage di tipo «in-memory key-value». Permette di inserire in cache strutture, valori, risultati di query complesse e rendere, quindi, il recupero degli stessi più veloce.

- Image Manipulation: permette di ridimensionare, ruotare ed effettuare operazioni basilari su immagini in formato JPEG e PNG. Ogni richiesta HTTP ha 30 secondi di tempo per essere evasa, l'SDK permette anche il deploy delle applicazioni per sviluppare in locale fino al momento dello spostamento sui server di Google, riducendo al minimo il carico della rete.

- Sandbox: è un ambiente sicuro, affidabile, indipendente dall'hardware, dal sistema operativo e dalla ubicazione fisica del sistema di servizio di web in cui vengono conservate le richieste di rete dei vari utenti. Questo metodo garantisce accessi sicuri, contemporanei e multipli alla rete GAE.

Comparazione di GAE e Azure

In questa sezione verranno comparati Google App Engine e Microsoft Azure sulle basi di alcune caratteristiche importanti per l'adozione di politiche Green volte al risparmio energetico, nell'ottica di un Cloud Computing "verde". Tale comparazione pone le basi su alcuni whitepapers e articoli rilasciati dalle stesse compagnie, nonché alcuni blog e siti di informazione non ufficiali.

Microsoft Azure

Microsoft riconosce che ci sia bisogno di educare maggiormente gli utenti su come migliorare l'ambiente attraverso il potenziale del software. Microsoft perciò si impegna sia a pubblicare le linee guida per una pratica migliore e a impegnarsi con partner governativi, non-governativi, industriali e di consumatori per indirizzare in questo verso l'impatto diretto e indiretto dell'industria tecnologica sull'ambiente.

L'approccio di Microsoft nel design dei propri datacenter [64] è di guardare alla facility come se fosse un grande computer in funzione 24 ore su 24, 7 giorni su 7. Come detto precedentemente, i computer lavorano meglio quando sono creati su misura per l'applicazione che devono eseguire. Questo semplice principio viene applicato da Microsoft al design del datacenter; la maggior parte dei design energicamente efficienti vanno incontro alle necessità degli utenti e della specifica locazione del datacenter.

Microsoft continuamente valuta molte differenti tecnologie per la distribuzione dell'elettricità, per i sistemi di raffreddamento e studia soluzioni di server rack e di sistemi di container. Per ottimizzare l'efficienza dei propri datacenter, Microsoft usa tools come Computational Fluid Dynamics per testare le differenti configurazioni.

Per la scelta della locazione del datacenter, Microsoft usa dei tools di software che creano mappe di calore in modo da poter scegliere più efficacemente la locazione ideale per il posizionamento del datacenter. Una volta che la locazione è stata scelta, Microsoft valuta il design dell'edificio e l'equipaggiamento per creare una configurazione efficiente con bassi costi (TCO) in rapporto alla vita dell'edificio. Piuttosto che dividere l'onere della scelta fra diversi team, Microsoft ha creato un unico team designato nella scelta della locazione geografica e nel design dell'edificio.

Per ovviare all'inefficienza dovuta ad un basso carico di lavoro, Microsoft ha implementato un design modulare nel quale solo una parte del datacenter lavora quando si ha basso carico.

Poiché uno dei primi passi verso un aumento dell'efficienza energetica è quello di rendere consapevoli dei costi i team che lavorano sul datacenter, Microsoft ha fatto sì che le metriche di efficienza energetica siano parte delle regolari comunicazioni per i servizi web ed ha sviluppato internamente strumenti per comunicare informazioni sulle operazioni dei datacenter. I team che applicano decisioni sui servizi Web di Microsoft ora ricevono un report delle performance di efficienza energetica in modo da apportare miglioramenti anche su questo frangente.

Microsoft inoltre ha iniziato lo sviluppo di un particolare software, chiamato Joulemeter [64], in grado di misurare quanta energia richiede un certo hardware o particolare software. I sensori hardware possono solo definire la quantità di energia consumata dalla macchina fisica, ma poiché i computer sono multitasking, è interessante conoscere la quantità di energia richiesta da un singolo frammento di software, come per esempio da una macchina virtuale. Questo può portare a una miglior politica di gestione costi per il Cloud Computing: per esempio si possono

far pagare ai clienti solo il consumo reale delle proprie macchine virtuali. Inoltre una migliore visibilità aumenta anche il controllo; per esempio, quando l'energia totale consumata dal datacenter diventa un problema si mettono alcune Virtual Machine a bassa priorità e si preservano le perfomance di altre Virtual Machine a priorità elevata. Sapere qual'è il consumo di ogni Virtual Machine impedisce di fare decisioni al buio.

All'interno dei propri datacenter, Microsoft usa alcune tecniche che gli hanno permesso di migliorare notevolmente il controllo delle temperature e la distribuzione di aria:

- Orientare le unità AC (Air Conditioning) perpendicolarmente alle corsie di aria calda (hot aisles), così da far rimanere l'aria calda nelle corsie calde ed evitare il ricircolo della stessa.

- Uguagliare la richiesta di aria degli equipaggiamenti IT con la disponibilità d'aria delle unità CRAC per diminuire il ricircolo dell'aria e rendere disponibile aria fresca a tutti i server.

- Configurare una pressione dell'aria uniforme, usando il pavimento rialzato di 0.8 — 1.0 m, mattonelle perforate movibili e mattonelle piene vicino alle unità AC.

Al fine di ridurre il ricircolo interno dell'aria all'interno dei proprio datacenter, Microsoft ha adottato alcune pratiche:

- dividere il datacenter in corsie calde e fredde (hot aisles, cold aisels)

- eliminare i vuoti tra le corsie

- disporre corsie più lunghe

- rendere vuoti alcuni pannelli di rack
- sigillare gli interruttori dei cavi

Per quel che riguarda la refrigerazione, Microsoft ha adottato tecniche di refrigeramento sia a liquido che ad aria, diverse in ogni datacenter a seconda della locazione geografica.

Inoltre Microsoft partecipa e condivide le migliori pratiche con The Green Grid, Climat Savers Computing, Environmental Protection Agency, Lawrence Berkeley National Labs, American Society of Heating Refrigeration and Air-Conditioning Engineer e Association for Computer Operations Management. L'iscrizione a queste organizzazione promuove la condivisione delle conoscenze tra le industrie e provvede a uno scambio di informazioni sulle diverse strategie e migliori messe in atto di data center.

Google App Engine

L'illustrazione 18 mostra come Google sia riuscita ad abbassare considerevolmente l'energia richiesta dai propri server e dalle unità di raffreddamento; questo è stato raggiunto diminuendo l'overhead di energia richiesta dalla facility del datacenter. Più specificamente, i datacenter di Google usano circa cinque volte meno l'energia richiesta da facility convenzionali per dare energia ai server e raffreddarli. Un risultato tangibile è, per esempio, che

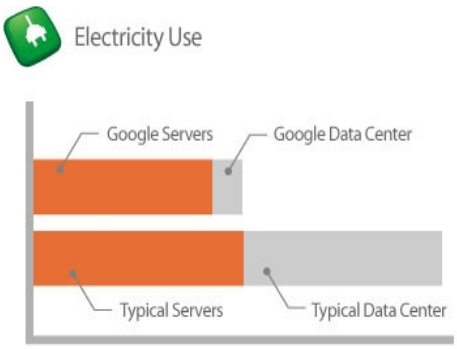


Illustrazione 18

l'uso di energia per fare una ricerca sul motore di ricerca google è minimale: nel tempo che la ricerca viene processata, il proprio personal computer usa più energia di quella che serve a Google per rispondere alla query.

Google attua nei suoi datacenter alcune configurazioni per aumentare l'efficienza energetica degli stessi; alcuni esempi di queste configurazioni sono le seguenti:

- Gestione intelligente del flusso d'aria: l'aria calda in uscita dai server non è fatta mischiare con l'aria fredda e il tragitto fino al refrigeratore è molto corto; in questo modo poca energia è usata per muovere per lunghe distanze l'aria calda o fredda.

- Maggiore temperatura dell'aria in ingresso ai server: nei datacenter Google basati su container, l'aria fredda all'interno dei container ha una temperatura di 27°C piuttosto che 18-20°C. Maggiore è la temperatura, maggiore è l'efficienza nel raffreddamento del datacenter.

- Uso di raffreddamento libero: diverse torri di raffreddamento dissipano il calore evaporando l'acqua, riducendo molto la necessità dell'uso di usare refrigeratori. Nei climi moderati, le torri di raffreddamento possono eliminare la maggior parte del tempo di uso dei refrigeratori. Il datacenter di Google in Belgio ha addirittura eliminato l'uso di refirgeratori, usando raffreddamento libero il 100% del tempo.

- Uso di UPS da 12V DC: ogni server contiene un piccolo UPS, essenzialmente una batteria che è messa a monte dell'alimentatore del server e che è efficiente per il 99.99%. Questi UPS possono eliminare la necessità di usare UPS a livello facility, aumentando l'efficienza di tutta l'infrastruttura da circa 90% al 99%.

Google ha deciso di non usufruire dei metodi di sleep dinamici di server; secondo Google infatti [66] l'uso di queste tecniche non può essere efficacemente usato anche per datacenter che ospitano soluzioni di Cloud Coputing poichè pagherebbero troppo spesso lo scotto di latenza ed energia per riesumare una macchina inattiva.

I nuovi datacenter di Google sono basti su container; è stato svelato [66] che Google ha costruito un datacenter basato su container che è attivo dal 2005 [67]: esso ha raggiunto efficienze energetiche molto alte, per questo Google ha deciso di usufruire delle tecnologie container per la costruzione dei suoi prossimi datacenter.

Google inoltre è attenta anche a rendere i propri datacenter ecologicamente sostenibili. Da diversi anni è alla ricerca di fonti di energia rinnovabile: una società secondaria di Google Inc., la Potter Drilling, ha sviluppato un sistema di turbine (le stesse turbine usate da jet) da usarsi in un sistema geotermico (Enhanced Geothermal Systems — EGS) [68]. Se questo sistema porterà a vantaggi economici, Google potrà costruire i suoi prossimi datacenter vicino a queste fonti geotermiche, in modo da sfruttare direttamente l'energia prodotta.

Inoltre Google è in collaborazione con Microsoft e HP per lo sviluppo di un sistema energetico basato sugli escrementi di animali da bestiame : esso può essere usato per creare biogas (quindi metano) da usare nella generazione di energia elettrica per i propri datacenter. Una mucca «media» può creare abbastanza energia da accendere una lampadina da 100W, perciò 10,000 mucche possono alimentare un piccolo datacenter (1MW). Inoltre, poiché la produzione di metano ha bisogno di molto calore, questo può essere preso gratuitamente dal calore in uscita dal datacenter, quindi può essere creato un circolo vincente dove il

biogas alimenta il datacenter e l'aria calda del datacenter alimenta la creazione di biogas.

6. Conclusioni

Nel mondo IT, il mercato del Cloud Computing è uno tra quelli che hanno avuto il maggior sviluppo negli ultimi anni: sempre più aziende, sia in Italia che nel mondo, si stanno interessando a questo nuovo fenomeno e i grandi colossi dell'informatica (tra cui Google, IBM, Amazon, Apple e Microsoft) stanno sviluppando nuove piattaforme per mettere a disposizione dei propri clienti servizi sempre più efficienti.

Ma come questo settore è in forte espansione, è in aumento anche la produzione indiretta di CO_2. In questa tesi si è valutata l'importanza dell'adozione di pratiche ecologicamente sostenibili, sia per una responsabilità etico-sociale che l'IT deve avere nei confronti della salvaguardia dell'ambiente, sia per un consistente risparmio di denaro. I costi energetici costituiscono un'ampia fetta nella totalità dei costi di un datacenter e sono destinati ad aumentare sempre di più: per questo una gestione efficiente dell'energia all'interno dei datacenter ricoprirà un ruolo sempre più importante. Tuttavia attualmente si è ben lontani dagli obiettivi prefissati: un datacenter nella media spreca i due terzi della propria energia.

In questa tesi si è studiato quanto sia importante adottare politiche e tecniche al fine di aumentare l'efficienza energetica di datacenter Cloud Computing, focalizzandosi sull'impatto ambientale dello stesso e i costi energetici. Inoltre sono state analizzate e approfondite alcune politiche e metodologie per migliorare l'efficienza energetica dei datacenter che ospitano Cloud Computing. Infine sono state comparate le scelte di due grandi aziende del settore (Google e Microsoft), sulla base di alcune caratteristiche importanti per l'adozione di politiche Green volte al risparmio energetico, nell'ottica di un «Green Cloud Computing».

Bibliografia

[1] Murugesan S.: Harnessing Green IT: Principles and Practices. IT Professional, Vol. 10, n. 1, 2008, p. 24-33.

[2] Restorick T.: An Inefficient Truth. Global Action Plan Report, 2007. www.globalactionplan.org.uk/research.aspx

[3] www.berr.gov.uk, IEA

[4] Renzi F.: L'innovazione che fa la differenza - la strategia IBM e la tecnologia a supporto della flessibilità d'impresa e dei risparmi energetici. ForumPA 2008.

[5] Microsoft to Google: My PUE is Getting Better Than Your PUE - http://www.treehugger.com/files/2008/10/microsoft-to-google-my-pue-is-getting-better-than-your-pue.php

[6] http://www.gartner.com

[7] Kumar R.: Important Power, Cooling and Green IT Concerns. Gartner report, Gennaio 2007.

[8] http://www.information-age.com/channels/development-and-Integration/news/1147613/green-it-a-convenient-truth-for-both-vendors-and-enterprise.thtml

[9] Energy Saving Trust press release, Ken Livingston's Energy Strategy Welcomed

[10] The New Scientist

[11] www.berr.gov.uk, IEA

[12] Gartner, 2007 Press Release

[13] IPCC Fourth Assessment Report, 16 November 2007

[14] Richard Barrington, head of Public Policy for Sun UK and Ireland and UK government advisor

[15] Restorick T.: An Inefficient Truth. Global Action Plan Report, 2007, www.globalactionplan.org.uk/research.aspx

[16] Kott Benjamin: Head Green Business Operations Europe, Middle East & Africa, Google: Tackling energy use at Google's data centres. OECD Conference ICTs, the environment and climate change, Helsingor, 2009.

[17] Forrester "Information Fabric: Enterprise Fabric Virtualization" 2006

[18] Report to Congress on Server and Data Center Energy Efficiency - Public Law 109-431 U.S. Environmental Protection Agency, August 2, 2007.

[19] UCSU Environmental Center: Green computer guide. http://ecenter.colorado.edu/energy/projects/green_computing.html

[20] Il Green IT e la sfida della sostenibilità, Eugenio Capra: http://www.economia.uniroma2.it/Public/files/eprocurement/file/CapraGreenIConsip23giu.ppt

[21] www.computerworlduk.com/TXPO `Gartner use intelligent IT to green the business, IT chiefs told!'

[22] Carbon Neutral Company

[23] Rakesh Kumar, Gartner Analyst, Sep 2006

[24] Report to Congress on Server and Data Center Efficiency, US Environmental Protection Agency

[25] Murugesan S.: Harnessing Green IT: Principles and Practices. IT Professional, Vol. 10, n. 1, 2008, p. 24-33.

[26] Hilty Lorenz M.: Environmental impact of ICT- A conceptual framework and some strategic recommendations. OECD workshop on ICT and environmental challenges, Copenhagen , 20-22 May 2008. http://www.oecd.org/dataoecd/42/13/40833380.pdf

[27] Mattern F.: Die technische Basis für das Internet der Dinge. In: Fleisch E., Mattern F. (Hrsg.): Das Internet der Dinge. Springer, 2005, p. 39-66.

[28] Noah, Timothy (Nov. 9, 2006). The GOP Triangulates. Slate.

[29] Legge carbon tax italiana - http://www.camera.it/parlam/leggi/98448l.htm

[30] The Market Transformation Programme

[31] M. Armbrust, A. Fox, R. Griffith, "Above the Clouds: A Berkeley View of Cloud Computing", 2009, http://radlab.cs.berkeley.edu

[32] Dean J., Hemawat S., Mapreduce S.: «Simplified data processing on large clusters». In OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation (Berkeley, CA, USA, 2004), USENIX Association, pp. 1010

[33] Hemawat S., Gobioff S. Leung S.: «The google file system». In SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles (New York, NY, USA, 2003), ACM, pp. 2943: http://portal.acm.org/ft_gateway.cfm?id=945450&type=pdf&coll=Portal&dl=GUIDE&CFID=19219697&CFTOKEN=50259492

[34] Chang F., Dean J., Hemawat S., Sieh H., Wallach D. Burrows M.: «Bigtable: A distributed storage system for structured data». In Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06) (2006)

[35] Decandia G., Hastorun D.: «Dynamo: Amazon's highly available key-value store». In Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles (2007), ACM Press New York, NY, USA, pp. 205220.

[36] Hamilton J.: «Cost of Power in Large-Scale Data Centers». November 2008: http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx

[37] J. Gozdechi, A. Jajszczyk, R. Stankiewicz, "Quality of Service Terminology in IP Networks", Marzo 2003, in IEEE Communications Magazine

[38] K. Keahey, M. Tsugawa, A. Matsunaga, J. A.B. Fortes, "Sky Computing", Settembre/Ottobre 2009, in IEEE Internet Computing

[39] G. Lodi, F. Panzieri, D. Rossi, E. Turrini, "SLA-Driven Clustering of QoS-aware Application Servers", Marzo 2007, in IEEE Transactions on Software Engineering

[40] J. Skene, D. Lamanna, W. Emmerich, "Precise Service Level Agreements", 2004 in Proceedings of the 26th International Conference on Software Engineering

[41] S. Ghemawat, H. Gobioff, and S-T. Leung, "The Google file system", in Proceedings of the 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October 2003 doi:10.1145/945445.945450

[42] U.S. Environmental Protection Agency, "Report to Congress on server and datacenter energy efficiency," Public Law 109-431, August 2, 2007

[43] M. K. Patterson and D. Fenwick, "The state of datacenter cooling," Intel Corporation White Paper. Available at http://download.intel.com/technology/eep/data-center-efficiency/state-of-date-center-cooling.pdf

[44] http://it.wikipedia.org/wiki/Glicol_etilenico

[45] Bechtolsheim: «A. Cloud Computing and Cloud Networking». talk at UC Berkeley, December 2008.

[46] Sissa G.: «Green Software» — Mondo Digitale, settembre 2009

[47] Clausen J., Fichter K., Hintemann R.: How to make computers green? Ressource-efficient innovation- sin schools. Einblicke n. 49/Fruejahr 2009 - Carl Von Ossietzky Universitaet, Oldenburg, p. 48-51.

[48] M. Kalyanakrishnam, Z. Kalbarczyk, and R. Iyer, "Failure data analysis of a LAN of Windows NT based computers," Reliable Distributed Systems, IEEE Symposium on, vol. 0, no. 0, pp. 178, 18th IEEE Symposium on Reliable Distributed Systems, 1999. doi:10.1109/RELDIS.1999.805094

[49] C. Patel et al., "Thermal considerations in cooling large scale high compute density datacenters": http://www.flomerics.com/flotherm/technical_papers/t299.pdf

[50] D. Nelson, M. Ryan, S. DeVito, K. V. Ramesh, P. Vlasaty, B. Rucker, B. Da y Nelson, et al., "The role of modularity in datacenter design". Sun BluePrints Online, http://www.sun.com/storagetek/docs/EED.pdf

[51] PG&E, "High performance datacenters". Available at http://hightech.lbl.gov/documents/DATA_CENTERS/06_DataCenters-PGE.pdf

[52] Green Grid, "Seven strategies to improve datacenter cooling efficiency". Available at http://www.thegreengrid.org/gg_content/

[53] C. D. Patel, C. E. Bash, R. Sharma, and M. Beitelmal. Smart Cooling of Data Centers. In Proceedings of the Pacific RIM/ASME International Electronics Packaging Technical Conference and Exhibition (IPACK03), July 2003

[54] Justin Moore, Jeff Chase , Parthasarathy Ranganathan , Ratnesh Sharma : Making Scheduling “Cool”: Temperature-Aware Workload Placement in Data Centers

[55] R. F. Sullivan. Alternating Cold and Hot Aisles Provides More Reliable Cooling for Server Farms. In Uptime Institute, 2000.

[56] D. Anderson, J. Dykes, and E. Riedel. More Than an Interface—SCSI vs. ATA. In Proceedings of the 2nd Usenix Conference on File and Storage Technologies (FAST), San Francisco, CA, March 2003.

[57] G. Cole. Estimating Drive Reliability in Desktop Computers and Consumer Electronics. In Technology Paper TP-338.1, Seagate Technology, November 2000

[58] Moore, Chase, Ranganathan, Sharma: Making scheduling «cool»: temperature-aware workload placement in data centers

[59] Heath, Centeno, George, Ramos, Jaluria, Bianchini: Mercury and Freon: tempearture emultaion and management for server systems

[60] W. Zhang: Linux Virtual Server for scalable network services, July 2000

[61] Chase, Anderson, Thakar, Vahdat: Managing Energy and Server Resources in Hosting Centers

[62] S. Sankar, S. Gurumurthi, and M. R. Stan, "Intra-disk parallelism: an idea whose time has come," in Proceedings of the ACM International Symposium on Computer Architecture, June 2008, pp. 303314.

[63] Microsoft's Professional Developers Conference

[64] http://download.microsoft.com/download/8/7/D/87D2D871-471E-44A3-BFD0-C02D3248B8CB/Energy%20Efficiency%20Best%20Practices%20in%20Microsoft%20Data%20Center%20Operations%20CeBIT.pdf

[65] http://gigaom.com/2010/04/25/green-software-qa-microsoft-research-joulemeter/

[66] «The Datacenter as a Computer»: An Introduction to the Design of Warehouse-Scale Machines - Luiz André Barroso and Urs Hölzle

[67] Google Inc., "Efficient Data Center Summit, April 2009": http://www.google.com/corporate/green/datacenters/summit.html

[68] Google Funds Ultra Efficient Jet Engine-Inspired Geothermal Drill by Ariel Schwartz: http://inhabitat.com/2010/05/19/google-funds-ultra-efficient-jet-engine-inspired-geothermal-drill/

[69] «Google, HP and Microsoft Consider Poo to Power Data Centers » by Yuka Yoneda: http://inhabitat.com/2010/05/20/google-hp-and-microsoft-consider-poo-to-power-data-centers/

Ringraziamenti

Ringrazio innanzitutto la mia famiglia che mi ha dato la possibilità di affrontare questi studi, incoraggiandomi e sostenendomi.

Ringrazio Elena per avermi sopportato, aiutato e incentivato durante questi anni.

Ringrazio il Professor Fabio Panzieri per avermi seguito con interesse durante la stesura di questa tesi.

Ringrazio il Dott. Mauro Minella per avermi procurato del materiale indispensabile per la stesura della tesi.

Ringrazio tutti i miei colleghi di facoltà che in questi anni mi hanno supportato ed aiutato negli studi; in particolare ringrazio Gianpietro, Davide, Pietro, Michael, Pasquale, Francesco, Rosa.

Ringrazio i miei amici Francesco e Luca che in questi anni mi hanno permesso di capire quanto sia importante il rispetto della natura e dell'ambiente.