# Performance Requirements and Testing

## Service Call Chains

To create an accurate test plan, the cascading service calls for each user function must be understood.

### List Retrieval Call Chain

1. **Client Pre-Check:** The client first calls the **Features API** to check for flags like PlanDocsMsgAvail to determine if a special message should be displayed immediately, bypassing the main document list call.
2. **Client to BFF:** The client sends a POST request to the new **OpenPlatform BFF** at /sa/plandocs/v1/list, including the JWT id\_token for user context .
3. **BFF to Backend:** The BFF calls the existing **Plan Docs Service** via POST /v1/plan-document-list/retrieve to fetch the document list.
4. **Backend to Cache/Core:** The backend Benefits Service checks a **Redis cache** for the list. On a cache miss, it calls the

**Core API** (/eie/header/v4/healthpolicies/communicationcontents/search/retrieve) and stores the result in Redis for one hour before returning it.

### Document (PDF) Retrieval Call Chain

1. **Client to BFF:** The client sends a POST request to the **OpenPlatform BFF** at /sa/plandocs/v1/retrieve with a specific documentId .
2. **BFF to Backend:** The BFF calls the existing **Benefits Service** at its /plan-document/retrieve endpoint.
3. **Backend to Cache/Core:** The Benefits Service first checks an **S3 cache** for the PDF. On a cache miss, it retrieves the file from the

**Core API** (/eie/header/v5/communicationcontents/{id}), stores it in S3, and then streams it back to the client.

## Required Performance Metrics

This section outlines the performance targets the solution must meet. These metrics serve as the acceptance criteria for the test plan.

### Volume and Throughput Requirements

The new BFF endpoints must support the existing backend service rate limits to ensure system stability.

* **Plan Document List (/sa/plandocs/v1/list):** 950 requests per minute.
* **Plan Document PDF (/sa/plandocs/v1/retrieve):** 2333 requests per minute.

### Response Time and Latency Requirements

The solution must provide "fast responses" and not degrade application performance. Specific latency targets (e.g., 95th percentile response time) should be defined and filled in below.

* **List Retrieval (Warm Cache):** \_\_\_\_\_ ms
* **List Retrieval (Cold Cache):** \_\_\_\_\_ ms
* **PDF Retrieval (Warm Cache):** \_\_\_\_\_ ms
* **PDF Retrieval (Cold Cache):** \_\_\_\_\_ ms

### Performance Test Plan and Results Grid

### Narrative and Methodology

The objective of this performance test plan is to validate that the end-to-end plan documents solution is responsive, stable, and scalable against the defined NFRs. The methodology focuses on simulating real-world user traffic against the OpenPlatform BFF endpoints to measure the cumulative latency of all downstream calls.

To ensure comprehensive coverage, tests will be executed under two distinct cache conditions:

* **Cold Cache:** All caches (Redis, S3) are cleared before the test to measure the worst-case performance scenario.
* **Warm Cache:** Caches are fully populated before the test to measure the best-case performance and validate the caching strategy.

Testing will use a variety of user profiles to cover scenarios such as users with zero, one, or multiple documents.

### Test Execution Grid

| **Test ID** | **User Action** | **API Endpoint** | **Test Type** | **Cache State** | **Scenario Description** | **Metric / Result** |
| --- | --- | --- | --- | --- | --- | --- |
| LT-LST-01 | List Retrieval | POST /sa/plandocs/v1/list | Latency | Cold | Single user requests a list of 10 documents. Cache is empty. | 95th Percentile Latency (ms): |
| LT-LST-02 | List Retrieval | POST /sa/plandocs/v1/list | Latency | Warm | Single user requests a list of 10 documents. Cache is populated. | 95th Percentile Latency (ms): |
| LD-LST-01 | List Retrieval | POST /sa/plandocs/v1/list | Load | Warm | Sustained load test with a mix of users (0, 1, and multiple docs). | Throughput (RPM): Error Rate (%): |
| LD-LST-02 | List Retrieval | POST /sa/plandocs/v1/list | Load | Warm | Stress test by increasing load beyond target RPM to find the breaking point. | Max RPM @ <1% Error Rate: |
| LT-PDF-01 | PDF Retrieval | POST /sa/plandocs/v1/retrieve | Latency | Cold | Single user requests a 1MB PDF document. Cache is empty. | 95th Percentile Latency (ms): |
| LT-PDF-02 | PDF Retrieval | POST /sa/plandocs/v1/retrieve | Latency | Warm | Single user requests a 1MB PDF document. Cache is populated. | 95th Percentile Latency (ms): |
| LD-PDF-01 | PDF Retrieval | POST /sa/plandocs/v1/retrieve | Load | Warm | Sustained load test with requests for various cached PDF documents. | Throughput (RPM): Error Rate (%): |
| LD-PDF-02 | PDF Retrieval | POST /sa/plandocs/v1/retrieve | Load | Warm | Stress test by increasing load beyond target RPM to find the breaking point. | Max RPM @ <1% Error |