Diapositiva 1: Titolo
Questa presentazione si basa su uno studio approfondito che ho condotto riguardo i benefici dell'Edge Computing nelle applicazioni IoT dove la latenza è il principale ostacolo per il funzionamento ottimale della rete. Lo studio analizza sia gli aspetti teorici che quelli pratici, includendo casi studio concreti.

Diapositiva 2: Indice
Questa presentazione si svilupperà in sei sezioni principali:

- Introduzione all'Internet of Things (IoT)
- I limiti del Cloud Computing tradizionale
- La nascita dell'Edge Computing
- Alcuni casi studio che provano l'efficacia di questo nuovo paradigma di calcolo
- Le sfide aperte nella ricerca

L'obiettivo è fornire una panoramica chiara del perché l'Edge Computing stia diventando una tecnologia fondamentale per il futuro delle applicazioni IoT a bassa latenza.

Diapositiva 3: IoT data flow
L'IoT rappresenta una delle più grandi rivoluzioni degli ultimi anni. Miliardi di dispositivi connessi tra loro: sensori industriali, wearable, sistemi di sorveglianza, veicoli autonomi, elettrodomestici smart. Si prevede che entro il 2025 avremo più di 75 miliardi di dispositivi IoT attivi.
Questi dispositivi interconnessi possono assumere il ruolo di sensori, capaci di percepire l'ambiente circostante ed effettuare misurazioni, o attuatori, in grado di agire fisicamente sull'ambiente. I dati generati dai sensori vengono aggregati tramite dispositivi appositi chiamati "gateway", posizionati vicino ai sensori e attuatori. I vari gateway si occupano poi di instradare tali dati al "cloud" che si occuperà dell'elaborazione degli stessi. La risposta post processo del cloud viene ritornata agli attuatori che agiscono sull'ambiente conseguentemente.
Come si può notare dallo schema proposto, un sistema IoT può essere suddiviso in due zone principali, separate dal cosiddetto "bordo" della rete: la parte caratterizzata dai diversi sensori e attuatori collegati tra di loro e a loro volta ai diversi gateway per la raccolta e l'invio dei dati (la rete stessa); la parte remota caratterizzata dai data center per l'elaborazione dei dati raccolti.
Questo scenario, tuttavia, presenta un problema: tutti questi dispositivi generano una mole enorme di dati, che, per alcuni casi d'uso, devono essere elaborati in tempo reale per essere utili. Questo ci porta al primo grande limite dell'infrastruttura cloud tradizionale.

Diapositiva 4: Limiti del Cloud Computing
Il Cloud Computing è stato per anni la soluzione principale per l'elaborazione dei dati, ma oggi sta mostrando i suoi limiti, specialmente in applicazioni dove i requisiti di elaborazione real-time e la bassa latenza sono fondamentali per garantire il corretto funzionamento del sistema.

- **Vincoli di banda**: l'invio di enormi quantità di dati ai server cloud può causare congestione della rete e relativo rallentamento del traffico.
- **Latenza**: la distanza fisica tra dispositivi e data center introduce ritardi inaccettabili per applicazioni critiche come i veicoli autonomi o la telemedicina.
- **Inefficienza energetica**: i dispositivi IoT spesso hanno batterie limitate e non possono permettersi di trasmettere continuamente l'enorme mole di dati raccolti a data center remoti.
- **Sicurezza e privacy**: i dati sensibili trasmessi ai server cloud possono essere intercettati o compromessi.

Tutte queste problematiche hanno portato alla nascita dell'Edge Computing.

Diapositiva 5: L'emergere dell'Edge Computing
Edge Computing consente di spostare l'elaborazione dei dati più vicino alla loro sorgente, riducendo la dipendenza dai server centralizzati. Per esempio, invece di inviare continuamente dati da una telecamera di sorveglianza al cloud per l'analisi, l'Edge Computing permette di elaborare i dati direttamente nel dispositivo o nel nodo di rete più vicino.
Questo nuovo approccio all'elaborazione dei dati presenta diversi punti di forza, tra cui:

- **Latenza ridotta**, perché i dati non devono viaggiare fino al cloud e ritorno. La latenza, infatti, è fortemente influenzata dalla distanza fisica tra il nodo della rete e il data center in cui i dati vengono processati, riducendo questa distanza è possibile ridurre notevolmente la latenza.
- **Risparmio di banda**, grazie ad una computazione "locale" è possibile ridurre notevolmente la mole di dati da inoltrare ad un cloud remoto, riducendo quindi il carico e la congestione della rete stessa.
- **Maggiore sicurezza**, poiché i dati sensibili possono essere elaborati localmente.
- **Migliore efficienza energetica**, riducendo il consumo di batteria dei dispositivi.

L'Edge Computing trova la sua applicazione in tutte quei casi d'uso dove la minima latenza è fondamentale per garantire il corretto funzionamento della rete IoT. Tra queste applicazioni troviamo: sistemi healthcare, video sorveglianza, monitoraggio di situazioni ambientali critiche.

Diapositiva 6: Mobile Edge Computing (MEC) - architettura
In questa diapositiva vediamo rappresentata l'architettura di un sistema Mobile Edge Computing (MEC), ideata principalmente a supporto delle reti mobili. Il tradizionale cloud computing viene scomposto su due livelli:

- **Livello 1 (Layer 1)**: Al livello inferiore troviamo i server MEC, che sono posizionati vicino alle stazioni base delle celle mobili per ridurre la latenza e aumentare l'efficienza. Questi server elaborano i dati localmente, evitando la necessità di inviarli al cloud centrale, il che porta notevoli vantaggi, come una maggiore velocità di risposta e un minor carico sulla rete.
- **Livello 2 (Layer 2)**: Il sistema include anche un server di backup centrale e risorse computazionali per supportare attività più intensive che non possono essere gestite direttamente dai server MEC.

La parte inferiore dello schema rappresenta la rete di base, composta da dispositivi mobili come smartphone, tablet, computer portatili e altri dispositivi IoT. Questi dispositivi si connettono a stazioni base che fungono da ponte tra i dispositivi e i server MEC. Durante la loro mobilità, i dispositivi possono scegliere di delegare la computazione al MEC server più conveniente tenendo in considerazione gli eventuali costi da sostenere in termini di computazione richiesta, tempo di esecuzione e grandezza del task. Dall'altro lato anche i server MEC cercano di bilanciare il carico delle richieste e le risorse disponibili. Questi due problemi di ottimizzazione sono stati risolti dagli autori attraverso un approccio appartenente alla teoria del gioco che, unito alla struttura del cloud a due livelli, si è rivelata molto promettente, anche in situazioni di notevole carico. La gestione della mobilità rende questo approccio particolarmente promettente per gestire reti di veicoli autonomi o altre reti mobili.

Diapositiva 7: Mobile Edge Computing (MEC) - risultati
Questa slide evidenzia i risultati ottenuti tramite l'architettura MEC, in termini di consumo energetico e la riduzione della latenza media durante l'esecuzione di alcuni task da parte dei dispositivi coinvolti.
Il primo grafico si concentra sullo studio del consumo energetico analizzando diverse strategie di offloading del calcolo, in particolare ne sono state considerate quattro:

- Nessun offloading: computazione gestita dal dispositivo.

- Graph-matching-based offloading: strategia alternativa a quella trattata nel paper dove i dispositivi e i server MEC vengono trattati come nodi di un grafo e l'obiettivo è quello di calcolare il percorso minimo tra un nodo di rete e il server MEC per ottimizzare i costi e ridurre la latenza.
- La strategia di offloading proposta con e senza supporto del server di backup centrale.

Il grafico mette a paragone il consumo energetico con il numero di dispositivi che formano la rete. Attraverso il grafico si può notare come la strategia proposta con server di backup (arancione) ottiene i migliori risultati, con il consumo energetico più basso, grazie a un'elaborazione efficiente e al supporto del server di backup. Al contrario, l'assenza di offloading (verde) risulta essere la più dispendiosa in termini energetici, poiché tutti i compiti vengono gestiti dai dispositivi stessi senza aiuto dai server MEC. In particolare, la strategia basata sul graph-matching-based offloading risulta essere meno efficace della soluzione proposta dagli autori poiché tale soluzione considera le risorse di calcolo e di comunicazione disponibili solamente nella locazione corrente dei dispositivi, senza considerare gli eventuali server MEC fuori raggio ma raggiungibili a mano a mano che il dispositivo si muove.
Questi risultati rendono il modello MEC, specialmente quando supportato da un server di backup, estremamente efficace nel ridurre il consumo di energia, rendendolo una soluzione ideale per applicazioni IoT ad alta intensità.

Il secondo grafico analizza invece come la velocità dei dispositivi mobili influisca sulla riduzione della latenza media. Qui si confrontano tre scenari: 100, 150 e 200 dispositivi connessi.
La riduzione della latenza media risulta più significativa a velocità più elevate, poiché i dispositivi possono accedere a un numero maggiore di server MEC durante il trasferimento dei task, migliorando l'efficienza dell'offloading. All'aumentare della velocità, il sistema sfrutta meglio le risorse disponibili, ottimizzando il bilanciamento del carico e riducendo i tempi di esecuzione.
Il sistema MEC gestisce meglio la latenza quando il numero di dispositivi è minore (100 dispositivi rispetto a 200), evidenziando la necessità di bilanciare il carico.

Questi risultati dimostrano chiaramente che l'architettura MEC proposta non solo riduce il consumo energetico, ma migliora anche la gestione della latenza, anche in condizioni più o meno critiche di mobilità. Questi vantaggi la rendono una soluzione ideale per supportare applicazioni IoT ad alta densità e in contesti dinamici, come veicoli connessi o smart cities.

Diapositiva 8: Studio sul Gaming Mobile
L'obiettivo di questo studio è stato quello di valutare l'impatto dell'Edge Computing sulla latenza in applicazioni di mobile gaming che richiedono un uso intensivo di risorse. Le reti di accesso considerate sono state Wi-Fi e LTE, mentre il gioco di riferimento, Neverball, è stato scelto per rappresentare applicazioni con ambienti complessi 3D.

Lo studio si è concentrato sulla response delay, ossia la misurazione sul tempo trascorso tra un'interazione del client e il corrispettivo risultato ritornato al client stesso. Questa metrica si compone di tre elementi fondamentali:

- Processing Delay (PD): il tempo di elaborazione lato server e rendering del corrispettivo frame.
- Network Delay (ND): il RTT tra client e server.
- Playout Delay (OD): il tempo necessario al client per decodificare e visualizzare i dati a video.

Gli autori hanno considerato tre strategie di implementazione: edge computing implementato attraverso server dedicati direttamente sulle base stations, un cloud remoto adibito al gaming e un cloud remoto commerciale.

Diapositiva 9: Studio sul Gaming Mobile
Lo studio ha messo in evidenza come l'Edge Computing offra vantaggi significativi rispetto al cloud centralizzato, sia specifico per il gaming che non:

- **Latenza di rete (ND)**: Lo scenario Edge, con il server posizionato presso la base station LTE, ha ottenuto una latenza inferiore ai 20 ms, superando di gran lunga i 50 ms registrati con le tradizionali infrastrutture cloud.
- **Tecnologie di virtualizzazione**: I container hanno mostrato prestazioni vicine a quelle di configurazioni bare-metal, mentre la virtualizzazione basata su hypervisor ha introdotto un overhead del 30% in più sul PD.
- **Risoluzione**: Per applicazioni caratterizzate da interazioni continue, ritardi oltre i 70 ms non sono accettabili. Gli autori, tramite l'edge setup, sono stati in grado di garantire ritardi accettabili in risoluzione HD, cosa non possibile con il tradizionale cloud centralizzato.

Lo studio dimostra che la vicinanza delle risorse computazionali agli utenti finali, grazie all'Edge Computing, è cruciale per fornire esperienze di gioco interattive e a bassa latenza. Per applicazioni ad alta intensità di interazione come il gaming, che richiedono tempi di risposta molto bassi, il modello Edge si pone come l'unica soluzione in grado di soddisfare tali requisiti.

(Tecniche di virtualizzazione)

- **Bare-Metal**: Esegue direttamente sul sistema operativo host senza alcuna tecnologia di virtualizzazione, fornendo accesso diretto alle risorse hardware.
- **Container-Based**: Utilizza istanze virtualizzate leggere che condividono le risorse del sistema operativo host senza la necessità di un sistema operativo separato per ogni istanza.
- **Hypervisor-Based**: Utilizza un livello software (hypervisor) per creare e gestire macchine virtuali, ognuna delle quali esegue il proprio sistema operativo guest sopra il sistema operativo host.

Diapositiva 10: Studio sull'Industria Manifatturiera
L'implementazione architetturale dell'edge computing in ambienti industriali IoT si basa su quattro livelli fondamentali:

- **Dispositivi**: Qui troviamo sensori, robot, strumenti di misura e macchinari interconnessi tramite protocolli industriali specifici come OPC UA e DDS, impiegati per raggiungere performance real-time. Per far fronte alle esigenze real-time qui i dispositivi possono implementare forme piuttosto basilari di computazione locale.
- **Network domain**: Questo livello collega i dispositivi ai sistemi di edge computing tramite l'implementazione di protocolli appartenenti all'insieme di standard noto come Time-Sensitive Networking (TSN). Protocolli come PTP assicurano la sincronizzazione tra dispositivi, supportando quindi i requisiti hard real-time dell'industria.
- **Data domain**: Rappresenta l'implementazione dell'Edge computing. Qui avviene la pulizia e l'estrazione delle caratteristiche direttamente alla fonte, riducendo la necessità di trasmettere dati grezzi al cloud. Il sistema può reagire in tempo reale agli eventi di produzione, migliorando l'efficienza e la qualità operativa.
- **Application domain**: Questo livello coordina i processi produttivi e integra tecnologie per la gestione dinamica delle apparecchiature, consentendo un sistema flessibile e interoperabile per l'industria 4.0.

Diapositiva 11: Studio sull'Industria Manifatturiera

L'obiettivo di questo studio era quello di evidenziare i benefici nelle implementazioni di edge computing nell'industria nel supportare le performance real-time. L'edge computing apporta numerosi benefici nel settore manifatturiero, tra cui:

- **Manutenzione attiva**: Un caso di studio su una linea di confezionamento di caramelle ha dimostrato un miglioramento significativo nell'efficienza produttiva con l'adozione dell'edge computing. Con un sistema di gestione automatizzata delle attività, la produzione ha mostrato una riduzione del traffico di rete del 60%, passando da 16-17 Mb/s a 5-6 Mb/s.
- **Cooperazione tra edge e cloud**: I due livelli lavorano in sinergia. L'edge gestisce l'elaborazione in tempo reale, la sicurezza dei dati e l'esecuzione della logica operativa. Il cloud si occupa di analisi dei big data, pianificazione della manutenzione e supporto alle decisioni a lungo termine.

Diapositiva 12: Sfide Aperte

Nonostante l'Edge Computing abbia dimostrato il suo potenziale, ci sono ancora delle sfide aperte che necessitano di attenzione e ricerca approfondita. Analizziamole una per una:

- **Eterogeneità**: L'ecosistema IoT è caratterizzato da una grande varietà di dispositivi con capacità diverse. È quindi fondamentale sviluppare modelli di programmazione standardizzati che possano supportare questa diversità in modo efficiente.
- **Gestione delle risorse**: L'allocazione delle risorse in ambienti dinamici e spesso limitati rappresenta una delle sfide principali. Ottimizzare l'utilizzo delle risorse e della rete in contesti mutevoli è essenziale per garantire prestazioni elevate.
- **Sicurezza e Privacy**: Proteggere i dati sensibili dagli attacchi e dalle minacce in continua evoluzione è cruciale. Gli ambienti edge, essendo distribuiti, sono particolarmente vulnerabili, richiedendo strategie innovative di sicurezza e crittografia.
- **Gestione dei dati**: Il volume di dati generato dalle applicazioni IoT è enorme. È fondamentale, pertanto, modellare attività di preprocessing e filtraggio che riducano il carico complessivo.
- **Affidabilità del sistema**: Infine, garantire che i servizi edge siano affidabili e scalabili è una priorità. Questo richiede soluzioni robuste in grado di mantenere un'erogazione del servizio costante anche in condizioni di carico elevato.

L'edge computing risulta essere un paradigma fondamentale per le applicazioni IoT particolarmente sensibili alla latenza. Risulta pertanto necessario uno studio continuo sul come migliorare questo paradigma.