

Inference Service Provider?

Il ruolo degli operatori nella distribuzione dell'inferenza

Alfredo Giordano

Settembre 2026



Dove entra l'ISP?



**L'inferenza costa meno.
La spesa può crescere.**

20%

Imprese UE che usano AI

2025, imprese con almeno 10 addetti (Eurostat)

÷280

Costo d'inferenza

Riduzione superiore a 280 volte a livello GPT-3.5
Novembre 2022 – ottobre 2024 (Stanford)

20–30%

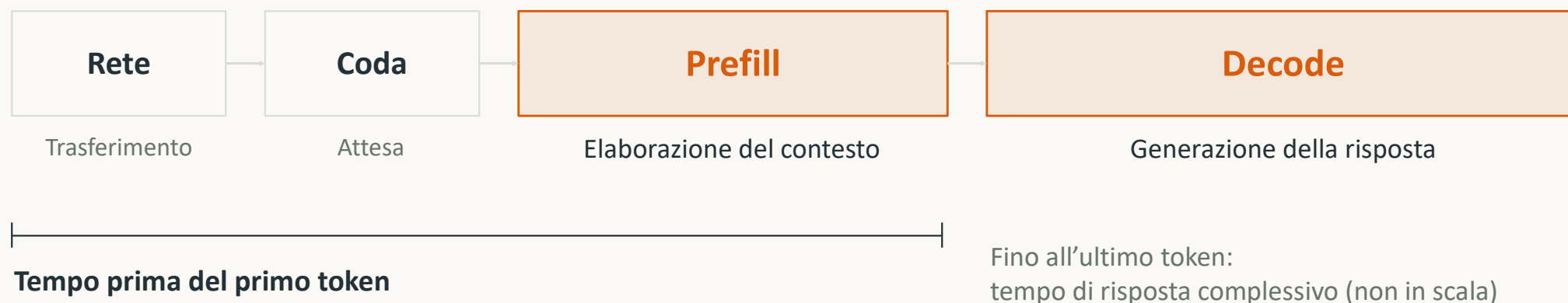
dell'opex telco

Agenti AI e token nello scenario Bain a 3–5 anni (Bain)

Quota della spesa complessiva.
Non un aumento dell'opex del 20–30%.

Dentro una richiesta

Un LLM come esempio. Il profilo di lavoro determina le risorse.



Memoria del modello + memoria delle richieste attive

Contesto, output e concorrenza cambiano il dimensionamento.

Vicino a che cosa?

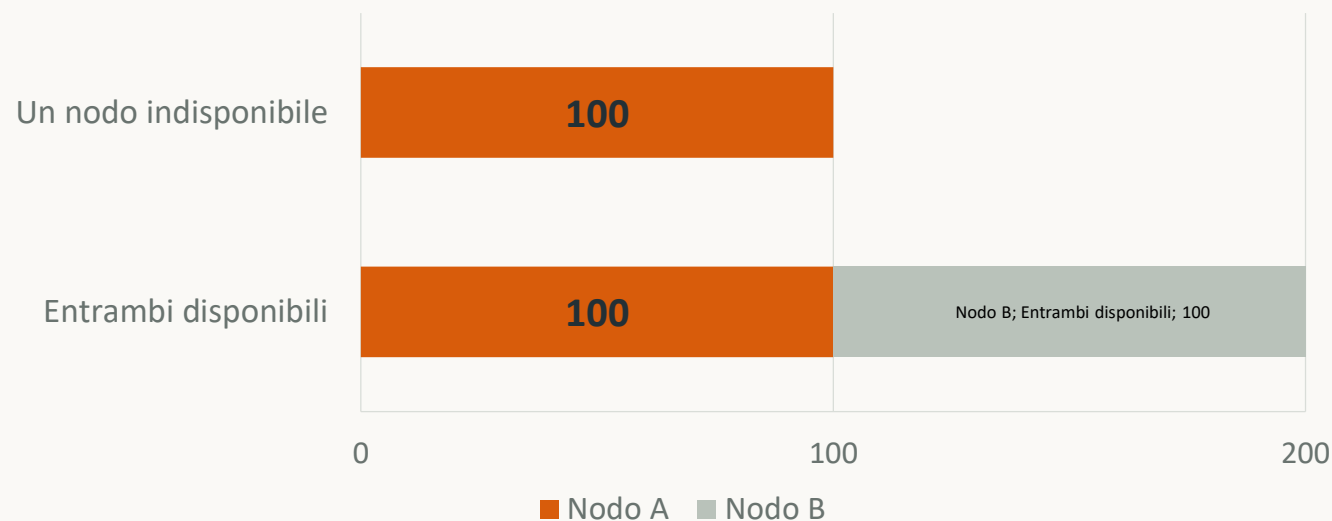
Punto d'ingresso API, luogo di esecuzione e dati possono trovarsi in sedi diverse.



Carico	Fattore da misurare	Collocazione da confrontare
Video locale	Volume dei dati da trasferire	Sito cliente / nodo vicino
Voce interattiva	Tempo di risposta complessivo	Regione / hub
Batch documentale	Costo e tempo di completamento	Hub / prossimità ai dati

La capacità che resta al guasto

Esempio didattico: due nodi indipendenti, ciascuno da 100 richieste/minuto.



100

richieste/minuto
con un nodo disponibile

La capacità di riserva entra nel prezzo.

Valori ipotetici. Stesso modello, profilo di richieste e obiettivo prestazionale. Capacità residua a regime, dopo il failover.

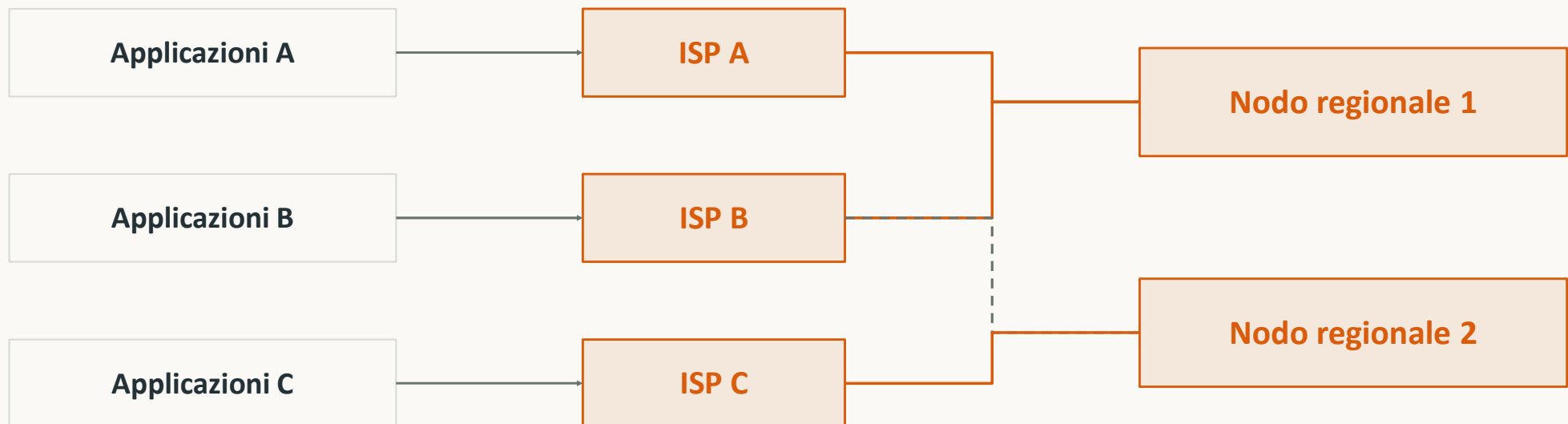
2030: il nodo segue la domanda

Ipotesi: aggregare carichi compatibili rende sostenibile capacità regionale.

Imprese e software house

ISP

Capacità regionale condivisa



Il rapporto con il cliente può precedere la capacità propria.

Per quale funzione ci pagheranno?

Tre posizioni possibili nella stessa catena del valore.



Ruolo dell'ISP	Che cosa acquista il cliente	Ricavo possibile
Distribuzione	Accesso e assistenza	Margine wholesale
Erogazione	Capacità entro condizioni concordate	Consumo / capacità riservata
Gestione	Integrazione e governo del servizio	Canone operativo

Le software house portano il servizio dentro le applicazioni.

La prova che precede l'investimento

Una sperimentazione deve consentire una decisione sulla capacità da acquistare o installare.

01

Applicazione reale

Cliente e software house definiscono un risultato utile.

02

Due co-localizzazioni

Stesso modello, stessi input e criteri di qualità.

03

Carico e guasto

Tempi di risposta, code e ripristino sotto pressione.

04

Margine completo

Capacità di riserva, esercizio e assistenza inclusi.

€/risultato

**Quanto costa
un lavoro riuscito?**

Che cosa può smentirci?



SaaS e dispositivi

Quanta inferenza rimarrà accessibile agli operatori?

Scelta del cliente

Quale vantaggio misurabile ci farà preferire?

Capacità condivisa

Quanto saranno sostituibili modelli e fornitori?

**Quale responsabilità sull'inferenza
siamo pronti a farci pagare?**

Grazie
Grazie

Alfredo.Giordano@Warian.net



DATA  SPACE[®]