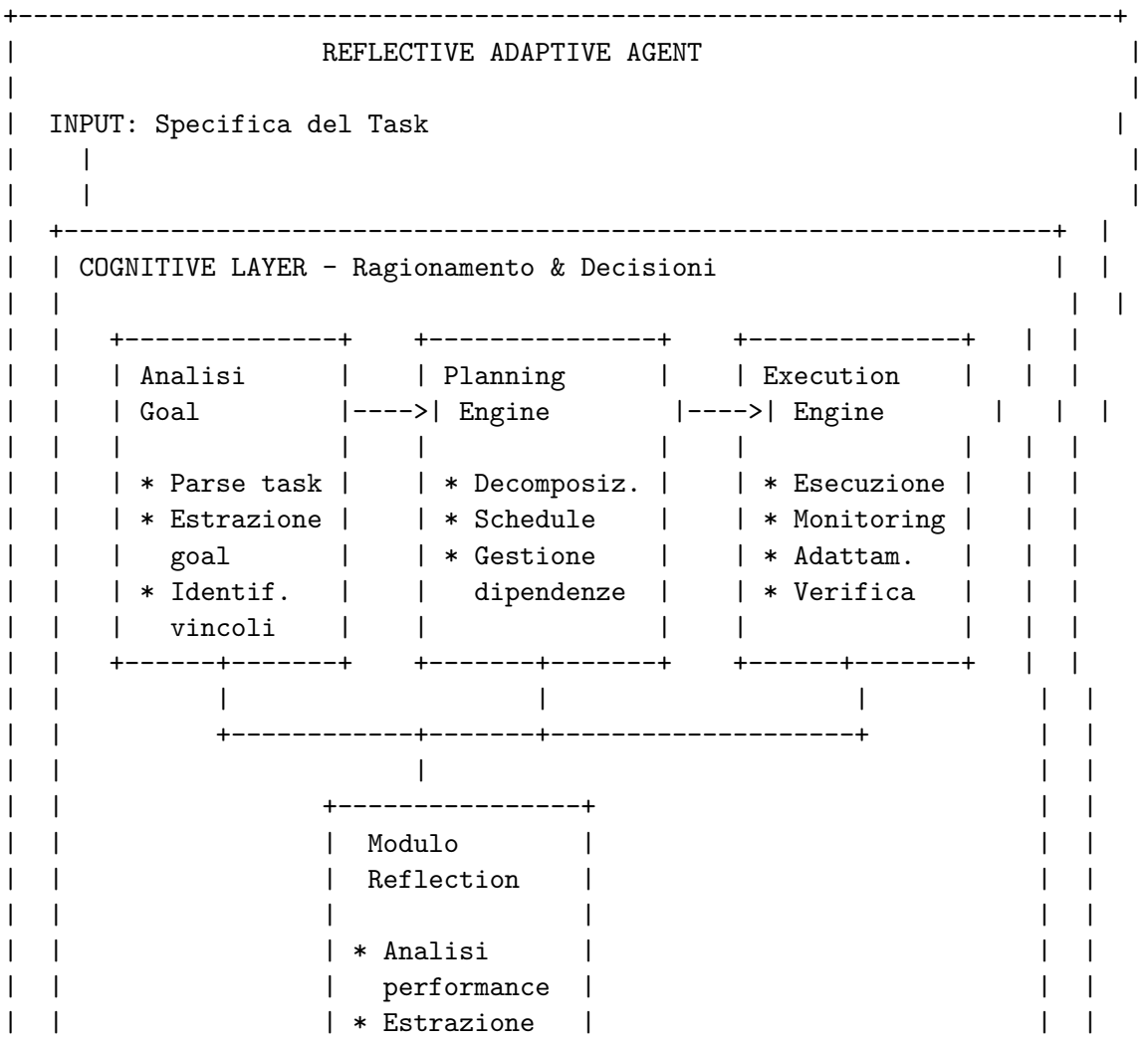
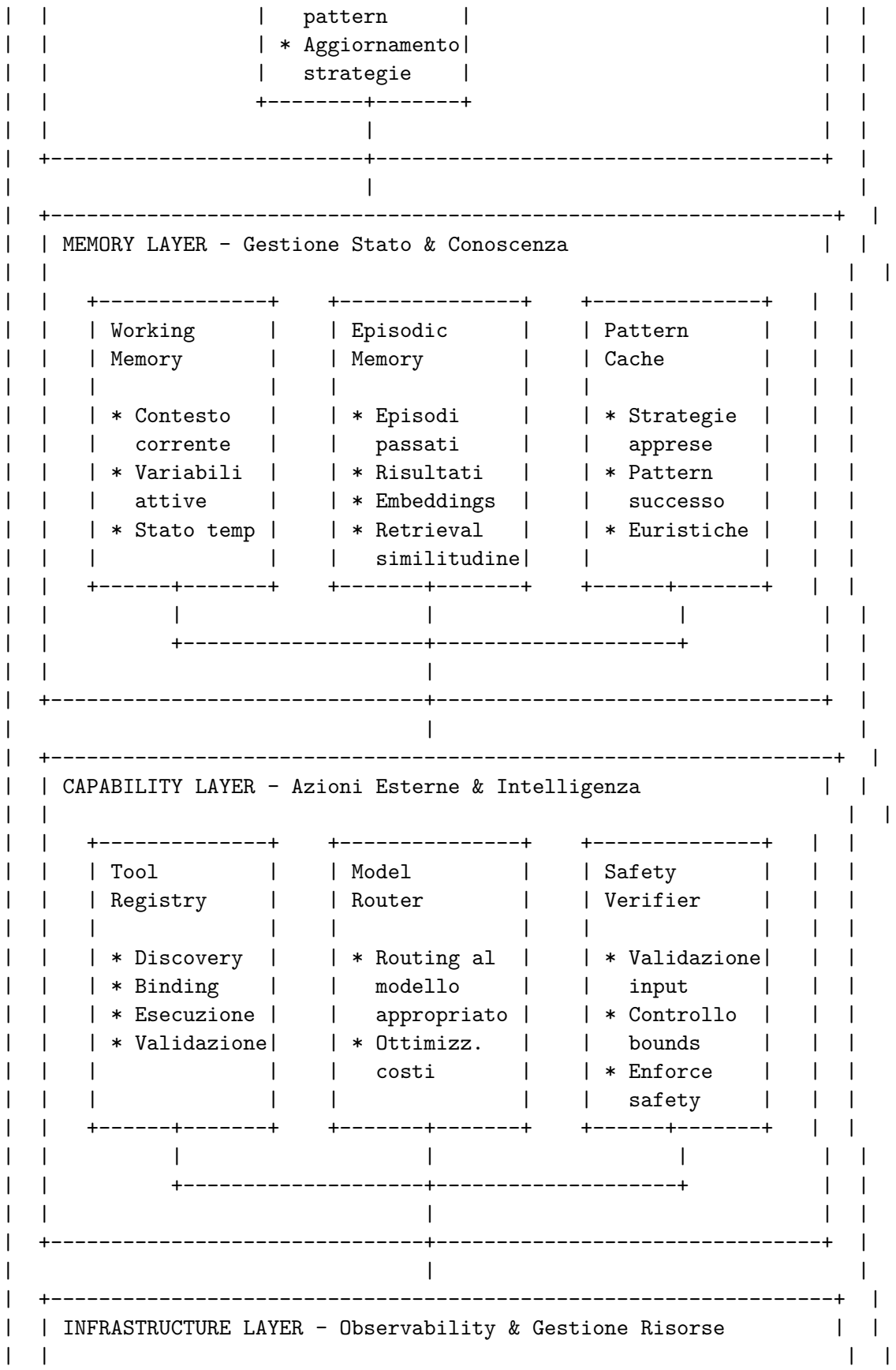
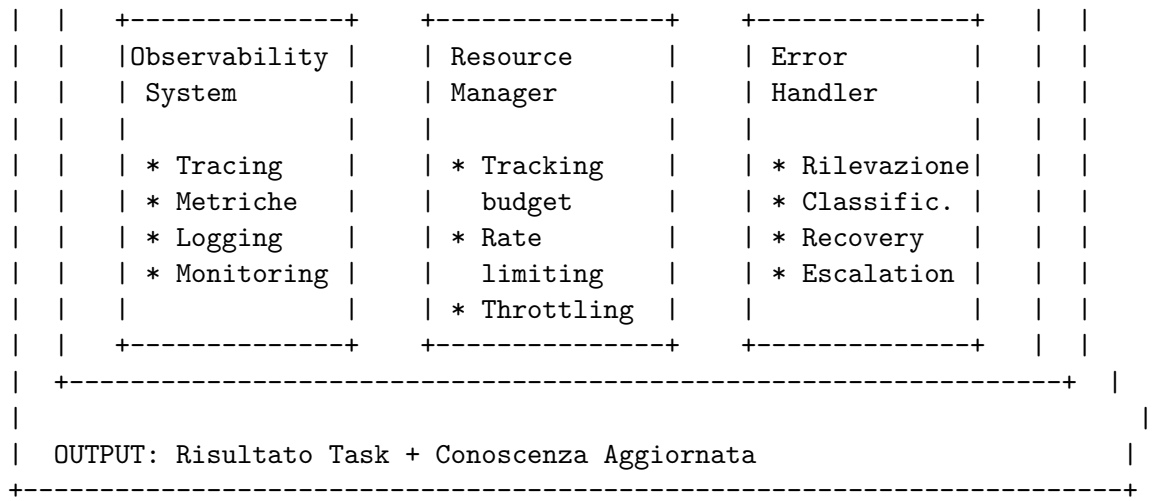
Contents

Architettura di Sistema - Specifiche Complete

Diagramma Generale




pattern
* Aggiornamento
strategie
MEMORY LAYER - Gestione Stato & Conoscenza
Working Memory
* Contesto corrente
* Variabili attive
* Stato temp
Episodic Memory
* Episodi passati
* Risultati
* Embeddings
* Retrieval similitudine
Pattern Cache
* Strategie apprese
* Pattern successo
* Euristiche
CAPABILITY LAYER - Azioni Esterne & Intelligenza
Tool Registry
* Discovery
* Binding
* Esecuzione
* Validazione
Model Router
* Routing al modello appropriato
* Ottimizz. costi
Safety Verifier
* Validazione input
* Controllo bounds
* Enforce safety
INFRASTRUCTURE LAYER - Observability & Gestione Risorse



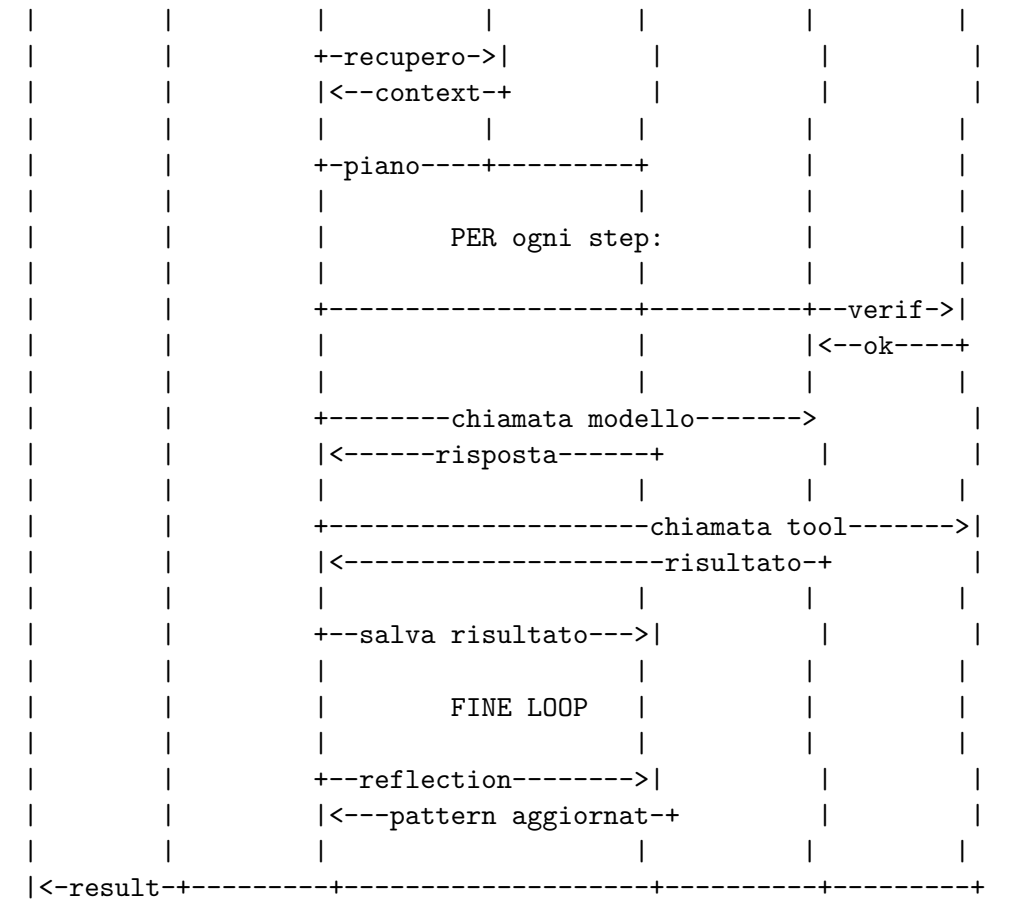
Flusso di Esecuzione - Loop Principale

```
INIZIO
  |
  +--> [1. RICEZIONE TASK]
  |      |
  |      +- Parse input
  |      +- Validazione formato
  |      +- Estrazione parametri
  |      |
  +--> [2. ANALISI GOAL]
  |      |
  |      +- Identificazione obiettivi
  |      +- Estrazione vincoli
  |      +- Definizione criteri successo
  |      +- Classificazione complessità
  |      |
  +--> [3. RECUPERO CONTESTO]
  |      |
  |      +- Query episodic memory (task simili passati)
  |      +- Controllo pattern cache (strategie apprese)
  |      +- Caricamento working memory (sessione corrente)
  |      |
  +--> [4. PIANIFICAZIONE]
  |      |
  |      +- Generazione piano alto livello
  |      +- Decomposizione in subtask
  |      +- Identificazione dipendenze
  |      +- Stima risorse (token, tempo, costo)
  |      +- Selezione strategia esecuzione
  |      |
  +--> [5. LOOP ESECUZIONE]
```

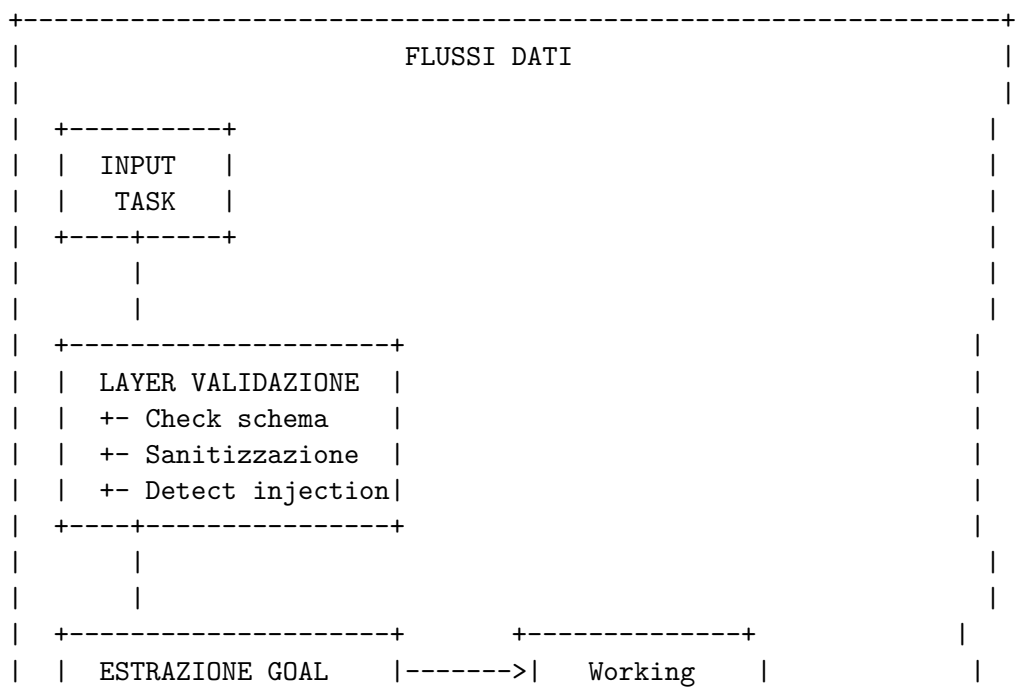
```
  |      |
  |      +- PER ogni step nel piano:
  |      |   |
  |      |   +--> [5a. Pre-esecuzione]
  |      |   |      +- Verifica sicurezza
  |      |   |      +- Controllo risorse
  |      |   |      +- Decisione routing modello
  |      |   |
  |      |   +--> [5b. Esecuzione]
  |      |   |      +- Reasoning LLM (se necessario)
  |      |   |      +- Esecuzione tool (se necessario)
  |      |   |      +- Cattura output
  |      |   |
  |      |   +--> [5c. Post-esecuzione]
  |      |   |      +- Verifica risultato
  |      |   |      +- Controllo criteri successo
  |      |   |      +- Aggiornamento working memory
  |      |   |
  |      |   +--> [5d. Adattamento]
  |      |          +- Se successo: Continua
  |      |          +- Se fallimento: Strategia recovery
  |      |          +- Se bloccato: Ripianificazione
  |      |
  +--> [6. VERIFICA]
  |      |
  |      +- Validazione output finale
  |      +- Controllo criteri successo
  |      +- Verifica sicurezza finale
  |      |
  +--> [7. REFLECTION]
  |      |
  |      +- Analisi performance episodio
  |      +- Estrazione pattern di successo
  |      +- Identificazione miglioramenti
  |      +- Aggiornamento pattern cache
  |      +- Memorizzazione in episodic memory
  |      |
  +--> [8. RESTITUZIONE RISULTATO]
```

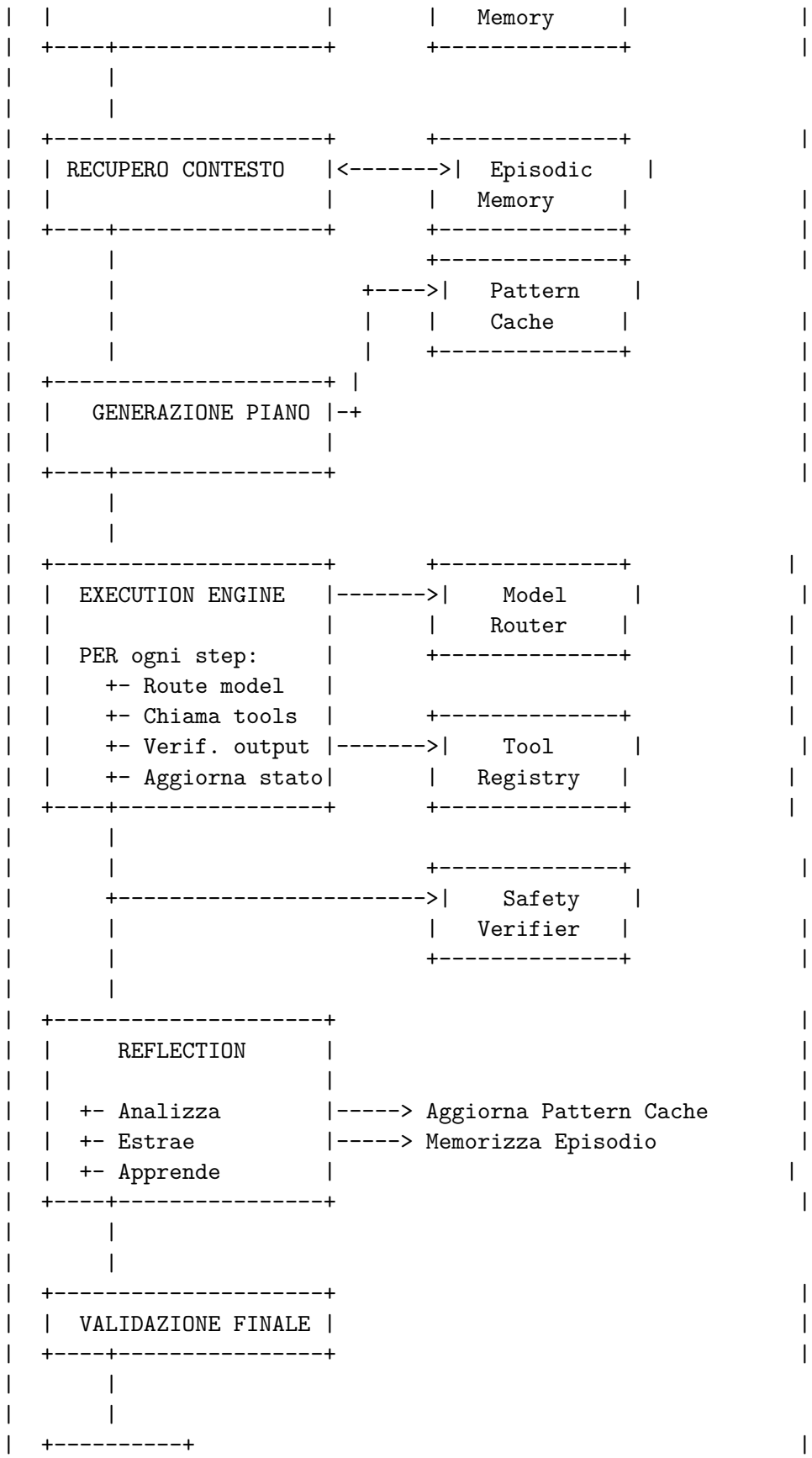
Interazione tra Componenti - Diagramma di Sequenza

```
User    Analisi   Planning   Memory    Model     Tool      Safety
 |      Goal      Engine     System    Router    Registry  Verifier
 |        |         |          |         |          |         |
 +--task-->|        |          |         |          |         |
 |        |         |          |         |          |         |
 |        +-goals--->|         |         |          |         |
```

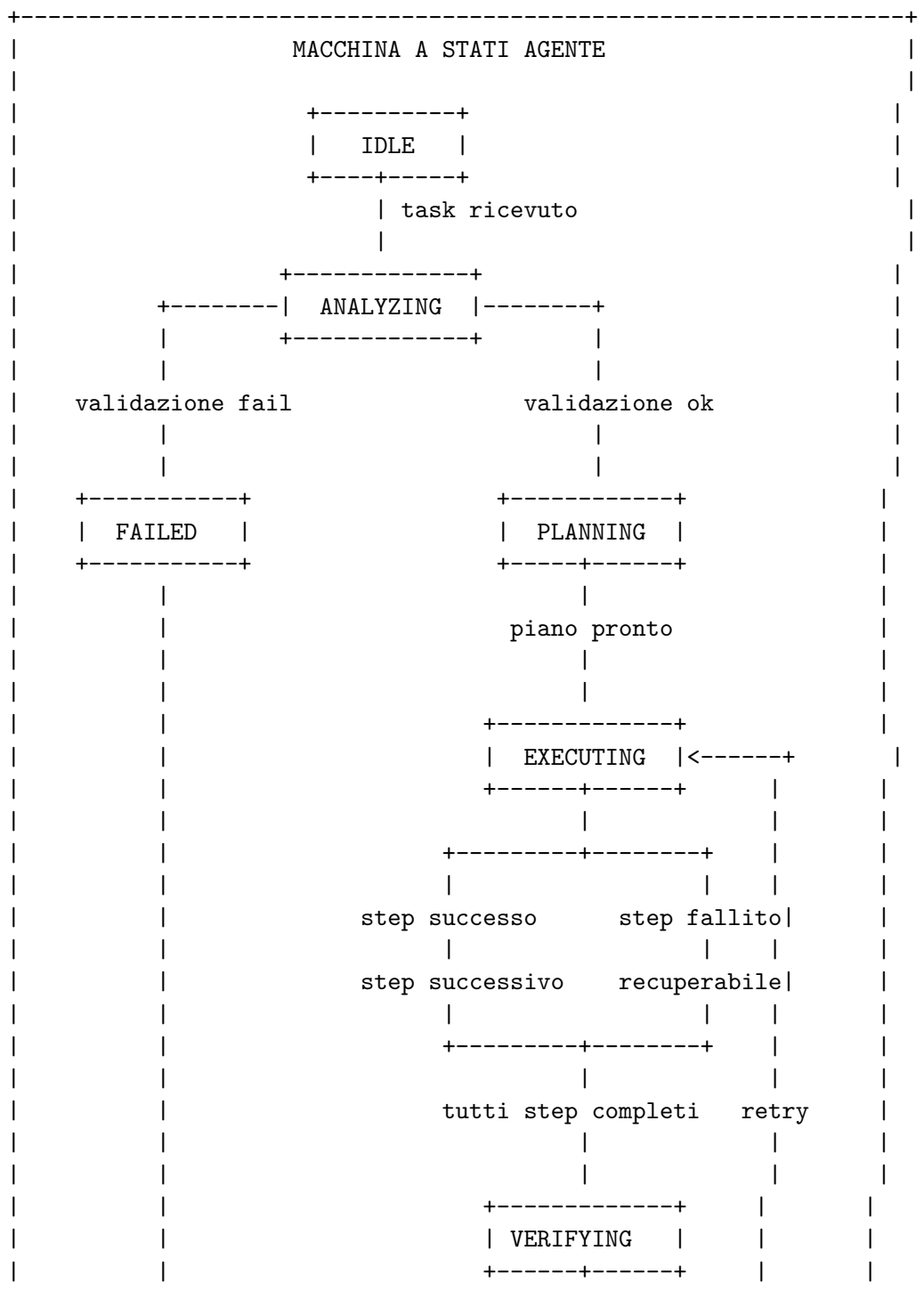


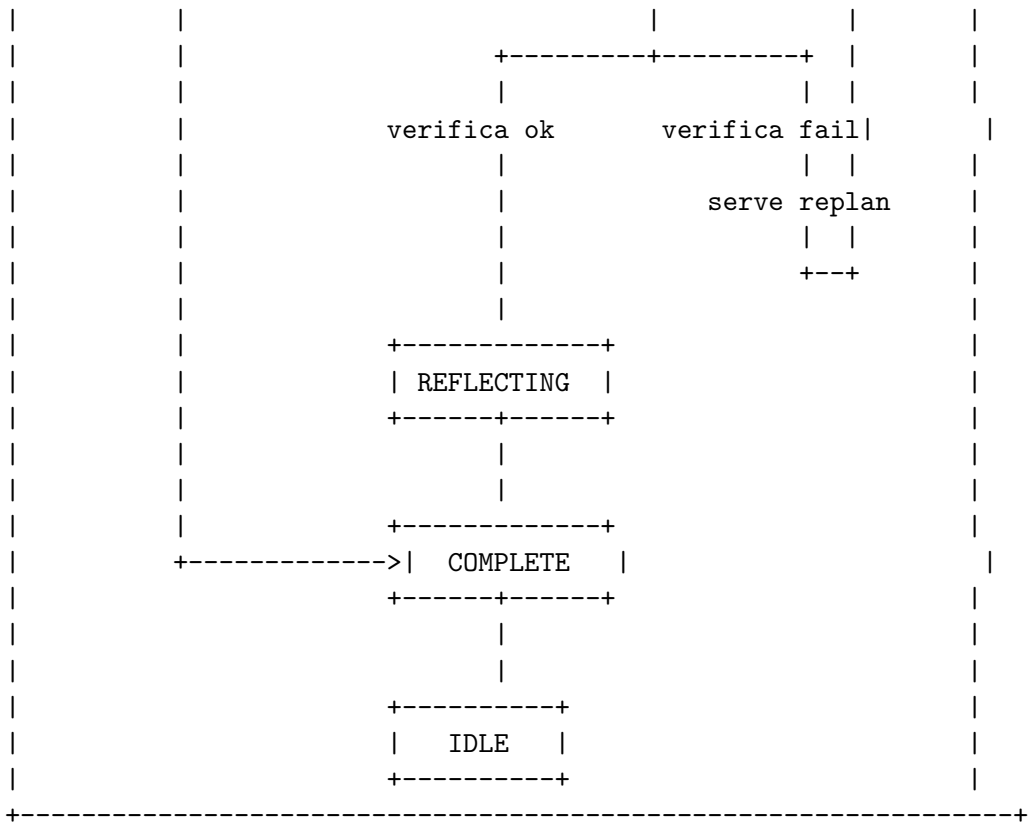
Architettura Flussi Dati



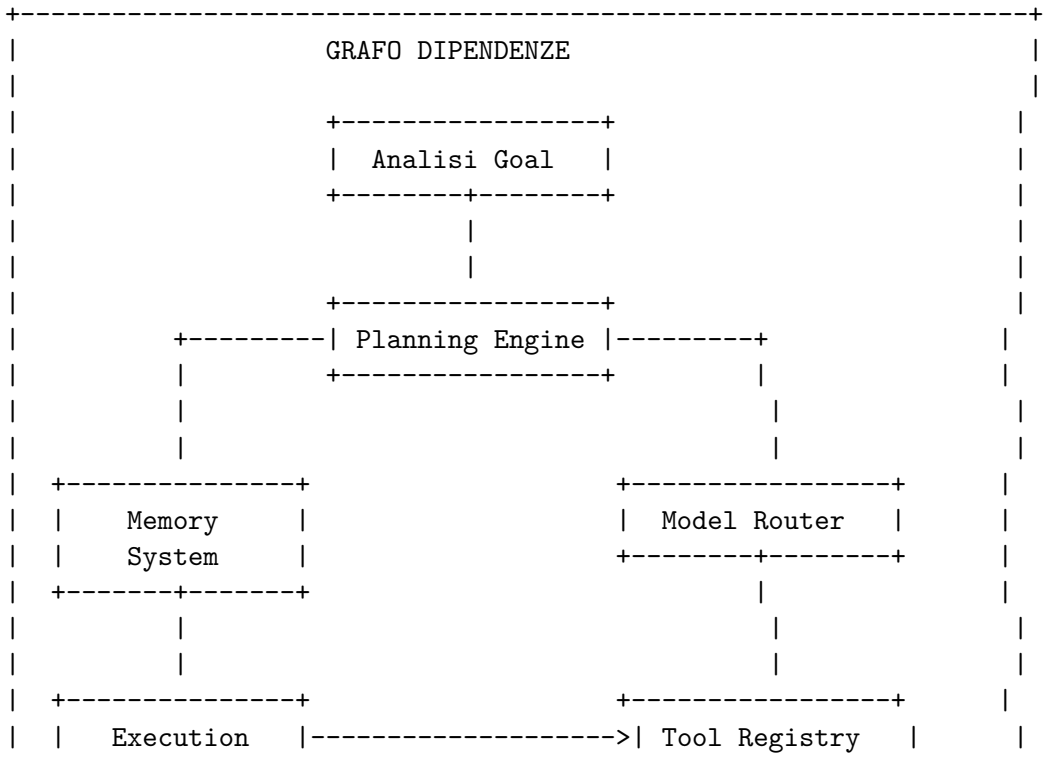

Memory
RECUPERO CONTESTO
Episodic Memory
Pattern Cache
GENERAZIONE PIANO
EXECUTION ENGINE
PER ogni step:
+- Route model
+- Chiama tools
+- Verif. output
+- Aggiorna stato
Model Router
Tool Registry
Safety Verifier
REFLECTION
+- Analizza
+- Estrae
+- Apprende
Aggiorna Pattern Cache
Memorizza Episodio
VALIDAZIONE FINALE

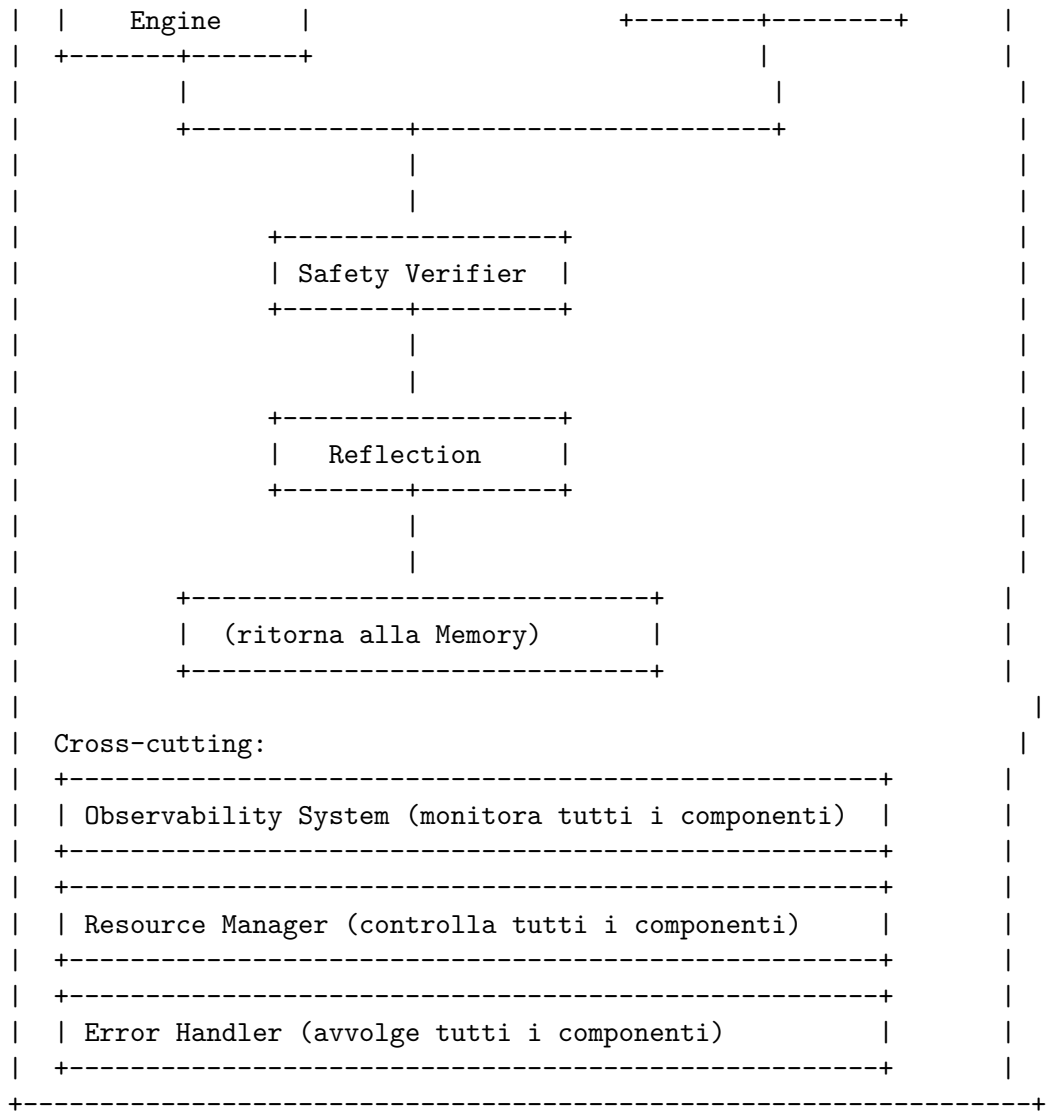
Macchina a Stati - Stati di Esecuzione



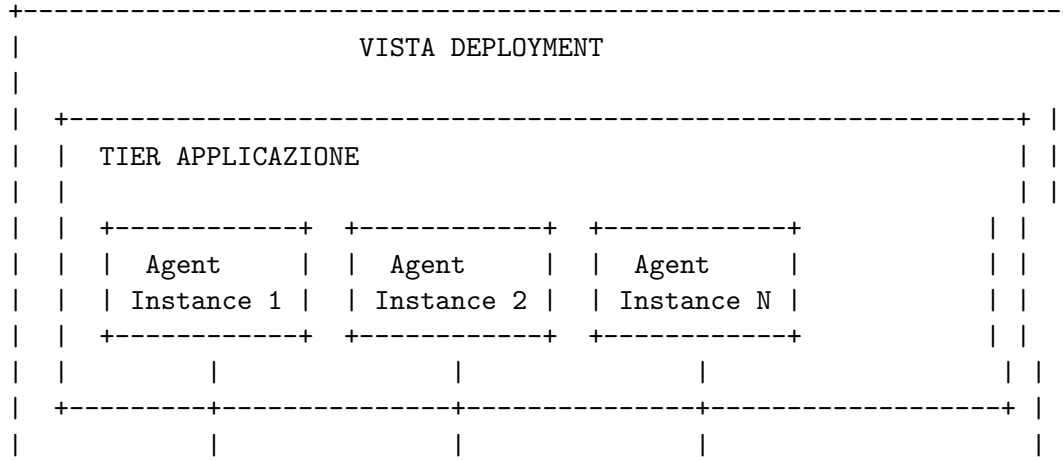


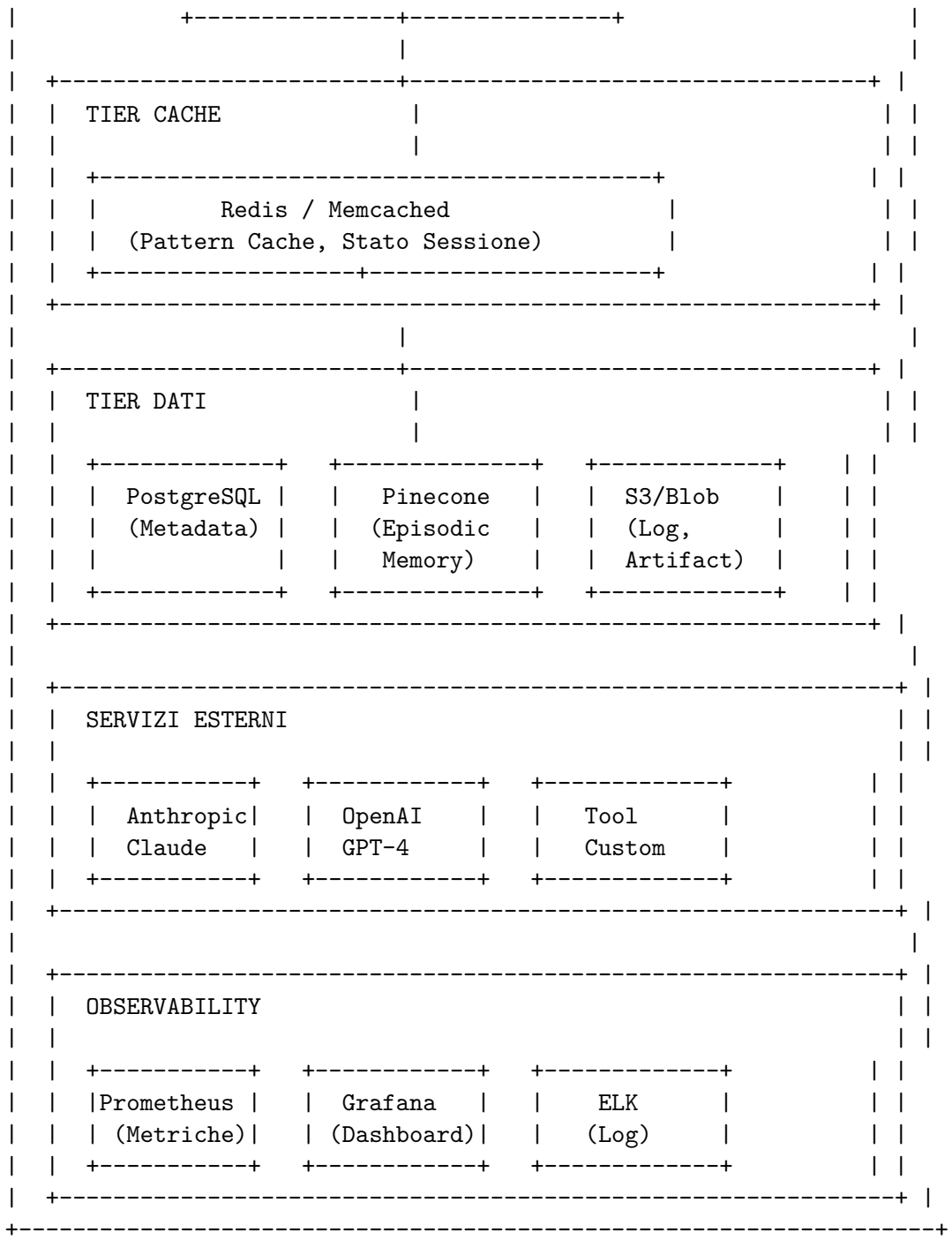
Dipendenze tra Componenti





Architettura di Deployment





Caratteristiche di Performance

Breakdown Budget Latenza

Target Totale: 5-30s per task complessità media

Componente	Tempo	% Totale

```
+---------------------------------------+
| Analisi Goal        0.5s      2%      |
| Recupero Contesto   1.0s      5%      |
| Pianificazione      2.0s      10%     |
| Esecuzione:                           |
|   - Chiamate model  15.0s     70%     |
|   - Esecuzione tool 3.0s      13%     |
| Verifica            0.5s      2%      |
| Reflection          1.0s      5%      |
+---------------------------------------+
| TOTALE              ~23s      ~100%   |
+---------------------------------------+

Strategie di ottimizzazione:
- Routing modelli (usa modelli veloci quando sufficiente)
- Esecuzione parallela tool
- Riuso pattern cached
- Streaming per latenza percepita
```

Caratteristiche di Scalabilità

```
+-----------------------------------------------------------+
| Dimensione        Corrente  Target    Strategia Scaling  |
+-----------------------------------------------------------+
| Utenti Concurrent 10-50     100-500   Orizzontale        |
| Task/Ora          100-500   5K-10K    Orizzontale        |
| Episodi Memory    10K       1M+       Scala Vector DB    |
| Pattern Cache     1K        10K+      Cache distribuita  |
| Latenza Media     5-30s     <30s      Ottimizz. modelli |
| Success Rate      85-95%    >90%      Apprendimento cont|
+-----------------------------------------------------------+
```

Prossimo: 02-cognitive-layer.md -> Specifiche dettagliate dei componenti di ragionamento