

# Network Slicing

Il Network Slicing è un concetto fondamentale nelle reti 5G e oltre, che permette la creazione di molteplici reti logiche virtuali su un'unica infrastruttura fisica condivisa, ognuna personalizzata per soddisfare specifici requisiti di servizio e applicazione.

## 1. Introduzione e Sfide nelle Reti Attuali

Le attuali reti basate su IP, come Internet, offrono un servizio “best-effort” senza garanzie sulla tempistica o sulla consegna dei dati, limitando la Qualità del Servizio (QoS) end-to-end. Nonostante esistano soluzioni QoS come IntServ e DiffServ, la loro applicazione su larga scala è complessa e spesso limitata a regioni di rete isolate.

Le applicazioni e i servizi sono in continua evoluzione, con requisiti molto diversi:

- Massive Machine Type Communications (mMTC): Grande numero di dispositivi con basso costo e lunga durata della batteria.
- Ultra-Reliable Low Latency Communications (URLLC): Bassa latenza e ultra-affidabilità simultanee.
- Enhanced Mobile Broadband (eMBB): Velocità di trasmissione più elevate e ampie aree di copertura.

Un'unica architettura di rete non può supportare simultaneamente tutti questi requisiti diversificati. Il Network Slicing emerge come una soluzione promettente per superare queste sfide, consentendo alle reti 5G di offrire una vasta gamma di servizi.

## 2. Definizione e Concetto di Network Slicing

Il Network Slicing implica la suddivisione delle risorse dell'infrastruttura di rete fisica per creare molteplici sottoreti indipendenti per diversi tipi di servizi e applicazioni. Ogni sottorete esegue lo slicing delle risorse fisiche per creare una rete indipendente per le sue applicazioni. È un'architettura di rete virtuale che permette la creazione di molteplici reti logiche atopiche di un'infrastruttura fisica comune, dove ogni “slice” è adattata ai requisiti specifici di servizi diversi. Il concetto di softwarizzazione della rete (Network Softwarization), tramite tecnologie come Network Function Virtualization (NFV) e Software Defined Networking (SDN), facilita la creazione di ambienti di rete flessibili e dinamici.

## 3. Visioni Architetturali del Network Slicing

Diverse alleanze e progetti hanno proposto architetture a strati per il Network Slicing:

- Architettura 5G-PPP (Public Private Partnership Project):
  - Service Layer: Definisce e gestisce i servizi offerti.
  - Business Function Layer: Si occupa degli aspetti commerciali (fatturazione, SLA).

- Orchestration Layer: Gestisce la coordinazione e l'automazione delle funzioni di rete.
- Network Function Layer: Gestisce la configurazione e il ciclo di vita delle funzioni di rete.
- Infrastructure Layer: Comprende l'infrastruttura fisica (core network e RAN).
- Architettura NGMN (Next Generation Mobile Network) Alliance:
  - Business Application Layer: Include applicazioni e servizi.
  - Business Enablement Layer: Fornisce strumenti per supportare le applicazioni (orchestration, policy management).
  - Infrastructure Resource Layer: Comprende risorse fisiche e virtuali.
- Framework Generico per il Network Slicing:
  - Service Layer: Definisce i servizi e le applicazioni.
  - Network Function Layer: Gestisce la configurazione e il ciclo di vita delle funzioni di rete, concatenando diverse funzioni per servizi end-to-end.
  - Infrastructure Layer: Si occupa dell'infrastruttura fisica (core network, RAN) e alloca le risorse.
  - Management and Orchestration (MANO) Entity: Responsabile della traduzione dei modelli di servizio in slice di rete.

#### 4. Benefici del Network Slicing

Il Network Slicing offre numerosi vantaggi:

- Flessibilità Operativa: Permette il funzionamento flessibile di diversi tipi di servizi sulla stessa infrastruttura fisica.
- Reti Logiche Indipendenti: Ogni slice è progettata per soddisfare esigenze specifiche e opera in modo indipendente, garantendo che la performance di una slice non influenzi le altre.
- Creazione On-Demand: Le slice possono essere create dinamicamente, consentendo una rapida implementazione di nuovi servizi e applicazioni.
- Isolamento: Garantisce performance e sicurezza tra le diverse slice e impedisce agli utenti di una slice di accedere o modificare altre slice.
- Elasticità: Permette la modifica dinamica delle risorse allocate alle slice, tramite la rilocalizzazione di funzioni di rete virtuali, lo scaling delle risorse e la riprogrammazione.
- Personalizzazione End-to-End: Assicura un efficiente utilizzo delle risorse condivise.

#### 5. Evoluzione e Concetti Chiave

Il concetto di Network Slicing ha radici nelle Overlay Networks degli anni '80 (primo esempio di slicing) e si è evoluto attraverso VLAN e VXLAN. Le moderne implementazioni sono rese possibili dalla virtualizzazione della rete e dall'astrazione delle risorse.

## 6. Tecnologie Abilitanti (Enablers)

Il Network Slicing è reso possibile da diverse tecnologie complementari:

- **Software Defined Networking (SDN):** Fornisce un controllo centralizzato delle risorse di rete, consentendo una configurazione dinamica e flessibile delle slice e garantendo l'isolamento. L'SDN controller virtualizza e astrae le risorse di rete, provvede a interfacce per applicazioni per usare le risorse di rete e astrae la topologia e le risorse attraverso il Southbound Interface.
- **Network Function Virtualization (NFV):** Permette il disaccoppiamento delle funzioni di rete dall'hardware, consentendo il loro dispiegamento come istanze virtualizzate e ottimizzando l'utilizzo delle risorse. Queste VNF sono facilmente scalabili. Si ha un miglioramento nei costi, nelle performance e nella flessibilità.
- **Cloud Computing:** Offre risorse di calcolo scalabili e on-demand, supportando la gestione centralizzata e lo scaling dinamico delle slice.
- **Edge Computing:** Permette l'elaborazione dei dati al bordo della rete, riducendo la latenza e migliorando le prestazioni in tempo reale, critiche per slice con requisiti stringenti.

## 7. Terminologia

- **Tenant:** Utenti o organizzazioni che hanno accesso a risorse di rete condivise con privilegi e diritti di accesso specifici.
- **Infrastructure Providers:** Operatori o entità responsabili della fornitura e manutenzione dell'infrastruttura di rete fisica.

## 8. Principi del Network Slicing

- **Isolamento della Slice:** Garantisce che ogni slice sia indipendente e che le sue prestazioni e sicurezza non siano influenzate dalle altre.
- **Elasticità:** Permette l'alterazione dinamica delle risorse assegnate a una slice.
- **Personalizzazione End-to-End:** Assicura un efficace utilizzo delle risorse condivise.

## 9. Allocazione delle Risorse

Le risorse vengono allocate dinamicamente a ciascuna slice, inclusi:

- **Banda:** Ogni slice ha una frazione di banda dedicata su un link.
- **Topologia:** Ogni slice ha la sua vista dei nodi di rete (switch, router) e della loro connettività.
- **CPU del Dispositivo:** Risorse computazionali adeguate sono assegnate a ciascuna slice.
- **Archiviazione:** Diversi livelli di capacità di archiviazione sono allocati.
- **Tabelle di Inoltro e Risorse del Piano di Controllo:** Vengono anch'esse "sliciate". Nel senso che ognuno ha le sue tabelle e le sue risorse del piano di controllo.
- **Traffico:** Specifiche porzioni di traffico sono associate alle slice per l'isolamento dalla rete sottostante.

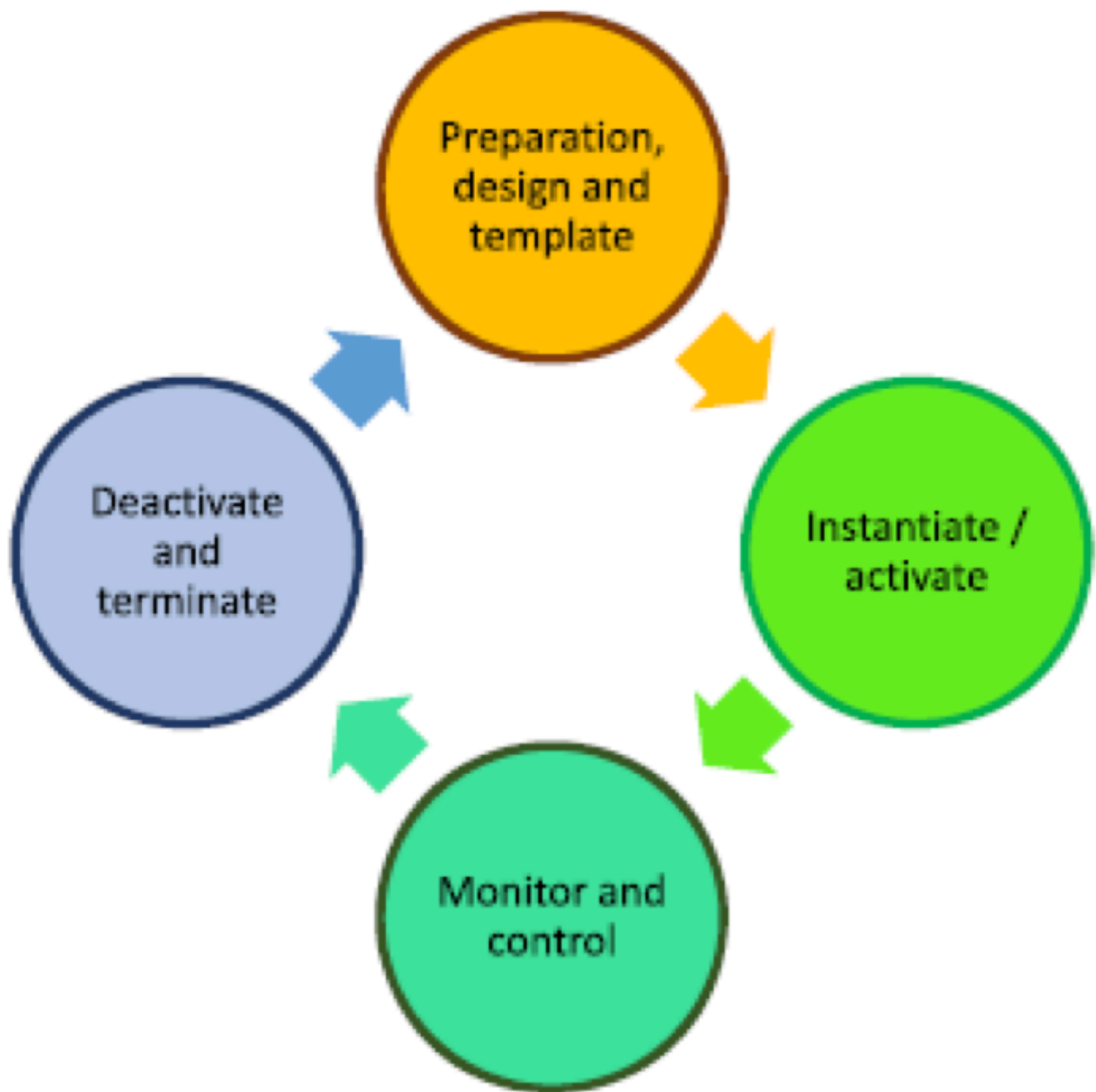
## 10. Creazione Dinamica e Ciclo di Vita delle Slice

Le slice di rete sono reti logiche end-to-end che possono essere create dinamicamente. Gli utenti possono accedere a più slice sulla stessa infrastruttura condivisa in base alle esigenze di servizio. Ci sono tre tipi di livelli:

- Service Instance Layer: questo livello ospita i servizi e le applicazioni che mette a disposizione dell'utente.
- Network Slice Instance Layer: dove un Network Slice Instance (NSI) è una collezione di risorse dal Resource Layer che forma una slice di rete.
- Resource Layer: ospita differenti istanze di sottorete, ogni istanza di sottorete rappresenta una risorsa di rete, di computazione o di memoria. Un NSI può essere composto da uno o più Network Slice Subnet Instances (NSSI), che a loro volta contengono funzioni di rete (VNF o PNF). Un NSI è composto da più NSSI e un NSSI può essere condiviso da più NSI, dove ogni NSSI consiste in NF - VNF/PNF.

Il Network Slice Controller agisce come un orchestratore, gestendo le richieste di slice e coordinando le funzionalità dei diversi strati per creare e riconfigurare le slice durante il loro ciclo di vita.

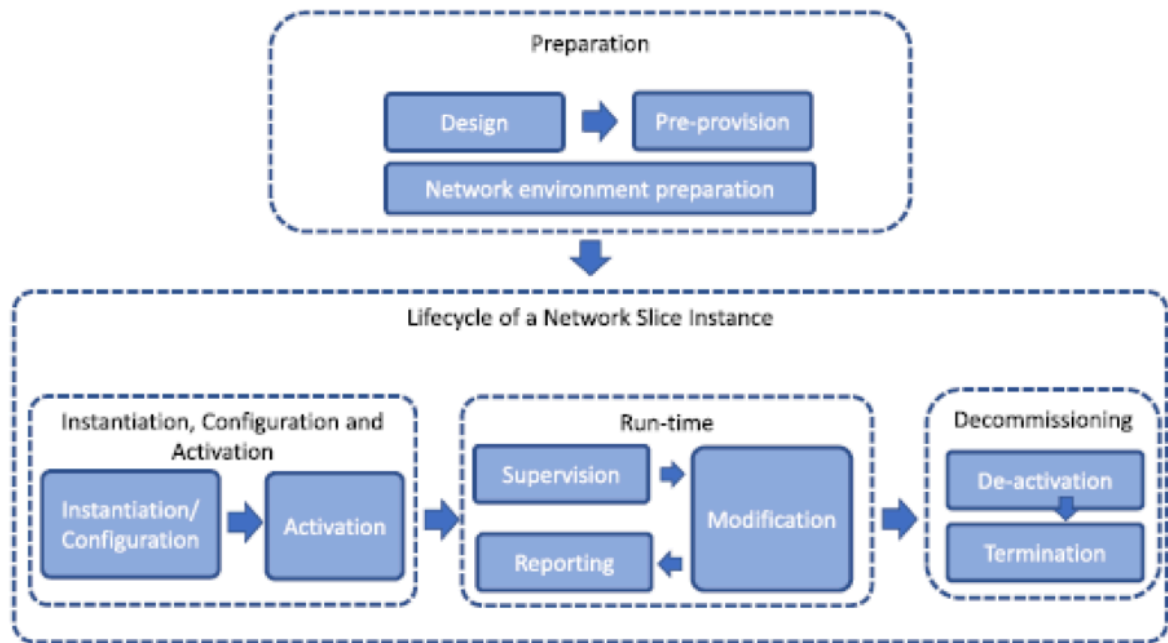
L'isolamento della slice è essenziale per la coesistenza di più slice che condividono la stessa infrastruttura di rete. Assicura infatti che ogni performance della slice non impatta sulle altre. Migliora la sicurezza, poichè impatta su una sola slice. Migliora la privacy.



Il ciclo di vita di un NSI include fasi di:

- Design/Preparazione: progetta e prepara l'ambiente della rete
- Istanziamento: crea e configura la NSI
- Attivazione: Attiva la NSI per l'operazione
- Operazione (monitoraggio e controllo): Monitora e controlla che la NSI incontri i requisiti QoS.
- Modifica: riconfigura la NSI in base ai bisogni.
- Disattivazione: disattiva NSI quando non ha più utilità.

- Terminazione: dealloca le risorse associate alla NSI.



#### 11. Architetture di Network Slicing:

Si affetta l'architettura originale della rete in multiple reti logiche e indipendenti e si configurano in base ai requisiti di servizi richiesti. Si può fare questo tramite la configurazione di NF che provvedono a esprimere funzionalità di rete. Queste funzioni vengono virtualizzate disaccoppiando la funzione di rete dall'hardware. Inoltre si ha un orchestratore che coordina le componenti delle reti che si hanno nel ciclo di una slice e si fa utilizzando un SDN in gradi di avere una configurazione dinamica e flessibile della slice.

Diverse alternative di implementazione esistono per l'architettura di slicing:

- Single Owner, Single Controller: Adatto per regioni di rete limitate sotto un unico proprietario, dove il controller SDN gestisce tutte le slice. Può presentare colli di bottiglia in performance e affidabilità.
- Single Owner, Multiple Tenants – SDN Proxy: L'SDN proxy (es. FlowVisor) permette a più tenant virtuali di implementare i propri controller SDN su un'infrastruttura condivisa, mantenendo l'isolamento tra le slice.
- Multiple Owners, Multiple Tenants: Richiede un layer di virtualizzazione avanzato per mappare le risorse dell'infrastruttura reale alle topologie virtuali richieste dai tenant, consentendo agli orchestratori SDN maggiore libertà.

#### 12. Esempio: Network Slicing nelle Reti Mobili 5G

Il Network Slicing è essenziale per supportare i diversi casi d'uso del 5G:

- eMBB (Enhanced Mobile Broadband): Richiede alta capacità e banda, per applicazioni come video 4K/8K UHD, olografia, AR/VR. Le funzioni di RAN e CN sono più centralizzate.

- URLLC (Ultra-Reliable Low-Latency Communications): Richiede latenza ultra-bassa (es. 1 ms) e alta affidabilità, per applicazioni critiche come chirurgia remota, guida autonoma, fabbriche automatizzate. Le funzioni (PHY, MAC, RLC, PDCP, UPF) sono eseguite al bordo della rete o nella RAN (es. BBU site).
- mMTC (Massive Machine Type Communications): Richiede connettività massiva e alta efficienza energetica, per sensori IoT e città intelligenti. Le funzioni sono posizionate su NOC e Multi-tenant Cloud.

Le Virtual Network Functions (VNFs) sono posizionate nel core o nell'edge cloud in base ai requisiti della slice e interconnesse tramite SDN. Ad esempio, per URLLC, le VNF possono essere spostate nell'Edge Cloud per minimizzare il ritardo di trasmissione e migliorare l'affidabilità. La gNB (Next Generation Node B) in 5G è flessibile e software-oriented, divisa in moduli funzionali (gNB-CU(Cloud Based), gNB-DU(supporta celle multiple livelli RLC,MAC,PHY), gNB-RU (Radio Unit)).

### 13. Posizionamento delle VNF nelle Reti 5G (Approccio Orientato all'Operatore)

Le componenti sono:

- Radio Access Network (RAN)
- Core Network

Un aspetto chiave è il posizionamento flessibile delle funzioni di rete (NF), sfruttando NFV, Network Slicing e l'Edge-to-Cloud Continuum(integrazione del MEC con il classico Cloud). Questo permette di supportare diverse classi di servizio e minimizzare un costo ponderato in base alle preferenze, rispettando i vincoli di rete e servizio.

Gli obiettivi sono quelli di supportare diversi classi di servizio rispettando i vincoli di rete e di servizio e minimizzando il costo di preferenze pesate (peso numerico che penalizza o favorisce certe scelte). Formato da più livelli gerarchici.

- Architettura Multi-Strato Distribuita C-RAN:
  - Layer 1 (L1): Remote Radio Heads (RRHs) al bordo della rete.
  - Layer 2 (L2): BBU (Baseband Unit) pool, che può essere suddiviso in L2a (vicino a RRH per URLLC) e L2b.
  - Layer 3 (L3): Edge Cloud, per l'esecuzione di alcune funzioni.
  - Layer 4 (L4): Network Operator Data Center (NOC), per funzioni dell'operatore.
  - Layer 5 (L5): Multi-Tenant Cloud Infrastructure (es. AWS), una soluzione cloud commerciale. La connettività tra questi strati avviene tramite link ad alta velocità.
- Tecnologie:
  - Wireless tra end device e RRH
  - collegamento in fibra ad alto rate tra RRH e sito BBU
  - collegamento in fibra scura ad alta velocità tra link in backhaul che portano a Edge, NOC o Multi-Tenant Cloud.
- Funzioni RAN e CN in 5G:

- Funzioni RAN: Radio Function (RF), Physical Layer (PHY), Medium Access Control (MAC), Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP).
- Funzioni CN: UPF, AMF, AUSF, SMF, NSSF, NEF, NRF, PCF, UDM. Il posizionamento di queste funzioni dipende dalla slice considerata.
- Funzione di Preferenza per il Posizionamento delle Funzioni: Viene introdotta una funzione di preferenza  $p(f, l, s)$  che “biases” (orienta) il posizionamento della funzione  $f$  sullo strato  $l$  per la slice  $s$ . Questo valore, compreso tra 0 e 1, permette agli operatori di utilizzare il proprio know-how per il dispiegamento dei servizi, bilanciando le richieste degli utenti con le esigenze dell'operatore.
- Obiettivo di Ottimizzazione: L'obiettivo è minimizzare una funzione di costo che considera il carico computazionale delle funzioni, il numero di utenti attivi e il tipo di traffico (UP/CP), rispettando i vincoli di risorse dei nodi e dei link di comunicazione, e i vincoli di ritardo massimo per UP e CP.

Le funzioni del piano Utente, ovvero quelle per il trasporto di dati sono  $F_{UP}$  mentre quelle per il piano controllo  $F_{CP}$ . Il costo della funzione è direttamente proporzionale al numero di utenti attivi nella slice e l'ammontare del traffico per le funzioni UP

Vediamo il costo della funzione  $f$ :

$$\eta'(f, s) = U_t(s) \cdot (u(f)r_s^u + (1 - u(f))r_s^c)$$

Dove  $U_t(s)$  è il numero medio di utenti collegati alla slice  $s$ .

$u(f)$  è una funzione binaria dove è 1 se  $f$  appartiene al UP altrimenti 0. Mentre poi si ha il rate di trasmissione del piano utente e piano di controllo.

Si ha poi il **Costo di Preferenza Pesato**:

$$\eta(f, l, s) = \begin{cases} \frac{1}{p(f, l, s)} \eta'(f, s), & p(f, l, s) \neq 0 \\ \text{N/A}, & p(f, l, s) = 0 \end{cases}$$

Se la preferenza è alta allora il costo è basso se la preferenza è a zero allora la funzione non è allocabile.

L'obiettivo è quello di minimizzare la funzione di costo:

$$P1 : \min\{C(A)\} = \min_A \left\{ \sum_f \sum_l \sum_s a(f, l, s) \cdot \eta(f, l, s) \right\}$$

Dove la matrice  $A$  è una matrice che ha 1 se la funzione  $f$  è piazzata in un nodo del layer  $l$  per la slice  $s$ .

Si deve stare attenti ai vincoli:



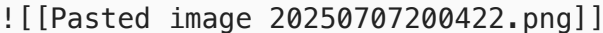
- Vincoli di Risorse:  $\gamma \sum_s \sum_f a(f, l, s) \eta'(f, l, s) < O_l$  dove gamma è il numero di FLOPS per bit e  $\eta'$  è il costo di preferenza pesato. Di conseguenza si vede se il numero di funzioni nel livello l per ogni slice eccede la capacità di computazione.
- Vincoli nel Piazzamento di Funzioni:  $\sum_l a(f, l, s) = 1$  ovvero che la funzione f deve essere posizionato in un solo layer. (Niente Ridondanza).
- Vincoli di Delay:

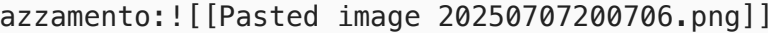
UP Delay Constraints:  $\tau_{UP_s} \geq \delta_{UP_s}(A) + \text{processing delay}$

CP Delay Constraints:  $\tau_{CP_s} \geq \delta_{CP_s}(A) + \text{processing delay}$

- Algoritmi di Soluzione: Per risolvere il problema del posizionamento delle funzioni, vengono utilizzati diversi algoritmi:

– Exhaustive Search Algorithm (ESA)(Soluzione Esatta): Trova la soluzione ottimale ma ha un'alta complessità computazionale, rendendolo impraticabile per implementazioni in tempo reale. Serve per vedere se soluzione dell'algoritmo GA è accurato. Trova una soluzione ottimale nello spazio delle soluzioni  $SP=\{A\}$  e provvede ad una matrice di posizionamento nella rete  $A_{\{ESA\}}$ .

– Maximum Preference Algorithm (MPA)(Soluzione Euristica): Semplice e veloce, ma non considera i vincoli di rete e servizio. Posiziona le funzioni di rete nel livello con più alto valore di preferenza  


– Modified Maximum Preference Algorithm (MMPA)(Soluzione Euristica): Migliora MPA considerando alcuni vincoli, ma la sua flessibilità è ancora limitata. Viene calcolato  $L_{UP_s} \rightarrow \max\{l: \sum \phi(l) \leq \tau_{UP_s}\}$  e  $L_{CP_s} \rightarrow \max\{l: \sum \phi(l) \leq \tau_{CP_s}\}$ . A questo punto si ha la funzione di piazzamento:  


Inoltre  $O_l$  viene aggiornato:  $O_l = O_l - \eta'(f, s)$ .

– Genetic Algorithm (GA): Un algoritmo meta-euristico basato sull'evoluzione che fornisce una soluzione quasi ottimale in tempi ridotti, adatto per il dispiegamento di funzioni di servizio in tempo reale e on-demand.

Algorithm	Computational complexity	Search flexibility	Type	Simulation time per solution point [sec]
MPA	$O(F \cdot L \cdot S)$	VeryLow	Heuristic	0.089
MMPA	$O(F \cdot L \cdot S)$	Low	Heuristic	0.1085
GA		Good	Metaheuristic	24.03
ESA	$O(L^{F \cdot S})$	Complete Search	Optimal	824.82