



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

Scuola di Ingegneria

Corso di Laurea in Ingegneria Informatica

Software Architecture and Methodologies

&

Quantitative Evaluation of Stochastic Models

# **Quarkus car-rental and E2E workflow execution time analysis**

*Edoardo Sarri*

7173337

Dicembre 2025

# Indice

<b>1 Car Rental Application</b>	<b>6</b>
1.1 Architettura . . . . .	6
1.2 Servizi Esterni . . . . .	7
1.3 Comunicazione . . . . .	8
<b>2 Deployment</b>	<b>10</b>
2.1 Environment . . . . .	10
2.2 Servizi Esterni . . . . .	12
2.2.1 Helm . . . . .	12
2.2.2 Deployment con Helm . . . . .	13
2.3 Health . . . . .	13
2.3.1 MicroProfile Health . . . . .	13
2.3.2 Implementazione . . . . .	14
<b>3 Tracing and Metrics</b>	<b>15</b>
3.1 Tracing . . . . .	15
3.1.1 OpenTelemetry . . . . .	15
3.1.2 Collector . . . . .	16
3.1.3 Jaeger . . . . .	16
3.1.4 OTel workflow . . . . .	17
3.2 Metrics . . . . .	17
3.2.1 Prometheus . . . . .	18
3.2.2 Grafana . . . . .	18
<b>4 Workflow</b>	<b>20</b>
4.1 Blocchi . . . . .	20
4.1.1 XOR . . . . .	21
4.1.2 AND . . . . .	21
4.2 Implementazione . . . . .	21
4.2.1 XOR . . . . .	21
4.2.2 AND . . . . .	22
4.3 Nuovo workflow . . . . .	23
<b>5 Analisi E2E</b>	<b>25</b>
5.1 Eulero . . . . .	25
5.1.1 Workflow complessi . . . . .	25
5.1.2 Tool e analisi bottom-up . . . . .	26

5.1.3	Dati necessari . . . . .	26
5.2	Tracce . . . . .	26
5.2.1	Generazione . . . . .	26
5.2.2	Elaborazione . . . . .	27
5.2.3	Risultati . . . . .	27
5.3	Distribuzioni . . . . .	28
5.3.1	Approssimanti di Whitt . . . . .	28
5.3.2	Risultati . . . . .	29
5.4	Distribuzione E2E . . . . .	29
5.5	Bande di confidenza . . . . .	30
5.5.1	Tecniche . . . . .	30
5.5.2	Risultati . . . . .	32
<b>6</b>	<b>Correlazione</b>	<b>34</b>
6.1	Implementazione . . . . .	34
6.1.1	Bassa correlazione . . . . .	35
6.1.2	Alta correlazione . . . . .	35
6.2	Risultati . . . . .	36
6.3	Bande di confidenza . . . . .	37

# Elenco delle figure

1.1	<i>car-rental</i> users use cases [14]. . . . .	6
1.2	<i>car-rental</i> communication [14]. . . . .	9
3.1	OpenTelemetry workflow. . . . .	17
3.2	Prometheus-Grafana Flow. . . . .	18
4.1	Workflow iniziale di Car-Rental. . . . .	20
4.2	Nuovo workflow di reserve mostrato da Jaeger. . . . .	23
4.3	Sequence diagram dei blocchi XOR e AND. . . . .	24
5.1	STPN del workflow [10]. . . . .	29
5.2	Risultati dell'esperimento. . . . .	29
5.3	Bande di confidenza. . . . .	33
6.1	Correlazioni iniziali, i.e. senza l'aggiunta esplicita nel codice. . . . .	34
6.2	Matrici di correlazione nei servizi con bassa correlazione. . . . .	35
6.3	Matrici di correlazione nei servizi con alta correlazione. . . . .	35
6.4	Distribuzioni negli esperimenti con correlazione diversa. . . . .	36
6.5	Bande di confidenza per le distribuzioni con bassa correlazione. . . . .	37
6.6	Bande di confidenza per le distribuzioni con alta correlazione. . . . .	37

# Listings

2.1	Kubernetes and Docker Configuration . . . . .	11
2.2	Reference in <i>users-service</i> . . . . .	12
2.3	MySQL Helm chart . . . . .	13
2.4	JDBC configuration. . . . .	14
3.1	Jaeger configuration for <i>users-service</i> . . . . .	16
3.2	Install Jaeger chart for <i>users-service</i> . . . . .	16
3.3	Prometheus configuration for <i>users-service</i> . . . . .	18

# 1 Car Rental Application

L'applicazione utilizzata durante l'intero progetto è stata [acme-car-rental](#). Si tratta di un'applicazione a micro servizi sviluppata all'interno del libro *Quarkus in Action* [14]. Esso mostra le caratteristiche principali di Quarkus e sviluppa le varie funzionalità dell'applicazione capitolo per capitolo, rilasciando su GitHub la versione finale e completa alla fine.

Lato utente l'applicazione permette, come si vede in Figura 1.1, di svolgere le attività di base che ogni applicazione di noleggio auto dovrebbe implementare: ricerca di auto libere in un determinato intervallo di tempo; prenotazione di un'auto definito tale intervallo; ricerca delle auto prenotate; aggiunta e rimozione di una'auto dall'inventario della auto disponibili.

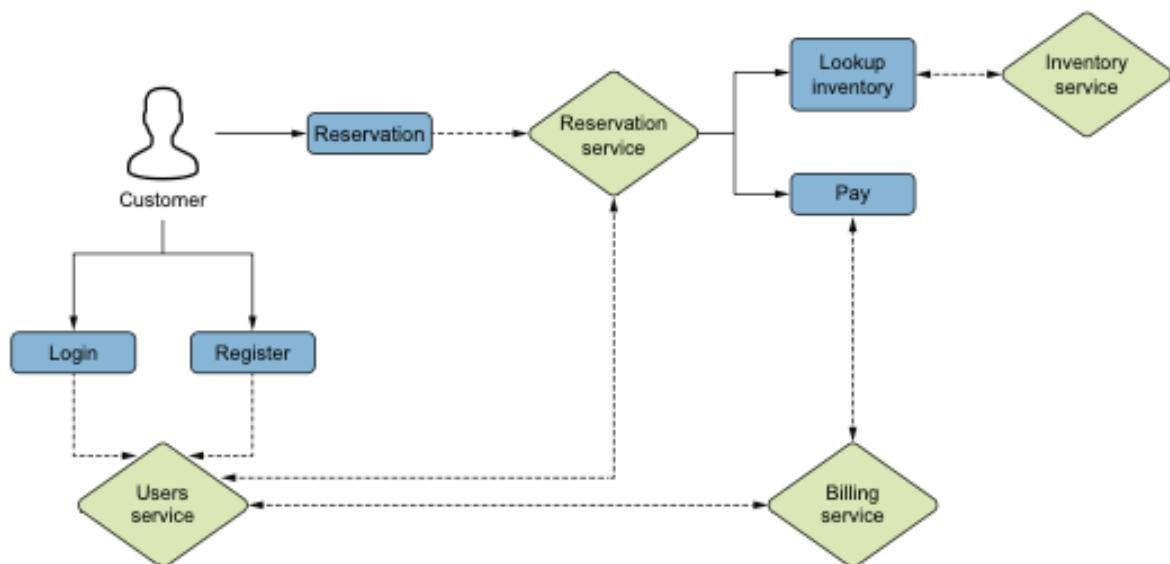


Figura 1.1: *car-rental* users use cases [14].

Oltre a quanto specificato in questo report e a quanto si trova nella [mia repo](#), la versione originale rilasciata conteneva anche degli elementi che sono stati eliminati visto che per il nostro scopo non erano necessari.

## 1.1 Architettura

Vediamo come prima cosa l'architettura dell'applicazione. I micro servizi che car-rental definisce sono cinque:

- **Billing-service:**

Gestisce i pagamenti e le fatture. Viene chiamato da *reservation-service* quando un utente effettua una prenotazione e da *rental-service* quando un'auto viene noleggiata.

- **Inventory-service:**

Gestisce l'inventario delle auto. Fornisce l'elenco delle auto disponibili e permette agli impiegati di aggiungere o rimuovere veicoli.

- **Rental-service:**

Gestisce il processo di noleggio effettivo. Un impiegato può avviare un noleggio attraverso questo servizio, che a sua volta interagisce con il *billing-service*.

- **Reservation-service:**

Gestisce le prenotazioni delle auto. Verifica la disponibilità dei veicoli tramite l'*inventory-service* e, in caso di prenotazione, avvia il processo di pagamento tramite il *billing-service*.

- **Users-service:**

Fornisce una semplice interfaccia per la gestione delle prenotazioni.

Nel codice originale questo servizio si appoggiava a *Keycloak* per la gestione degli utenti. Per problemi di deployment e gestione, ho preferito rimuovere questa dipendenza; l'applicazione adesso permette a un solo utente, *guest*, di effettuare prenotazioni.

## 1.2 Servizi Esterni

Oltre ai servizi business dell'applicazione, descritti nella Sezione 1.1, essa si basa anche su dei servizi esterni, non sviluppati dai produttori ma che devono essere configurati e da cui i servizi principali dipendono. Questi sono:

- **Kafka:**

È una piattaforma di streaming di eventi distribuita, utilizzata per la comunicazione asincrona e mantenuta dalla Apache Software Foundation.

Gli eventi in Kafka vengono aggiunti in fondo a una coda e sono i subscriber che si devono occupare di gestire i messaggi già letti e quelli da leggere.

È ideale per streaming di dati di grandi dimensioni.

- **RabbitMQ:**

È un sistema di message broker che implementa il protocollo AMQP (Advanced Message Queuing Protocol), usato per la comunicazione asincrona.

Rispetto a Kafka, in RabbitMQ è il publisher che si occupa di distribuire i messaggi e che si assicura della loro ricezione.

È ideale per creare code di lavoro, dove ogni messaggio è un compito da eseguire.

- **MongoDB:**

È un database NoSQL orientato ai documenti, che memorizza i dati in file simili ai JSON (detti BSON).

È ideale per quelle applicazioni i cui requisiti sui dati non sono già stati ben definiti e che vogliono scalare orizzontalmente.

- **MySQL:**

È un database relazionale open-source.

È ideale per la memorizzazione di dati strutturati, per la rapidità e la semplicità d'uso.

- **PostgreSQL:**

È un database relazionale object-oriented open-source.

È più conforme agli standard SQL e offre un set di funzionalità per la gestione di query complesse più ricco rispetto a MySQL.

## 1.3 Comunicazione

Vediamo infine come i micro servizi, sia di business che esterni, comunicano tra loro. La Figura 1.2 illustra i vari flussi di comunicazione; da notare che *inventory-CLI* è uno di questi servizi rimossi rispetto alla versione rilasciata da Quarkus in Action, vista la sua inutilità nel nostro contesto.

- **REST:**

È lo stile architetturale di comunicazione più utilizzato nel web e si basa sul protocollo HTTP. È ideale quando si cerca una comunicazione sincrona tra servizi, dove un client invia una richiesta e attende una risposta.

Nell'applicazione viene usato ad esempio da *reservation-service* per interrogare *inventory-service* sulla disponibilità dei veicoli.

- **GraphQL:**

È un'alternativa più flessibile a REST, basata sempre sul principio client-server, che consente ai client di richiedere esattamente i dati di cui hanno bisogno; in questo modo il client non deve filtrare tutto quello che riceve.

Nell'applicazione viene usato da *users-service* per aggregare dati da diversi servizi con una singola richiesta, semplificando l'interazione dal lato client.

- **gRPC:**

Protocollo di comunicazione che si basa su HTTP/2, utilizza il formato binario, che permette di inviare più richieste contemporaneamente sullo stesso canale e di ricevere le relative risposte contemporaneamente. È ottimo nelle situazioni in cui si richiede una latenza minima con una comunicazione molto efficiente.

In questo progetto, *inventory-service* espone un endpoint gRPC che viene utilizzato da *reservation-service* per ottenere informazioni sulle auto in modo performante.

- **Kafka:**

Nell'applicazione viene usata per notificare eventi come l'inizio o la fine di un noleggio tra il *rental-service* e il *billing-service*. Questo garantisce una comunicazione resiliente anche in caso di fallimenti temporanei di uno dei due.

- **RabbitMQ:**

Viene usato da *reservation-service* per comunicare la creazione di una nuova prenotazione a *billing-service*, che si occuperà di creare la fattura.

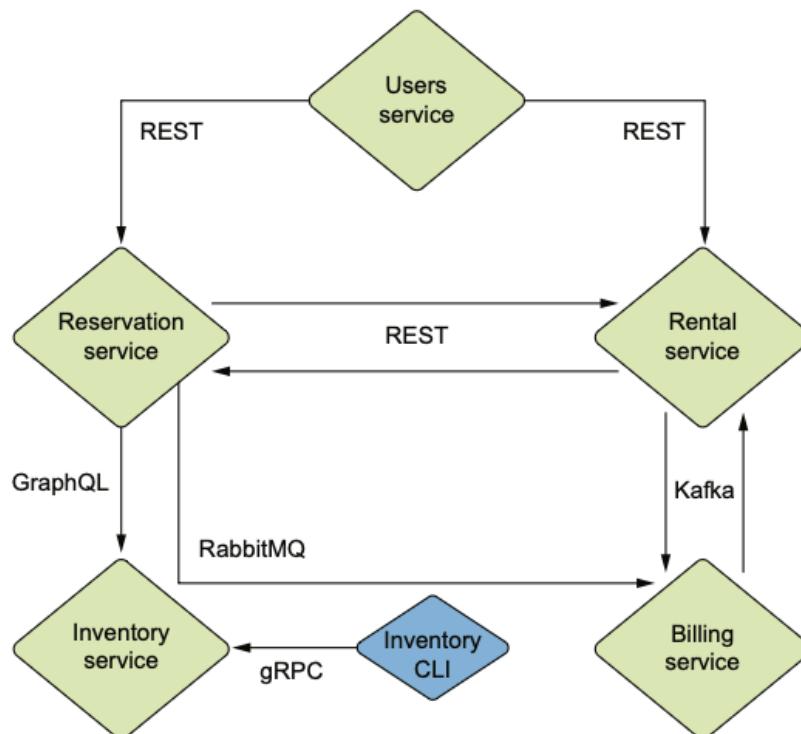


Figura 1.2: *car-rental* communication [14].

# 2 Deployment

In questo capitolo descriviamo il processo di deployment, più complesso di quello che ci aspettavamo. Nonostante infatti l'applicazione fosse già sviluppata in ogni suo aspetto, essa era predisposta per essere rilasciata su OpenShift e usare Quay.io come registry per le immagini.

Per questi motivi il file di configurazione Quarkus *application.properties* sono stati completamente rivisti e adattati alla nostra configurazione, che come abbiamo detto nel Capitolo ?? comprende Minikube come ambiente di deployment e Docker come runtime per i container.

Per automattizzare il deployment, nella root directory del progetto è presente una cartella *exec*, all'interno della quale è presente il file *deployment.sh*. Se eseguito a partire dalla root directory e con il Docker demon in esecuzione, allora verrà creato un cluster Minikube con tutti i micro servizi e le loro dipendenze. Il corretto funzionamento di questo script è garantito su MacOS con ARM64.

All'interno dello script viene usato, per ogni micro servizio, il comando *quarkus build*. Questo permette di costruire l'immagine Docker in automatico, senza scrivere a mano il manifesto Kubernetes; l'unico requisito è l'utilizzo dell'estensione *quarkus-container-image-docker*. Inoltre, cosa che in Minikube non è permesso, sarebbe consentito anche eseguire il deployment della nuova immagine appena costruita nell'ambiente di Deployment; questo può essere fatto tramite la configurazione *quarkus.container-image.push=true*.

## 2.1 Environment

L'ambiente per il rilascio dell'applicazione è stato Minikube, un ambiente locale di sviluppo basato su Kubernetes. In questo modo è stato possibile testare il deployment in un ambiente simile a quello di produzione senza la necessità di risorse hardware dedicate. Come abbiamo detto, nonostante l'applicazione rilasciata su GitHub sia completa e configurata per un tipo di deployment, ci sono stati vari aspetti da considerare e che hanno richiesto attenzione e tempo:

- **OpenShift**

L'applicazione era stata pensata per essere distribuita su OpenShift, la piattaforma per container basata su Kubernetes e sviluppata da Red Hat. Nonostante essa sia appunto Kubernetes-based, in Quarkus si richiede di specificare quale ambiente di deployment si intende utilizzare all'interno del file di configurazio-

ne *application.properties*.

Un esempio di configurazione corretta per Kubernetes, usando Docker come runtime di container, è quella mostrata nel Listing 2.1: questo è il contenuto del file di configurazione Quarkus per *inventory-service*, ma generalizzando è lo stesso usato in tutti gli altri microservizi. Oltre ai comandi che sono naturalmente interpretabili, possiamo specificare alcuni concetti più complessi: `quarkus.container-image.build=true` permette di costruire l'immagine in automatico quando eseguiamo il comando `quarkus build`; `quarkus.kubernetes.service-type=NodePort` istruisce Kubernetes a creare un Service di tipo *NodePort* (di default è *ClusterIP*) e quindi accessibile anche dall'esterno del cluster (meno corretto in produzione di *LoadBalancer*, ma giusto per test locali); `quarkus.kubernetes.image-pull-policy=Never` istruisce Kubernetes a far fallire l'avvio del pod se l'immagine Docker su cui si basa non è presente localmente (a differenza di *Always* che la scarica sempre da un registry).

- **Registry**

Una volta costruita l'immagine docker, il file di configurazione era progettato per eseguire il push su un registry, Quay.io, registry di Red Hat che fa parte dello stesso ecosistema di OpenShift. Siccome il nostro ambiente per il deployment era Minikube, sono state fatte delle modifiche anche in questo senso in modo da rilasciare l'immagine del microservizio all'interno di Minikube stesso.

In questo scenario è fondamentale la configurazione `quarkus.container-image.push=false` all'interno del Listing 2.1.

- **Riferimenti**

Usando un ambiente di produzione diversi (OpenShift vs Minikube) rispetto a quello per cui l'applicazione era stata pensata, si sono dovuti cambiare gli URL con cui un microservizio identificava le proprie dipendenze. Un esempio di quello che è stato fatto si trova nel Listing 2.2, relativo al micro servizio *users-service*: i riferimenti per l'ambiente di sviluppo sono stati lasciati invariati, mentre sono state configurate correttamente le variabili d'ambiente all'interno del container in modo che sia presente il riferimento alla dipendenza usata.

```
# container
quarkus.container-image.build=true
quarkus.container-image.push=false
quarkus.container-image.group=edoardosarri
quarkus.container-image.name=inventory-service
quarkus.container-image.tag=2.0
# kubernetes
quarkus.kubernetes.name=inventory-service
```

```
quarkus.kubernetes.deployment-target=kubernetes  
quarkus.kubernetes.service-type=NodePort  
quarkus.kubernetes.image-pull-policy=Never
```

Listing 2.1: Kubernetes and Docker Configuration

```
# reference  
%dev.quarkus.rest-client.reservations.url=http://localhost  
:8081  
%prod.quarkus.rest-client.reservations.url=http://reservation-  
service
```

Listing 2.2: Reference in *users-service*

## 2.2 Servizi Esterni

Come abbiamo già detto nella Sezione 1.2, i micro servizi business dell'applicazione utilizzano al loro interno altri servizi ausiliari. Questi devono eseguire all'interno dello stesso cluster Minikube e quindi devono essere in un qualche modo rilasciati.

La prima soluzione esplorata è stata farsi scrivere da un LLM il manifesto Kubernetes per l'ambiente Minikube. Nonostante questa fosse una soluzione funzionante, mi sembrava chiaro che ci potesse essere un'alternativa off-the-shelf che mi permettesse di non dover fare configurazioni manuali; l'alternativa adottata è stata Helm.

L'unico servizio esterno per cui è stato generato (da un LLM, in particolare Gemini) il manifesto invece che usare un chart è stato MongoDB. Il problema, che non sono riuscito a risolvere, è stato quella della compatibilità tra la versione di MongoDB e quella di Minikube.

### 2.2.1 Helm

Helm è un gestore di pacchetti Kubernetes che semplifica il ciclo di vita delle applicazioni all'interno di Kubernetes.

Tramite i Chart possiamo definire, installare e aggiornare anche le applicazioni Kubernetes più complesse. Essi permettono infatti di non generare manifesti complessi e ridondanti a mano, ma di utilizzare una struttura predefinita e riutilizzabile. Nello stesso modo in cui esiste un database di immagini Docker, esiste un repository di Chart Helm, ArtifactoryHUB; oltre a fornirci molte applicazioni, il repository ci fornisce tutte le istruzioni necessarie per configurazioni avanzate [2].

## 2.2.2 Deployment con Helm

Per capire come Helm semplifica il deployment di servizi terzi da cui un nostro micro servizio dipende, prendiamo come esempio l'installazione del database MySQL, necessario per *inventory-service*. Tramite un comando, seguito da pochi parametri, come si vede nel Listing 2.3, viene rilasciato un pod MySQL nel cluster Minikube.

```
helm install mysql-inventory bitnami/mysql \
--set auth.rootPassword=root-pass \
--set auth.database=mysql-inventory \
--set auth.username=user \
--set auth.password=pass
```

Listing 2.3: MySQL Helm chart

## 2.3 Health

Durante la fase di deployment, e in particolare durante il deployment automizzato con *deployment.sh*, è stato notato che il servizio *reservation-service* non funzionava correttamente. In particolare il pod partiva, andava in stato *running* e *ready*, ma non si riusciva a connettersi al suo database PostgreSQL. Dopo numerose e lunghe investigazioni, analizzando i log dei pod, il problema è stato risolto eliminando manualmente il pod: Kubernetes riavvia una nuova istanza e i problemi di connessione vengono risolti.

Questo ci può portare ad affermare che, per qualche motivo, il pod necessitava che il database PostgreSQL sia disponibile e pronto per accettare connessioni prima di avviarsi correttamente.

### 2.3.1 MicroProfile Health

Per evitare di riavviare il pod manualmente o di introdurre nello script del deployment una *wait* è stato deciso di forzare il pod relativo a *reservation-service* ad attendere la disponibilità del database prima di avvarsi.

Questo viene fatto tramite la specifica Health di MicroProfile implementata poi da SmallRye; essa estende le specifiche di Jakarta per adattare quest'ultima all'architettura dei micro servizi. Permette di esporre lo stato di un'applicazione, cioè un valore binario (a differenza delle metriche che sono valori numerici qualunque), che viene solitamente controllato dall'orchestratore che gestisce il ciclo di vita delle applicazioni (e.g., Kubernetes) per decidere se un'applicazione è sana e può essere usata.

I tre maggiori controlli di salute che MicroProfile mette a disposizione sono *Startup*, *Readiness* e *Liveness* [9]: la prima verifica che l'applicazione sia stata avviata correttamente; la seconda controlla se l'applicazione è pronta a ricevere traffico; la terza monitora lo stato dell'applicazione durante il suo funzionamento. Se uno di questi controlli fallisce solitamente il pod viene riavviato.

### 2.3.2 Implementazione

Considerando che il nostro problema viene risolto facendo ripartire il pod del micro servizio una volta che il database è disponibile, l’Health Check utile è lo *StartUp Check*. È stata definita la classe **DatabaseConnectionHealthCheck.java**, annotata con l’annotazione `org.eclipse.microprofile.health.Startup`, che implementa una SELECT sul database: se il comando ha successo allora il check termina con successo e il pod è in esecuzione correttamente, altrimenti il pod viene fatto ripartire.

Nella classe la connessione viene gestita da un oggetto di tipo `AgroalDataSource`[12]: `DataSource` è un’interfaccia che rappresenta una connessione a un database standard in Java (JDBC); `Agroal` è un’implementazione di questa interfaccia ed è quella che funziona meglio con Quarkus.

Quarkus permette con molta facilità di gestire le metriche di Health con le apposite estensioni [13]. Basta infatti aggiungere `quarkus-smallrye-health`, `quarkus-agroal` e `quarkus-jdbc-postgresql` e il tutto funiona completamente. L’unica cosa aggiuntiva a cui dobbiamo pensare è a fornire l’url a cui l’oggetto `AgroalDataSource` deve connettersi; questo è facilmente configurabile nell’`application.properties`, come si vede dal Listing 2.4. È stata aggiunta anche la configurazione [11] per gestire dopo quanti tentativi falliti riavviare il pod.

```
# health
quarkus.kubernetes.startup-probe.failure-threshold=1
quarkus.datasource.jdbc.url=jdbc:postgresql://postgresql-
reservation:5432/reservation
```

Listing 2.4: JDBC configuration.

# 3 Tracing and Metrics

## 3.1 Tracing

In applicazioni con migliaia di micro servizi, seguire il workflow, cioè la cascata di chiamate, o valutare in che punto si è verificato un errore può essere molto complesso. Tramite il tracing non solo possiamo capire quale micro servizio viene chiamato e da chi, ma possiamo anche analizzare i tempi con cui ogni richiesta è stata servita, sia E2E che all'interno del singolo servizio.

Gli elementi principali che compongono un sistema di tracing sono: ogni singola unità di lavoro (i.e., chiamata a un microservizio) rilevata dal sistema è detta *span*; l'intera operazione E2E è composta da più span ed è detta *traccia*; ad ogni span è associato un ID univoco che rappresenta la traccia a cui appartiene; i diversi span che compongono una traccia hanno una relazione gerarchia, per cui si parla di span padre e/o figlio. Il funzionamento solitamente è abbastanza semplice: durante le varie chiamate nel workflow viene trasmesso l'ID della traccia; il trasporto di questo ID avviene tramite qualche funzionalità del protocollo di trasporto, che in HTTP è il relativo header; quando la chiamata ritorna allora vengono unite gli span con lo stesso ID e viene generata la traccia.

### 3.1.1 OpenTelemetry

OpenTelemetry, noto anche come OTel, è un insieme di tool e API che permette di collezionare ed esportare le telemetrie di un'applicazione a microservizi. Oltre alle telemetrie in realtà possono anche essere gestite metriche e logs, ma in questo campo non è stabile e si preferisce usare altro (e.g., MicroMeter).

La vera forza di OpenTelemetry però non è tanto la raccolta di tracce, quando la definizione di OTLP, un protocollo standard che permette di esportare le telemetrie verso vari e diversi backend (e.g., Jaeger per la visualizzazione).

Per permettere il funzionamento di Opentelemetry con un qualche backend ci sono due possibili:

- Direttamente: OpenTelemetry una volta raccolte le tracce le invia al backend in questione. È una strategia più semplice, ma non scala quando i backend aumentano e se dobbiamo fare un filtraggio delle tracce.
- Collector: I dati vanno all'Otel-Collector, che è il responsabile dell'invio e di operazioni intermedie (e.g., filtraggio, il batching e la gestione dei fallimenti). È la strategia più complesso, però in applicazioni che devono scalare è necessario.

### 3.1.2 Collector

Nella nostra applicazione è stato usato il collector perché si voleva inviare le tracce al backend di Jaeger e anche salvarle in un file per eseguire ulteriori analisi in un secondo momento.

La raccolta delle tracce con Quarkus diventa estremamente semplice: aggiungendo l'estensione *quarkus-opentelemetry* il framework instrumenta il codice e in automatico le telemetrie vengono collezionate. Una volta che le tracce sono raccolte vengono inviate al collector, dopo la configurazione ne Listing 3.1 (relativo a *users-service*), dove l'unica configurazione da spiegare è forse *quarkus.otel.traces.sampler=always\_on*: essa permette di non fare nessun campionamento delle chiamate, ma di considerarle tutte.

Il collector è un pod all'interno dell'ambiente Minikube che è stato deployato tramite una **configurazione helm**. Quando riceve le tracce le invia a due destinazioni: il pod di Jaeger per la visualizzazione e il file */traces/traces.json* che si trova nel PVC (Persistant Volume Claim) *pvc-traces* sempre all'interno del cluster. Quando eseguiamo l'analisi dei dati allora dovremmo eseguire l'analyser in un pod a cui è collegato questo PVC; in questo modo lo script esegue sui propri dati locali.

```
# jaeger
quarkus.otel.service.name=users-service
quarkus.otel.exporter.otlp.traces.endpoint=http://otel-
    collector:4317
quarkus.otel.traces.sampler=always_on
```

Listing 3.1: Jaeger configuration for *users-service*

### 3.1.3 Jaeger

Come abbiamo detto Jaeger è un backend per la visualizzazione del work di una chiamata in un'applicazione a microservizi.

Come per i servizi terzi elencati nella Sezione 1.2, anche il backend Jaeger deve essere rilasciato nello stesso cluster Minikube di tutta l'applicazione. Anche in questo caso è stato utilizzato il chart Helm Jaeger [8]. Si può vedere il comando nel Listing 3.2.

```
helm install jaeger jaegertracing/jaeger \
--set allInOne.enabled=true \
--set agent.enabled=false \
--set collector.enabled=false \
--set query.enabled=false \
--set provisionDataStore.cassandra=false \
--set storage.type=memory
```

Listing 3.2: Install Jaeger chart for *users-service*

### 3.1.4 OTel workflow

Fino a questo momento in questa sezione abbiamo parlato dei componenti in modo separato; vediamo adesso di mettere tutto insieme per capire come i vari pod interagiscono nella nostra applicazione.

Analizziamo il flusso mostrato in Figura 3.1:

- Viene chiamata una qualche API dell'applicazione.
- OpenTelemetry raccoglie le telemetrie e invia le tracce al collector, che eventualmente fa qualche operazione.
- Il collector invia le tracce ai vari backend. Nel nostro caso Jaeger per la visualizzazione e il file system per l'analisi successiva.

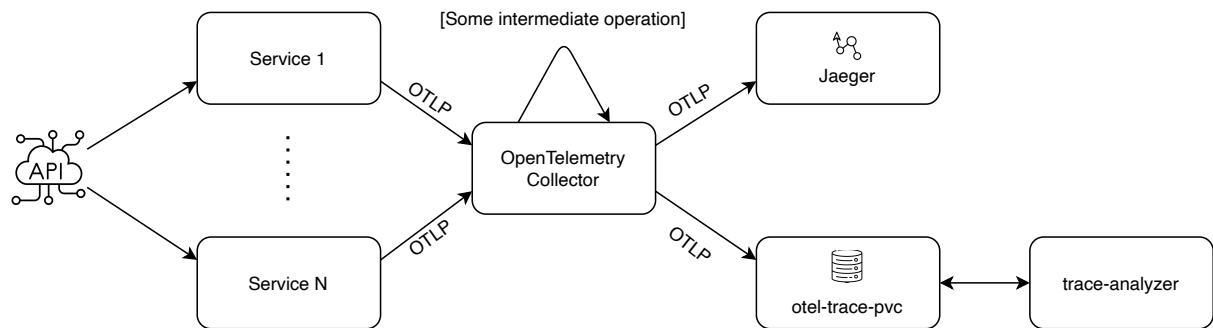


Figura 3.1: OpenTelemetry workflow.

## 3.2 Metrics

Oltre a tracciare la sequenza delle varie chiamate, come abbiamo mostrato nel Capitolo 3.1, in un applicazione a micro servizi è importante anche essere in grado di catturare metriche, valori reali che espimono una qualche informazione.

In generale possiamo esporre metriche framework o business: le prime sono standard fornite da un qualche framework e sono generali e valide in un qualunque ambito (e.g., carico della cpu, uso della memoria o richieste per secondo); le seconde sono metriche specifiche per il dominio dell'applicazione.

La specifica MicroProfile per le metriche si chiama Metrics. Nonostante questo solitamente viene usata MicroMeter, un'implementazione che non segue la specifica MicroProfile. Quarkus permette in ogni caso di usare sia implemetazioni della specifica Metrics che MicroMeter, nonostante la seconda sia quella preferibile.

### 3.2.1 Prometheus

Promethues è un sistema open-source per il monitoraggio e allerta di applicazioni. Svolge tre compiti principali: raccoglie metriche, le memorizza e fornisce un endpoint utile per fare querying.

In Quarkus utilizzare MicroMetrics e Prometheus insieme è molto semplice: dobbiamo solo aggiungere l'estensione `quarkus-micrometer-registry-prometheus`: senza strumentalizzare il codice, MicroMeter inizia a raccogliere metriche standard (e.g., utilizzo CPU e della memoria) nel formato richiesto da Prometheus, mentre quest'ultimo periodicamente interroga gli endpoint creati, raccogliendo e salvando tali metriche.

Dobbiamo a questo punto definire un modo per dire a Prometheus di raccogliere le metriche che MicroMetrics definisce. In Kubernetes questo metodo è detto `ServiceMonitor` e può essere semplicemente implementato nel file di configurazione, come vediamo nel Listing 3.3

```
# prometheus and grafana
quarkus.kubernetes.prometheus.generate-service-monitor=true
quarkus.kubernetes.labels.release=prometheus
```

Listing 3.3: Prometheus configuration for `users-service`

Infine serve installare nel cluster Minikube un container in cui esegue il server Prometheus. Questo è stato fatto, come nel resto del progetto, tramite Helm e il suo chart c. Questo non solo installa Prometheus, ma è un singolo pacchetto che comprende anche Grafana.

### 3.2.2 Grafana

Il sistema di query di Promethues permette di visualizzare le metriche raccolte in forma testuale tramite terminale. Quando le metriche sono complesse, ma anche in generale, questo può non essere molto comodo.

Grafana è un backend che permette di visualizzare in modo grafico le metriche raccolte da Promethues, come si può vedere nella Figura 3.2

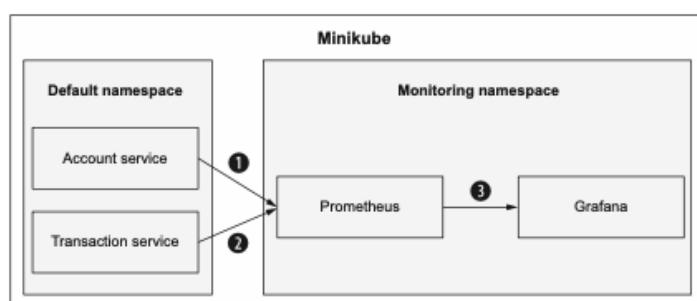


Figura 3.2: Prometheus-Grafana Flow.

Come dette nella Sezione 3.2.1 l’installazione all’interno del cluster Minikube è stata fatta con il chart di Helm *kube-prometheus-stack* [6], che comprende sia Prometheus che Grafana.

# 4 Workflow

Nei Capitoli 2 e 3 abbiamo visto come l'applicazione car rental, descritta nel Capitolo 1, viene rilasciata su Minikube e come possiamo analizzare metriche e workflow. In questo capitolo ci concentriamo invece sul workflow, cioè su come i microservizi si chiamano durante l'esecuzione di una funzionalità; in particolare la funzionalità che andremo ad esaminare sarà la prenotazione di una macchina.

Prima di qualunque modifica, come si nota dalla Figura 4.1, *Car-Rental* aveva un workflow esclusivamente sequenziale: quando un utente invocava una qualunque operazione, e in particolare `reserve`, tramite `users-service`, la sequenza di chiamate agli altri microservizi era puramente sincrona e deterministica, cioè era sempre la stessa data l'operazione chiamata; ovviamente i tempi di esecuzione variano di chiamata in chiamata.

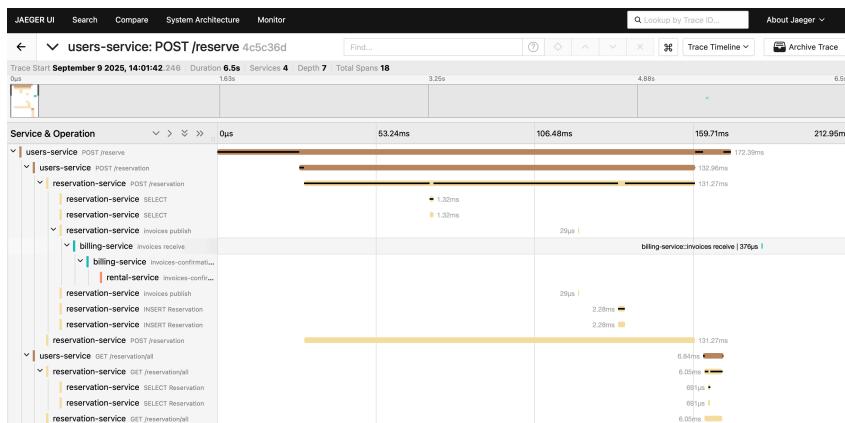


Figura 4.1: Workflow iniziale di Car-Rental.

Il nostro obiettivo è, facendo riferimento sempre all'operazione `reserve`, complicare questo workflow. Facendo riferimento alla terminologia usata nell'articolo del tool Euler [5], che useremo più avanti, l'applicazione prima della modifica combina solamente blocchi di tipo sequenziale (SEQ), mentre a noi interessa aggiungere degli blocchi XOR e blocchi AND, che sono descritti nella Sezione 4.1 e implementati nella Sezione 4.2.

## 4.1 Blocchi

Vediamo come prima cosa quali blocchi sono stati aggiunti e in che punto dell'attuale workflow sono stati eseguiti.

### 4.1.1 XOR

I blocchi XOR rappresentano scelte casuali esclusive definite da una certa probabilità. Questo vuol dire che se abbiamo  $n$  possibili strade che il flusso può prendere, ne sarà scelta una e una soltanto una in base alla probabilità assegnata.

### 4.1.2 AND

I blocchi AND rappresentano esecuzioni indipendenti che possono essere eseguite in parallelo. Questo vuol dire che se abbiamo  $n$  strade, tutte saranno eseguite contemporaneamente e si sincronizzeranno alla fine.

## 4.2 Implementazione

Adesso che abbiamo visto quali blocchi vogliamo aggiungere, vediamo come sono stati implementati e in che punto del workflow vengono eseguiti.

La cosa importante da osservare è che l'obiettivo di questa modifica non è quello di aggiungere funzionalità all'applicazione, ma solamente quello di complicare il workflow. Per questo motivo ogni blocco aggiunto sarà composto da microservizi che non fanno altro che eseguire una busy wait per un tempo campionato da una distribuzione esponenziale; la busy wait è implementata con un ciclo `while` che controlla il tempo trascorso.

### 4.2.1 XOR

Il blocco XOR, che si trova nella cartella **XOR** su GitHub, implementa la scelta esclusiva tra tre microservizi. Questo blocco è inserito nel flusso di *users-service*: quest'ultimo microservizio chiama quello responsabile della scelta randomica e una volta che il blocco XOR termina riprende il controllo; in questo modo al codice dell'applicazione precedente è stata solo aggiunta la chiamata a questo blocco.

Il flusso del blocco XOR implementato è il seguente:

- L'inizio del blocco è definito da *start-choice-service*. Questo esegue una busy wait per un tempo campionato da una distribuzione esponenziale con tasso 1. Esegue poi una scelta randomica tra tre microservizio, dove: il primo ha probabilità 0.2, il secondo 0.5 e il terzo 0.3. Questo momento è utile per osservare un altro vantaggio di Quarkus: le probabilità con cui sono definite le scelte non sono build-in nel codice, ma sono definite del file *application.properties*; questo permette di modificarle senza dover mettere mano al codice stesso.
- I tre microservizi chiamati sono *first-choice-service*, *second-choice-service* e *third-choice-service*. Ognuno di questi esegue una busy wait per un tempo campionato

da una distribuzione esponenziale con tasso 1. Al termine di uno di questi il controllo ritorna a *users-service*

#### 4.2.2 AND

Il blocco AND, che si trova nella cartella **AND** su GitHub, implementa l'esecuzione parallela di due servizi chiamati dallo stesso microservizio client; i due server eseguono contemporaneamente e si sincronizzano nel client.

Anche questo blocco è inserito nel flusso di *users-service*, esattamente dopo la chiamata al servizio *start-choice*: il microservizio chiama *start-parallel* che è il responsabile del fork; una volta che il blocco AND termina riprende il controllo; in questo modo al codice dell'applicazione precedente è stata solo aggiunta la chiamata a questo blocco.

Per l'implementazione del blocco AND è stato usato gRPC, già descritto nella Sezione 1.3. In questo caso ha fatto la differenza l'utilizzo del framework Quarkus, che fornisce l'integrazione nativa con gRPC e permette di definire servizi asincroni in modo semplice.

La pipeline usata per implementare la comunicazione asincrona di gRPC con Quarkus è la seguente:

- Il server dichiara un file *.proto*, che definisce i servizi offerti e i messaggi scambiati. Questo permette di lavorare in modo API-first: si definisce cosa il server espone e si lavora a partire da questo.
- Con l'estensione *quarkus-grpc*, quando si chiama *quarkus build*, il framework genera automaticamente tutto quello che ci serve per implementare il server. In particolare la classe che espone il servizio deve implementare un'interfaccia e sovrascrivere i metodi che nel file *.proto* sono i messaggi scambiati.
- Il client deve memorizzare il file *.proto*, compilarlo in modo che Quarkus generi lo stub del server e intrumentare il codice in modo da poter chiamare i metodi esposti dal server come se fossero locali.

gRPC consente quindi di implementare una comunicazione asincrona tra client e server tramite canali asincroni. Questo si può fare usando l'interfaccia nativa di gRPC o le API di SmallRye.

In gRPC Java classico questi canali sono gestiti dall'interfaccia *StreamObserver<T>*, dove T è il tipo del messaggio; ogni comunicazione è quindi un oggetto che implementa tale interfaccia. Quando si invia un messaggio tramite il canale (usando il metodo *onNext(T message)*), gRPC avvia un thread del suo pool di thread dedicati, il quale si occupa della gestione. Questa interfaccia ha però numerosi limiti: è imperativa e quindi tutto deve essere gestito a mano; si basa su callback, cioè molte funzioni

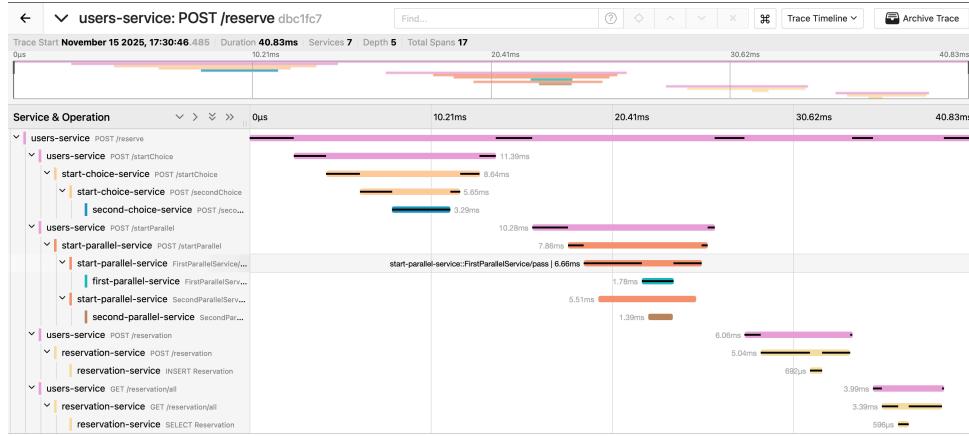


Figura 4.2: Nuovo workflow di reserve mostrato da Jaeger.

vengono passate come parametro (uno *StreamObserver* deve implementare tre metodi) e saranno chiamate in un secondo momento; è difficile da testare; non permette di gestire più risponde come un unico oggetto.

Quarkus, ovviamente, non utilizza questa interfaccia, ma sfrutta quella di SmallRye, *Mutiny*. Si tratta di una libreria reattiva e dichiarativa: si deve definire come il sistema si comporta quando i dati arriveranno; non dobbiamo dire come fare le cose ma solo cosa vogliamo ottenere (e.g., nello stesso modo delle lambda). L'interfaccia di *Mutiny* si basa su due tipi, *Uni<T>* e *Multi<T>*: il primo permette l'invio di un solo messaggio, mentre il secondo consente gli stream. I vantaggi che si ottengono sono: codice più leggibile; composizione di chiamate diverse facile da implementare; facile da testare e debuggare.

## 4.3 Nuovo workflow

Nel nuovo workflow, oltre ad includere i blocchi XOR e AND, è stata eliminata la parte di fatturazione. Il motivo di questa scelta è in dovuto al nostro obiettivo che ci ha portato a complicare questo workflow: vogliamo valutare il tempo E2E di una prenotazione. La prenotazione e la fatturazione sono in car-rental due operazioni indipendenti: una prenotazione ha successo, e quindi ritorna all'utente, indipendentemente dal completamento della chiamata di *reservation-service* a *billing-service*.

Usando ancora Jeager per visualizzare la traccia della chiamata a *reserve*, possiamo notare dalla Figura 4.2 la presenza dei blocchi XOR e AND e l'esclusione dei microservizi *billing-service* e *rental-service*. In Figura 4.3 è mostrato invece il sequence diagram UML.

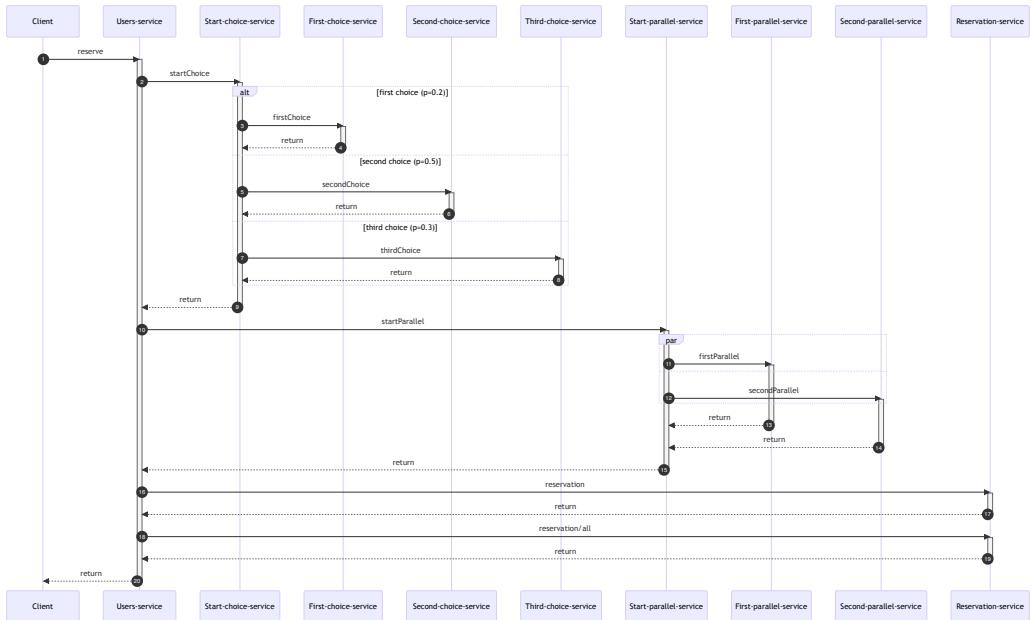


Figura 4.3: Sequence diagram dei blocchi XOR e AND.

# 5 Analisi E2E

In questo capitolo andremo ad eseguire un’analisi del tempo end to end (E2E) di una prenotazione da parte dell’utente, cioè dell’intervallo di tempo che trascorre dall’inizio della chiamata della funzionalità fino al suo termine.

Nei casi in cui si voglia garantire un tempo di risposta medio di una funzionalità, l’analisi E2E può essere fatta semplicemente osservando media e varianza dei tempi di esecuzione dei servizi che compongono tale funzionalità. Ci sono delle situazioni in cui però questo non è sufficiente e si vuole garantire un determinato Service Level Agreements (SLA), cioè si vuole lavorare con un upper-bound sul tempo di esecuzione. Il nostro obiettivo è quindi più stringente rispetto al tempo medio: ogni chiamata a una qualche funzionalità deve sempre ritornare entro un tempo all’interno di una soglia. Questo richiede non solo di avere dati statistici descrittivi, ma di essere in possesso dell’intera Probability Density Function (PDF) o della Cumulative Density Function (CDF) del workflow che ci interessa analizzare.

Un esempio di metrica che possiamo ottenere seguendo questo approccio e basandoci sulla PDF è la probabilità che il tempo di esecuzione sfiori una data deadline  $T_{max}$ : semplicemente si ottiene integrando la PDF in  $[T_{max}, +\infty]$ . Questo è un buon esempio di osservazione che non potremmo fare con le sole statistiche descrittive.

## 5.1 Eulero

Per ottenere una distribuzione del tempo E2E di un workflow ci sono due tecniche: la prima è un’analisi numerica esatta, fattibile quando il workflow non è troppo complesso e i tempi di esecuzione seguono distribuzioni markoviane (i.e., esponenziali); la seconda è un’approssimazione basata su un approccio compostionale, che permette di analizzare anche workflow complessi con distribuzioni generali (GEN). Eulero utilizza questo secondo approccio bottom-up.

### 5.1.1 Workflow complessi

Quando un workflow è complesso dobbiamo ricorrere a tecniche approssimate. La complessità di un workflow nella vita reale deriva dal fatto che i tempi di esecuzione dei microservizi non seguono distribuzioni esponenziali, cioè markoviane.

Quando i tempi di esecuzione sono espostizionali, l’unica cosa che condiziona il futuro del sistema è lo stato attuale. I possibili percorsi che il sistema può prendere da un certo punto in poi non considerano il passato e non considerano il tempo trascorso in quello stato. Quando però introduciamo distribuzioni non markoviane, il futuro

viene condizionato anche dal passato. Questo provoca un’esplosione dello spazio degli stati e rende l’analisi numerica esatta impossibile.

### 5.1.2 Tool e analisi bottom-up

Visto quanto detto sopra, per ottenere la PDF del tempo di esecuzione di un servizio abbiamo usato Eulero<sup>[5]</sup>, un tool dell’SILAB dell’università di Firenze che esegue un’analisi bottom-up tramite un metodo compositivo: si divide il workflow in sub-workflow, fino ad unità elementari; si calcolano le PDF di queste unità; si compongono le varie PDF per ottenere la PDF del workflow completo.

Il vantaggio di questo approccio è la riduzione della complessità: si elimina infatti il condizionamento del futuro dagli stati precedenti anche per distribuzione GEN e questo porta a una riduzione del numero degli stati che devono essere analizzati. Il lato negativo è che si introduce un’errore di approssimazione nel tempo E2E calcolato.

### 5.1.3 Dati necessari

Senza scendere nel dettaglio, Eulero supporta workflow con strutture di controllo sequenziali (SEQ), parallele (AND), condizionali (XOR) e modellate con Direct Acyclic Graph (DAG). Le strutture parallele e condizionali sono quelle descritte nella Sezione 4.1 e che sono state poi implementate in car-rental.

Quello che dobbiamo fornire al tool per ottenere la distribuzione del tempo E2E è: la struttura del workflow modellata tramite blocchi, eventualmente annidati; la distribuzione marginale del tempo di esecuzione di ogni blocco elementare. Osserviamo che i blocchi e la loro composizione possono essere osservati dalla Figura 4.2 e dalla Figura 4.3; le distribuzioni sono invece ottenute nella Sezione 5.3.

## 5.2 Tracce

Nella Sezione 3.1 abbiamo descritto il funzionamento del collector di OpenTelemetry e come questo lavora all’interno della nostra applicazione. In questo momento ci interessa lavorare sulle tracce che il collector salva nel Persistent Volume Claim (PVC) *otel-traces-pvc* e che sono accedibili dal pod *analyzer*; l’obiettivo è eseguire molti esperimenti, ottenere i tempi di esecuzione delle chiamate ai microservizi che poi possiamo usare per estrarre i dati necessari per Eulero.

### 5.2.1 Generazione

Per il nostro obiettivo ci servono molti dati e quindi molte tracce da analizzare. Siccome generare queste tracce a mano non è molto pratico, è stato usato K6 come generatore di carico.

Il generatore di carico K6 può essere installato localmente su una macchina Linux, MacOS o Windows, ma ha anche il grande vantaggio di poter essere eseguito all'interno di un'infrastruttura cloud. Nel nostro progetto è utile e sensato distribuire K6 all'interno del cluster Minikube dove è in esecuzione la nostra applicazione; in questo modo può chiamare ogni risorsa di ogni servizio senza esporre tale servizio all'esterno del cluster. Per distribuire K6 in un cluster Kubernetes ci diversi modi, ma nel nostro caso è stato usato, come in gran parte del progetto, Helm. Su GitHub è possibile vedere l'installazione del [job di K6](#): helm utilizza la cartella *templates/* per generare le risorse Kubernetes; il file *templates/configmap.yaml* serve per iniettare lo script che il job dovrà eseguire; il file *job.yaml* crea una Job Kubernetes che esegue tale script.

L'implementazione questo job è molto semplice: logicamente si tratta di un ciclo `for` che esegue per un certo intervallo di tempo specificato chiamando l'entry point del servizio di prenotazione. Prenota sempre la stessa macchina spostando l'intervallo della prenotazione in avanti di un giorno a ogni iterazione.

### 5.2.2 Elaborazione

Le tracce sono salvata in un singolo file *traces.json*, che all'interno del pod *analyzer* si trova nella directory *data*. Si tratta di un JSON Lines, cioè un file dove ogni linea è un oggetto JSON indipendente. L'elaborazione avviene portando i dati dal pod in memoria locale e poi eseguendo uno script costruito in modo da leggere ogni riga del file e analizzarla come json. L'output dell'analisi contiene delle statistiche descrittive (i.e., minimo, massimo, media, varianza e coefficiente di variazione) dei tempi di esecuzione di ogni microservizio. Queste statistiche saranno poi necessarie per ottenere le distribuzioni del tempo di esecuzione di ogni blocco del workflow.

Grazie a una caratteristica di OpenTelemetry è possibile anche ottenere la latenza di rete tra le varie chiaate. Per ogni microservizio viene raccolto sia il tempo *server* che *client*: il primo misura il tempo in cui il servizio è in esecuzione; il secondo misura il tempo che ci mette una chiamata a un altro servizio a tornare. Combinando insieme queste due informazioni, per due servizi consecutivi, si può ottenere il tempo in cui l'informazione è stata trasmessa.

### 5.2.3 Risultati

Possiamo osservare tutti le informazioni raccolte in [extract\\_data/result](#). Come si nota sono disponibili, per ogni microservizio, i tempi di server (i.e., per quanto il servizio esegue), i tempi di ogni chiamata e la latenza di rete.

## 5.3 Distribuzioni

Per eseguire Eulero e ottenere la nostra PDF (o CDF) del tempo E2E del workflow ci servono le distribuzioni dei tempi di esecuzione di ogni blocco di tale workflow. Al momento, come descritto nella Sezione 5.2, abbiamo le statistiche descrittive dei tempi di esecuzione di ogni operazione che compone il workflow. Dobbiamo a questo punto utilizzare questi dati per ottenere delle distribuzioni che abbiano il miglior fitting sui nostri dati osservati tramite gli esperimenti.

### 5.3.1 Approssimanti di Whitt

Se abbiamo una distribuzione del tempo di esecuzione di un servizio, allora possiamo ottenere i dati semplicemente campionando da tale distribuzione. La nostra situazione è però l'opposta: abbiamo i tempi di esecuzione e ci serve una distribuzione che fitti bene i dati.

Una prima approssimazione può essere usando una distribuzione esponenziale, che inoltre, come abbiamo detto nella Sezione 5.1.1, ci semplificherebbe la vita da un punto di vista dell'analisi. Sappiamo che una distribuzione esponenziale ha un coefficiente di variazione ( $CV$ ) pari a 1; questo però nella realtà non succede (praticamente) mai. I coefficienti di variazione sono tutti  $CV < 1$  e questo implica che i nostri microservizi hanno una varianza maggiore di quella di una distribuzione esponenziale. In queste situazioni per trovare una distribuzione che approssimi bene i nostri dati si possono usare le approssimazioni di Whitt. I casi più comuni e che solitamente si usano sono:

- HyperExp

Si usa per dati che hanno un coefficiente di variazione  $C_V > 1$ . Si tratta di uno XOR di due esponenziali, cioè di una distribuzione definita come  $X = \begin{cases} Exp(\lambda_1) & \text{con prob } p \\ Exp(\lambda_2) & \text{con prob } 1 - p \end{cases}$ ; osserviamo che se  $\lambda_1 = \lambda_2$  allora abbiamo un'esponenziale.

- HypoExp

Si usa per dati che hanno un coefficiente di variazione  $C_V \in [\frac{1}{\sqrt{2}}, 1]$ . Si tratta di una sequenza di esponenziali con tassi diversi, cioè di una distribuzione definita  $X = Exp(\lambda_1) + \dots + Exp(\lambda_n)$ , con  $\lambda_i \neq \lambda_j$ .

- Erlang

Si usa per dati che hanno un coefficiente di variazione  $C_V = \frac{1}{\sqrt{k}}$ ,  $\exists k$ . Si tratta di una sequenza di  $k$  esponenziali con tassi uguali, cioè di una distribuzione definita  $X = k \cdot Exp(\lambda)$ .

- GeneralizedErlang

Si usa per dati che hanno un coefficiente di variazione  $C_V < \frac{1}{\sqrt{2}}$  [1]. Si tratta

di una sequenza di  $k$  esponenziali con tasso  $\lambda_1$  seguita da un'esponenziale con tasso  $\lambda_2$ , cioè di una distribuzione definita  $X = k \cdot \text{Exp}(\lambda_1) + \text{Exp}(\lambda_2)$ . Questo permette di fissare dati che hanno un  $CV \in (\frac{1}{\sqrt{k+1}}, \frac{1}{\sqrt{k}})$ .

### 5.3.2 Risultati

I risultati, cioè le distribuzioni che meglio fissano i dati ottenuti in precedenza, possono essere calcolate in funzione del valore atteso e del coefficiente di variazione. I risultati si trovano nel file `trace_analysis_stats.json` presente su GitHub.

A questo punto abbiamo tutto per eseguire l'analisi finale e calcolare la distribuzione dei tempi E2E del nostro workflow.

## 5.4 Distribuzione E2E

A questo punto abbiamo il workflow con i microservizi che vengono chiamati, mostrati in Figura 4.3, e le distribuzioni dei tempi di esecuzione di ogni microservizio: possiamo eseguire la simulazione su Eulero, che modella il nostro workflow con qualcosa di simile a quello che possiamo osservare in Figura 5.1.

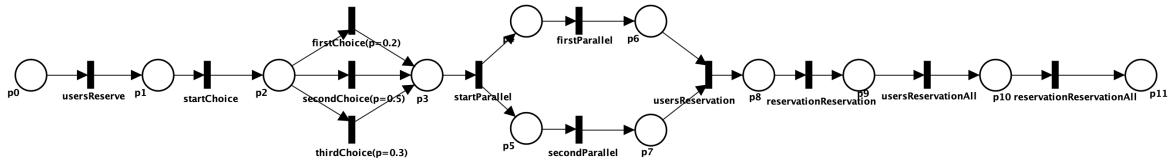
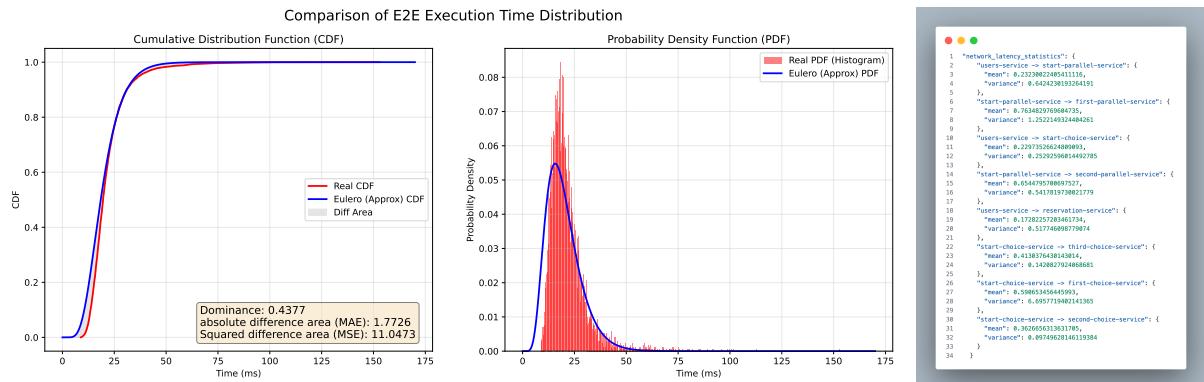


Figura 5.1: STPN del workflow [10].

In Figura 5.2a possiamo osservare il risultato ottenuto. È interessante mostrare anche, in Figura 5.2b le statistiche (i.e., media e varianza) della latenza di rete che c'è nelle varie chiamate.



(a) Eulero and real distribution.

(b) Latenza di rete.

Figura 5.2: Risultati dell'esperimento.

Per valutare la differenza delle due CDF ottenute sono state usate due misure:

- Dominance: Si tratta di una misura che quantifica la probabilità che una random variable  $X$  sia minore di una random variable  $Y$ , cioè  $\delta(X, Y) = P\{X \leq Y\}$ : se  $\delta(X, Y) \geq \frac{1}{2}$  allora  $X$  arriva dopo  $Y$  [4].
- Difference area: Per ogni punto si calcola il valore assoluto della differenza tra la vera CDF e quella ottenuta da Eulero. Il valore è dato dalla somma di queste differenze. È stato calcolato sia l'errore L1 (assoluto) che quadratico.

## 5.5 Bande di confidenza

Quando eseguiamo degli esperimenti lavoriamo con i dati provenienti dalle nostre simulazioni, ma non con la loro vera distribuzione: possiamo definire la distribuzione solo in modo campionario e non assoluto. Per trattare in modo statistico i dati e trarre precise conferme (e.g., in casi di performance analysis) ci serve capire quanto le distribuzioni campionarie sono precise e quanto possiamo fare affidamento su di esse. Un modo per fare questo è usare le bande di confidenza. Possiamo in questo modo verificare, con una fissata confidenza di  $1 - \alpha$ , quanto è accurata la distribuzione ottenuta e se è necessario aumentare i campioni dell'esperimento perché vogliamo una maggiore accuratezza.

Per ricavare queste bande di confidenza ci sono solitamente due approcci:

- Pointwise band

Le bande di confidenza sono definite analizzando ogni punto della CDF in modo isolato. Sono tecniche valide quando ci interessa capire l'affidabilità di valore specifico (e.g., quale è la probabilità che il mio esperimento restituisca un valore minore di  $x$ ), oppure quando si vuole analizzare i percentili. In questi contesti sono meno conservative e quindi più precise.

- Simultaneus band

Fornisco delle bande probabilistiche di confidenza della CDF della distribuzione da cui derivano i dati e garantiscono che la vera curva della CDF resti all'interno delle bande con confidenza  $1 - \alpha$ . Permettono quindi di ottenere delle garanzie sull'intera popolazione senza sapere nulla della vera distribuzione.

### 5.5.1 Tecniche

Le tecniche più comuni sono:

- Approssimazione normale

Si tratta della tecnica base più citata in letteratura [16]. Utilizza la distribuzione binomiale per determinare gli intervallie di confidenza di un punto della CDF. Il

metodo approssima la distribuzione a una gaussiana; per una binomiale questo vuol dire supporre che il numero di esperimenti tenda a infinito o che la probabilità sia  $p = 0.5$ . Questo fa sì che questa tecnica sia buona quando il numero di dati è molto elevato.

I problemi principali che si possono incontrare sono:

- Copertura a salti

Se chiediamo un grado di confidenza di  $1 - \alpha\%$  può succedere che per alcuni punti questo sia rispettato, per altri più alto e per altri ancora più basso.

- Asimmetria

Per eventi che hanno una probabilità molto bassa o molto alta la distribuzione diventa molto asimmetrica, ma i metodi standard si assumono che essa sia una campana simmetrica.

- Numero di successi limite

Se osserviamo valore di  $k = 0$  o  $K = N$  allora otteniamo i metodi standard portano a divisioni per zero e altri problemi.

- Clopper-Pearson

Si basa sulla distribuzione binomiale e risolve i problemi base che si hanno con l'approssimazione normale o tecniche simili [7]. L'idea è cercare tutti i valori di  $p$  per cui le formule sotto funzionano. Questo porta a: avere costo computazionale alto, per cui si usa la relazione tra distribuzione binomiale e distribuzione beta; essere conservativo; garantire una copertura completa. Gli intervalli sono definiti da due valori:

- $p_{inf}$

È il valore di  $p$  della binomiale per cui la probabilità di osservare almeno  $k$  successi è esattamente  $\alpha/2$ , cioè  $\sum_{i=k}^n \binom{n}{i} p_{inf}^i (1 - p_{inf})^{n-i} = \frac{\alpha}{2}$ . Usando la distribuzione beta si ha poi  $B(\alpha/2, X, n - X + 1)$ .

- $p_{sup}$

È il valore di  $p$  della binomiale per cui la probabilità di osservare al più  $k$  successi è esattamente  $\alpha/2$ , cioè  $\sum_{i=0}^k \binom{n}{i} p_{sup}^i (1 - p_{sup})^{n-i} = \frac{\alpha}{2}$ . Usando la distribuzione beta si ha poi  $B(1 - \alpha/2, X + 1, n - X)$ .

- Dvoretzky-Kiefer-Wolfowitz (DKW)

È una tecnica che definisce delle bande parallele alla CDF campionaria di ampiezza costante  $\epsilon = \sqrt{\frac{\ln(2/\alpha)}{2n}}$  [17]. Siccome l'ampiezza è fissa in ogni punto della CDF, tende a essere troppo conservativa e quindi è una buona tecnica quando si lavora in contesti di sicurezza. La matematica ci dice che la vera CDF  $F(x)$  è in

$L(x) \leq F(x) \leq U(x)$ , dove  $U(x) = \min(F_n(x) + \epsilon, 1)$  e  $L(x) = \min(F_n(x) - \epsilon, 0)$ , con  $F_n(x)$  che è la CDF campionaria.

- Non-parametric Bootstrap

Si tratta di un approccio data-driven [3]: invece che usare formule teoriche che solitamente fanno assunzioni come il tipo di distribuzione o il numero di campioni infiniti, in questo caso si usano i dati per studiare la variabilità dei dati. L'idea è quella di campionare  $n$  elementi dalla distribuzione campionaria  $F$  in modo da definire una CDF empirica  $F^*$ ; se ripetiamo questo processo per  $N$  volte, per ogni valore  $x$  all'interno del range che i nostri dati assumono avremo molti valori di  $F^*(x)$ . A questo punto possiamo definire le bande di confidenza in due modi; in entrambi i casi la in entrambi i casi le bande ottenute sono meno conservative di quelle derivate dai metodi Pearson-Clopper e DKW.

- Pointwise

Se abbiamo intervallo di confidenza del 95% allora prendiamo il 2.5° percentile e il 97.5° percentile dei valori assunti da tutte le  $F^*(x)$  ottenute. Avremo in questo modo una banda di confidenza che varia in funzione di  $x$ .

- Simultaneus

Si condiera il valore massimo di  $|F(x) - F^*(x)|$  per ogni  $x$ . Si prendono poi i percentili di questi valori massimi. Avremo in questo modo una banda di confidenza di ampiezza costante.

### 5.5.2 Risultati

Il nostro numero di campioni è stato abbastanza elevato da produrre una CDF campionaria abbastanza precisa. I risultati, che mostrano questa precisione, sono ottenuti richiedendo una confidenza del 99%.

- Le bande ottenute con il metodo di Clopper-Pearson sono mostrate in Figura 5.3a.
- Le bande ottenute con il metodo DKW sono mostrate in Figura 5.3b.
- Le bande pointwise ottenute con il metodo del bootstrapping sono mostrate in Figura 5.3c.
- Le bande simultanee ottenute con il metodo del bootstrapping sono mostrate in Figura 5.3d.

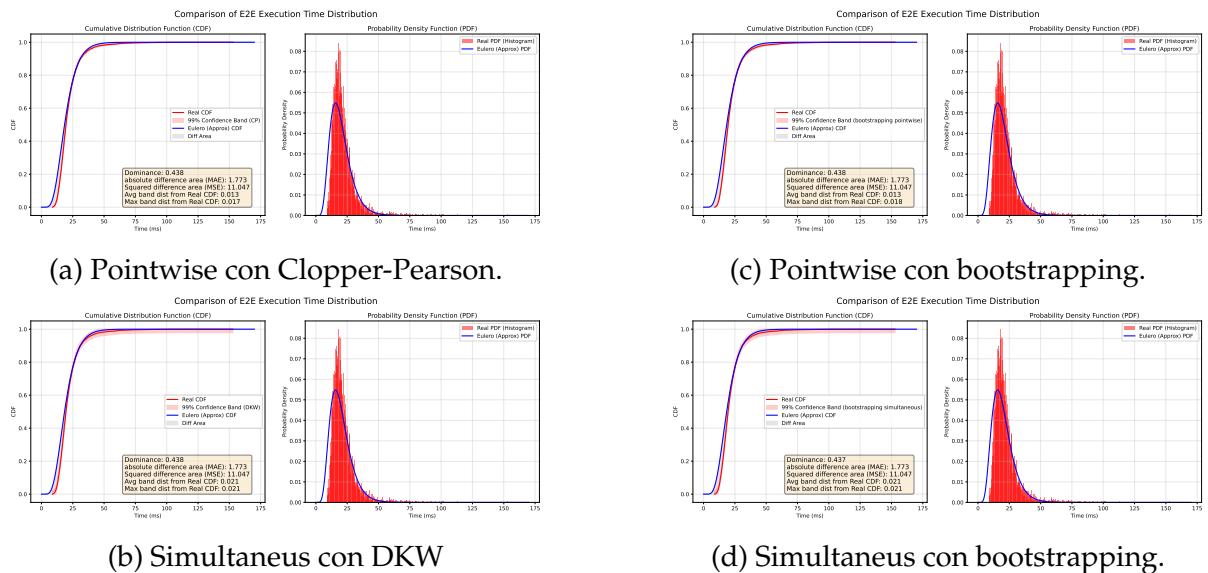


Figura 5.3: Bande di confidenza.

# 6 Correlazione

Fino a questo momento i nostri microservizi non avevano nessun tipo di correlazione: il tempo di esecuzione del microservizio  $A$  è totalmente indipendente dal tempo di esecuzione di  $B$ , per qualunque  $A$  e  $B$ . In questo capitolo andremo ad aggiungere correlazione tra qualche microservizio per cercare di dimostrare un aspetto tipico della tecnica di analisi composita di Eulero: siccome le distribuzioni (i.e., PDF e CDF) del tempo di esecuzione E2E sono costruite con un approccio bottom-up, cioè compонendo le distribuzioni marginali fino ad arrivare al risultato, il modello perde tutte le correlazioni che ci possono essere; quando i tempi di esecuzione sono in un qualche modo correlati, allora il sistema di analisi dovrebbe funzionare peggio, cioè l'area tra le due curve sovrrebbe aumentare.

Per capire il livello di correlazione iniziale possiamo osservare la Figura 6.1: notiamo come tutti i servizi hanno una correlazione a due a due molto bassa (i.e., vicina a 0).

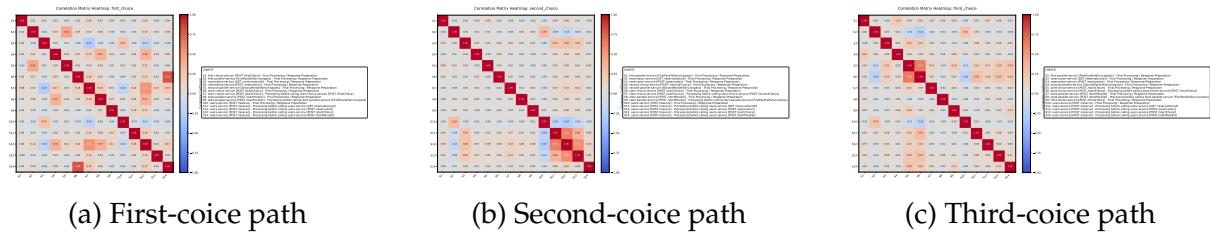


Figura 6.1: Correlazioni iniziali, i.e. senza l'aggiunta esplicita nel codice.

È importante osservare che il numero di tracce dei tre esperimenti, senza correlazione, con basse e con alta correlazione, sono eseguiti su un numero di tracce simile, circa 5500.

## 6.1 Implementazione

Vediamo come implementare la dipendenza tra microservizi in modo semplice ma funzionale. Osserviamo che vogliamo aggiungere una correlazione tra i tempi di esecuzione di due microservizi; è importante non aggiungere una dipendenza funzionale, ma temporale tra una coppia.

Andremo ad aggiungere due livelli: nella prima avremo una correlazione solamente tra due microservizi; nella seconda aggiungeremo alla precedente anche un altro tipo di correlazione. Ci aspettiamo che gli esperimenti sulla correlazione maggiore siano meno precisi.

### 6.1.1 Bassa correlazione

In questo esperimento si è aggiunta una sola dipendenza, in particolare tra *users-service* e *reservation-service*.

Prima di chiamare *reservation-service/reservation*, il microservizio *users-service* campiona un tempo di esecuzione  $t$ , in cui fa una busy-wait, da una distribuzione esponenziale con tasso  $\frac{1}{7}$ ; il servizio *reservation* farà una busy wait per un tempo di dipendente da quanto campionato precedentemente da *users-service*, secondo la funzione

$$f(t) = \begin{cases} 5ms & \text{se } t < 7 \\ 15ms & \text{altrimenti} \end{cases}.$$

L'implementazione avviene passando un parametro booleano nella chiamata di *users-service* a *reservation-service*.

La matrice di correlazione diventa, rispetto alla precedente, quella mostrata in Figura 6.2.

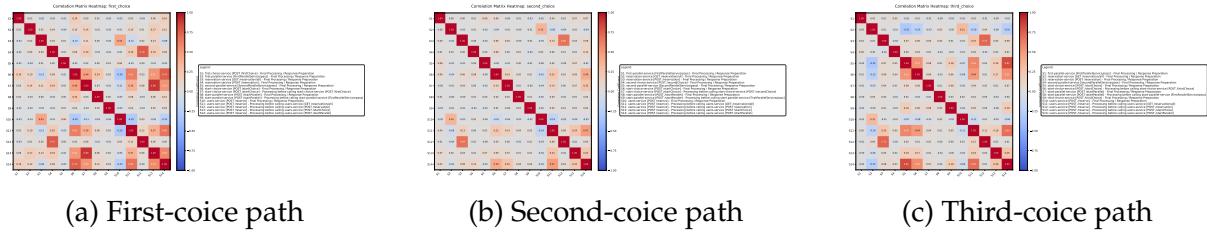


Figura 6.2: Matrici di correlazione nei servizi con bassa correlazione.

### 6.1.2 Alta correlazione

L'implementazione con alta correlazione è definita sulla base di quella precedente a bassa correlazione, a cui è stata aggiunta un'ulteriore dipendenza tra i tempi di esecuzione di microservizi.

In particolare dopo che i microservizi paralleli *first-parallel-service* e *second-parallel-service* ritornano, il chiamante *start-parallel-service* esegue per un tempo che è definito dalla somma dei tempi di esecuzione dei due servizi paralleli. Questo viene implementato facendo ritornare ai due servizi un double che rappresenta il tempo per essi hanno eseguito la loro busywait; *start-parallel-service* eseguirà una busy-wait per questo tempo.

La matrice di correlazione in questo caso è mostrata Figura 6.3.

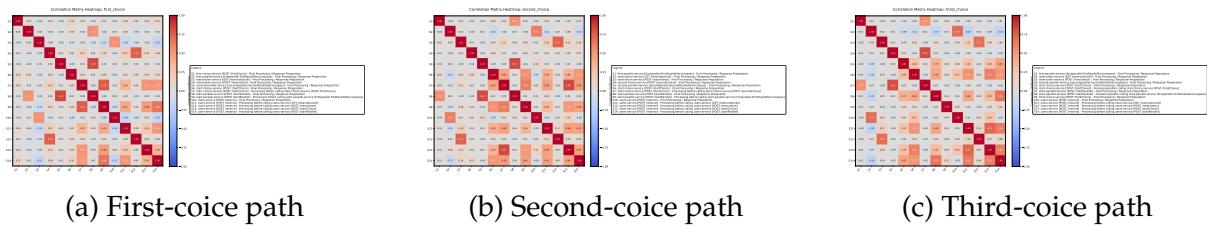


Figura 6.3: Matrici di correlazione nei servizi con alta correlazione.

## 6.2 Risultati

Vediamo adesso i risultati ottenuti con le due simulazioni, cioè i grafici delle distribuzioni reali e ottenuti con Eulero.

- I risultati ottenuti per la bassa correlazione sono mostrati in Figura 6.4a.
- I risultati ottenuti per l'alta correlazione sono mostrati in Figura 6.4b.

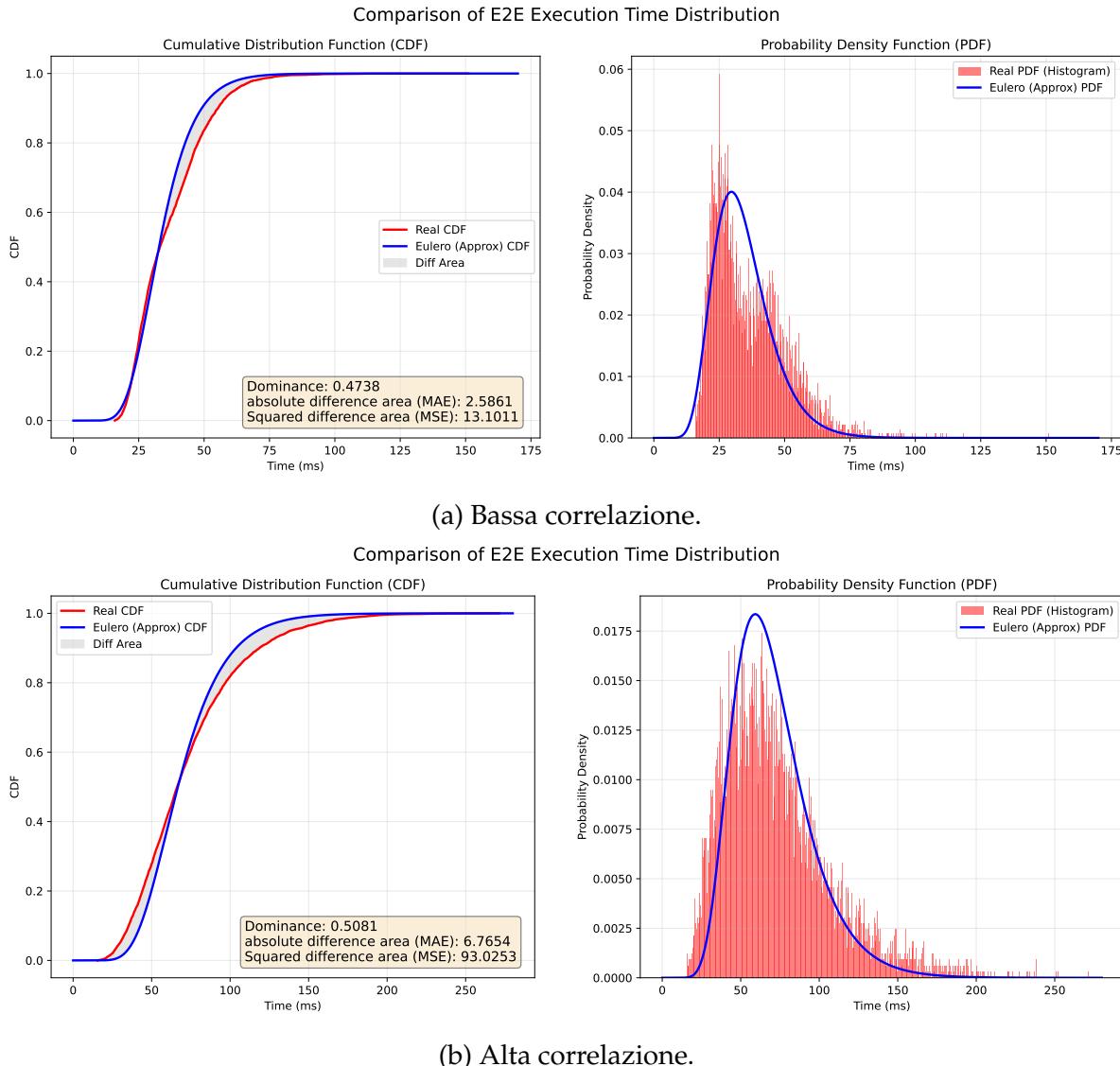


Figura 6.4: Distribuzioni negli esperimenti con correlazione diversa.

Possiamo notare come l'area delle curve aumenta all'aumentare della correlazione, segno che l'analisi compositiva eseguita da Eulero non riesce a catturare tali dipendenze.

## 6.3 Bande di confidenza

Analizziamo adesso esattamente le stesse bande di confidenza viste nella Sezione 5.5.

- Per le distribuzioni con bassa correlazione possiamo osservare la Figura 6.5.
- Per le distribuzioni con alta correlazione possiamo osservere la Figura 6.6.

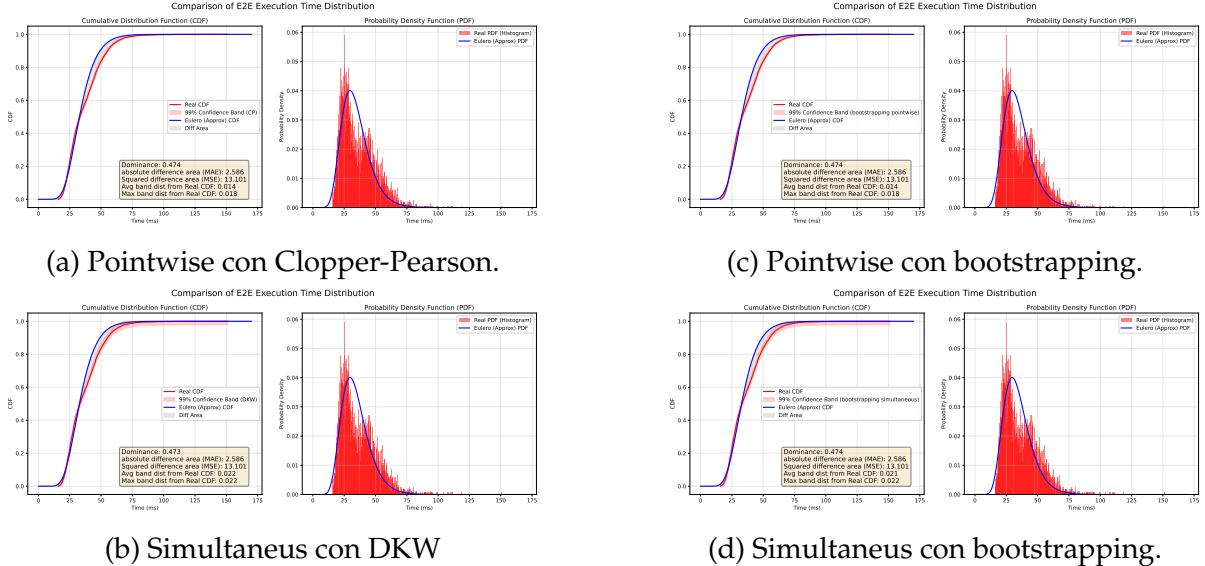


Figura 6.5: Bande di confidenza per le distribuzioni con bassa correlazione.

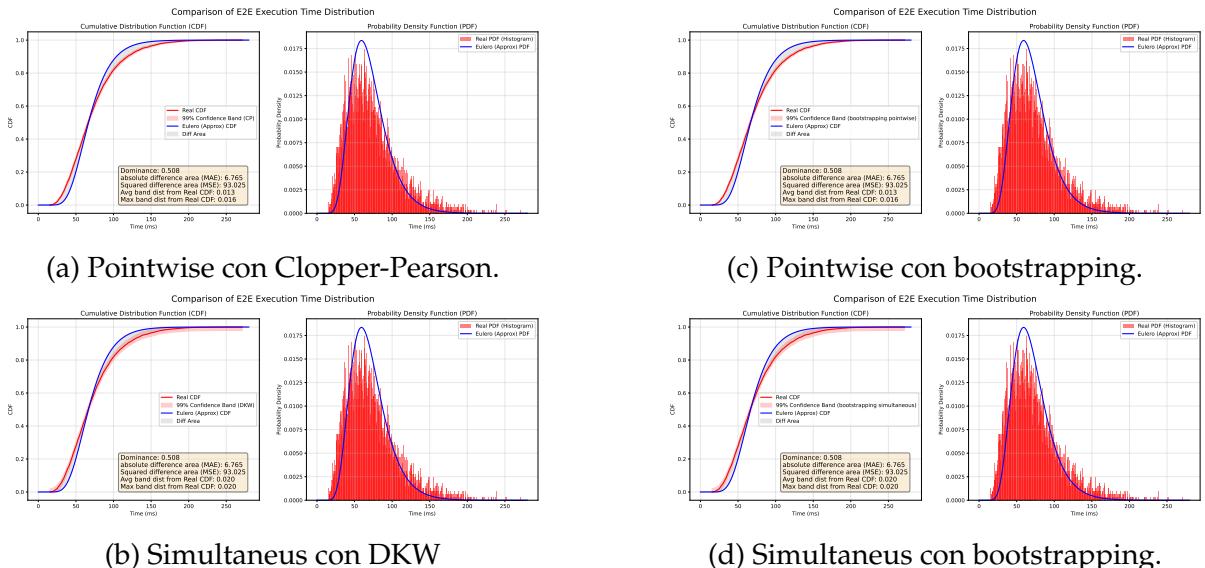


Figura 6.6: Bande di confidenza per le distribuzioni con alta correlazione.

# Bibliografia

- [1] Marco Biagi et al. «A Continuous-Time Model-Based Approach for Activity Recognition in Pervasive Environments». In: *IEEE Transactions on Human-Machine Systems* 49.4 (2019), pp. 293–303. DOI: [10.1109/THMS.2019.2903091](https://doi.org/10.1109/THMS.2019.2903091).
- [2] Bitnami Helm Charts for MySQL. URL: <https://artifacthub.io/packages/helm/bitnami/mysql> (visitato il giorno 03/09/2025).
- [3] Patrick Breheny. *The Bootstrap (Lecture Notes from BIOS 7640: Advanced Biostatistics)*. <https://myweb.uiowa.edu/pbreheny/uk/764/notes/12-6.pdf>. Accessed: 2025-12-19. 2016.
- [4] Laura Carnevali et al. «Compositional Coordinated Resource Provisioning in Workflows With Stochastic Durations». In: *IEEE Transactions on Parallel and Distributed Systems* 36.9 (2025), pp. 1937–1954. DOI: [10.1109/TPDS.2025.3585821](https://doi.org/10.1109/TPDS.2025.3585821).
- [5] Laura Carnevali et al. «Compositional safe approximation of response time probability density function of complex workflows». In: *ACM Transactions on Modeling and Computer Simulation* 33.4 (2023), pp. 1–26.
- [6] Prometheus Community. *Kube Prometheus Stack*. 2025. URL: <https://artifacthub.io/packages/helm/prometheus-community/kube-prometheus-stack#configuration> (visitato il giorno 04/09/2025).
- [7] Stephanie Glen. *Clopper-Pearson Exact Method*. Statistics How To. 2024. URL: <https://www.statisticshowto.com/clopper-pearson-exact-method/> (visitato il giorno 18/12/2024).
- [8] JaegerTracing. *Jaeger Helm chart*. URL: <https://artifacthub.io/packages/helm/jaegertracing/jaeger> (visitato il giorno 03/09/2025).
- [9] Kubernetes Authors. *Liveness, Readiness and Startup Probes*. 2025. URL: <https://kubernetes.io/docs/concepts/configuration/liveness-readiness-startup-probes/> (visitato il giorno 03/09/2025).
- [10] Marco Paolieri et al. «The ORIS Tool: Quantitative Evaluation of Non-Markovian Systems». In: *IEEE Trans. Software Eng.* 47.6 (2021), pp. 1211–1225. DOI: [10.1109/TSE.2019.2917202](https://doi.org/10.1109/TSE.2019.2917202).
- [11] Quarkus - All configuration options. URL: <https://quarkus.io/guides/all-config> (visitato il giorno 04/09/2025).
- [12] Quarkus Team. *Configure data sources in Quarkus*. URL: <https://quarkus.io/guides/datasource> (visitato il giorno 03/09/2025).

- [13] Quarkus Team. *SmallRye Health - Quarkus*. URL: <https://quarkus.io/guides/smallrye-health> (visitato il giorno 03/09/2025).
- [14] Martin Štefanko e Jan Martiška. *Quarkus in Action*. <https://www.manning.com/books/quarkus-in-action>. Data ultima visualizzazione: 2026-07-30. Manning Publications, 2025.
- [15] Burr Sutter et al. *Kubernetes Native Microservices*. O'Reilly Media, 2020.
- [16] Wikipedia contributors. *Binomial proportion confidence interval*. En línia; accedit el 18-desembre-2024. Wikipedia, The Free Encyclopedia. 2024. URL: [https://en.wikipedia.org/wiki/Binomial\\_proportion\\_confidence\\_interval](https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval) (visitato il giorno 18/12/2024).
- [17] Wikipedia contributors. *Dvoretzky–Kiefer–Wolfowitz inequality* — Wikipedia, The Free Encyclopedia. [Online; accessed 19-December-2024]. 2024. URL: [https://en.wikipedia.org/wiki/Dvoretzky%20%93Kiefer%20%93Wolfowitz\\_inequality](https://en.wikipedia.org/wiki/Dvoretzky%20%93Kiefer%20%93Wolfowitz_inequality).