PLATAFORMA DE BANCA

Nombre: Oscar Fernando Ibarra Torres

Descripción: Se describe la propuesta de una plataforma de banca por internet con énfasis en la seguridad.

Modelo C4

Nivel 1: Contexto

En este nivel se identifican los actores y las interacciones de alto nivel. Los usuarios finales (web y móvil) se conectan al sistema a través de un API Gateway protegido por WAF, que canaliza las solicitudes a los backends específicos (BFF Web y BFF Mobile). Los microservicios internos gestionan clientes, movimientos, transferencias, onboarding, notificaciones y fraude. Se integran con sistemas externos como el Core Bancario, ACH, proveedores biométricos y de notificaciones, y el sistema complementario. Todos los servicios registran acciones en la base de auditoría inmutable. Este contexto permite decidir dónde colocar controles perimetrales (WAF, API Gateway) y cómo cumplir estándares como ISO 27001 y OWASP ASVS.

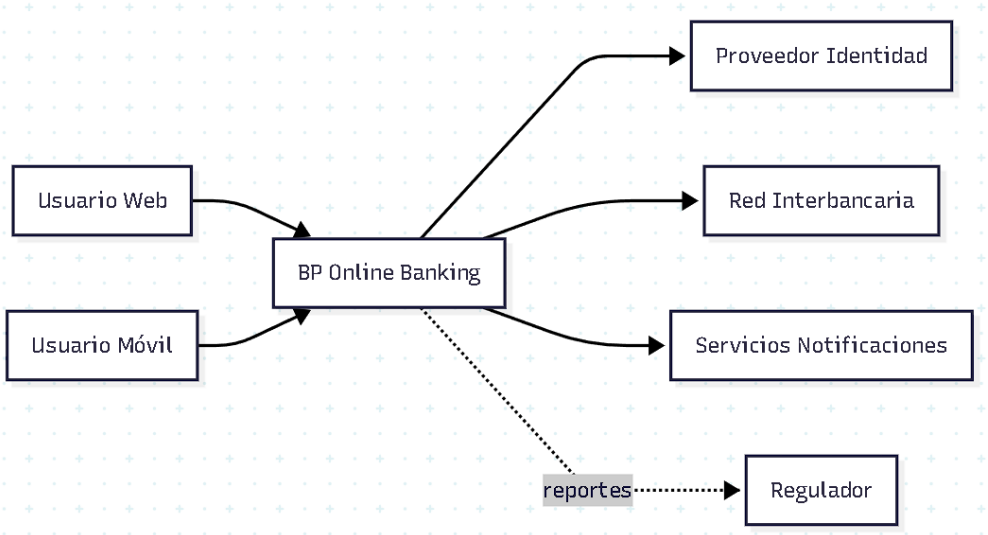


Figura 1

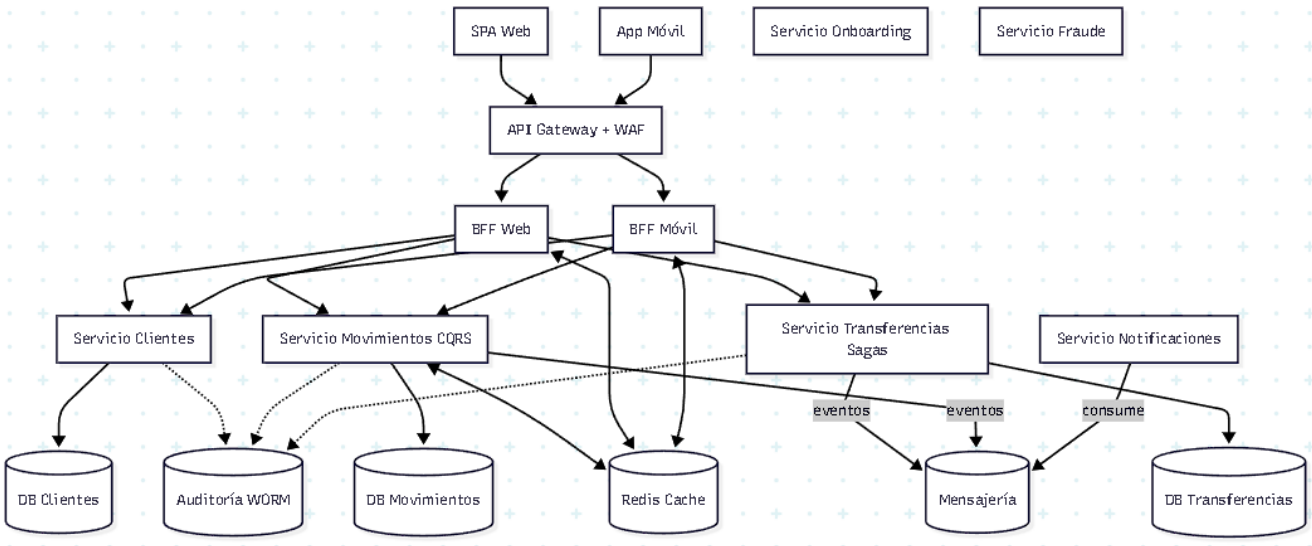
Este diagrama de contexto ubica a BP Online Banking en su ecosistema y define con claridad los límites de confianza y las dependencias externas que condicionan la arquitectura. Los nodos UWeb y Umovil representan a los clientes que se conectan desde un navegador (SPA) y desde la app móvil, respectivamente. Ambos canales comparten casos de uso (consultar movimientos, transferir, pagar), pero plantean riesgos distintos: en web se priorizan cookies httpOnly, SameSite y CSP para proteger tokens y contenido; en móvil se emplean Keychain/Keystore y atestación de dispositivo para reducir el fraude por apps modificadas. El nodo BP es el sistema bajo nuestra responsabilidad, que expone APIs seguras y aplica autorizaciones por scopes y políticas.

La flecha hacia IdP señala que la autenticación se delega a un proveedor de identidad OIDC/OAuth2, encargado de MFA, federación y emisión de tokens; este vínculo exige SLA estrictos, monitoreo sintético y alertas por latencia/errores, pues su caída se traduce en indisponibilidad. La conexión con ACH evidencia la dependencia de la red interbancaria para transferencias externas; aquí impactan ventanas de procesamiento, confirmaciones diferidas y requisitos de conciliación. Noti agrupa servicios de email, SMS y push; se opta por una integración desacoplada mediante colas y reintentos para que la experiencia del usuario no dependa del éxito inmediato del envío. Por último, Reg representa al regulador y a las obligaciones de reporte, trazabilidad y retención: el sistema debe producir evidencias confiables (auditoría, sellos de tiempo, hashing encadenado) con controles de acceso y segregación de funciones.

Las direcciones de las flechas resumen el flujo de interacción: clientes → BP para operaciones; BP → IdP/ACH/Noti para capacidades especializadas; BP → Reg para informes. Este mapa guía decisiones posteriores: dónde ubicar WAF y API Gateway, qué telemetría recolectar, cómo dimensionar rate limiting y backoffs ante terceros, qué contratos (SLA, RPO/RTO) pactar, y qué estrategias de degradación funcional aplicar. En síntesis, el contexto asegura una visión compartida de actores, relaciones y riesgos que alimenta el diseño lógico y operativo del sistema.

Nivel 2: Contenedores

La solución se compone de contenedores frontend (SPA web y app móvil) que se comunican con un API Gateway. En el backend se encuentran los BFF diferenciados por canal, servicios de dominio desacoplados (clientes, movimientos con CQRS, transferencias con Sagas, onboarding biométrico, notificaciones y fraude), y componentes de infraestructura como colas de mensajería, cache distribuida y bases de datos independientes por servicio. La auditoría utiliza almacenamiento WORM y la gestión de secretos se realiza mediante KMS o Key Vault. Cada componente está alineado con prácticas de seguridad como segregación de datos por servicio y cifrado en reposo/transito, siguiendo normativas de protección de datos (LOPDP, GDPR) y estándares de cifrado (AES-256, TLS 1.3). El diseño asegura cumplimiento con requisitos regulatorios de continuidad operativa y resiliencia.



El diagrama de contenedores traduce la visión del contexto a una estructura de ejecución. SPA y APP son clientes ligeros que consumen APIs; su tráfico pasa por CDN + WAF, primera capa de defensa y rendimiento (cache de estáticos, mitigación de bots, OWASP CRS). El API Gateway (GW) concentra terminación TLS, validación básica de JWT/DPoP, rate limiting per usuario/dispositivo, control de cuotas y enrutamiento; unifica telemetría en el perímetro y reduce superficie de ataque. Los BFF —BFFW para web y BFFM para móvil— adaptan payloads, agregan datos entre servicios y evitan el chatty entre front y backend; permiten versionar UI/experiencias sin afectar dominios.

Los servicios de dominio encapsulan responsabilidades: CLI integra datos del core y del sistema complementario; MOV implementa CQRS con read store especializado para consultas masivas; TRF orquesta transferencias internas y ACH usando Sagas, idempotencia y Outbox; ONB gestiona KYC y biometría; NOT estandariza el envío multi-canal con plantillas y reintentos; FRD ejecuta reglas y puntajes de riesgo para frenar operaciones sospechosas. La infraestructura transversal incluye MQ (colas/tópicos para desacoplar y absorber picos), CACHE

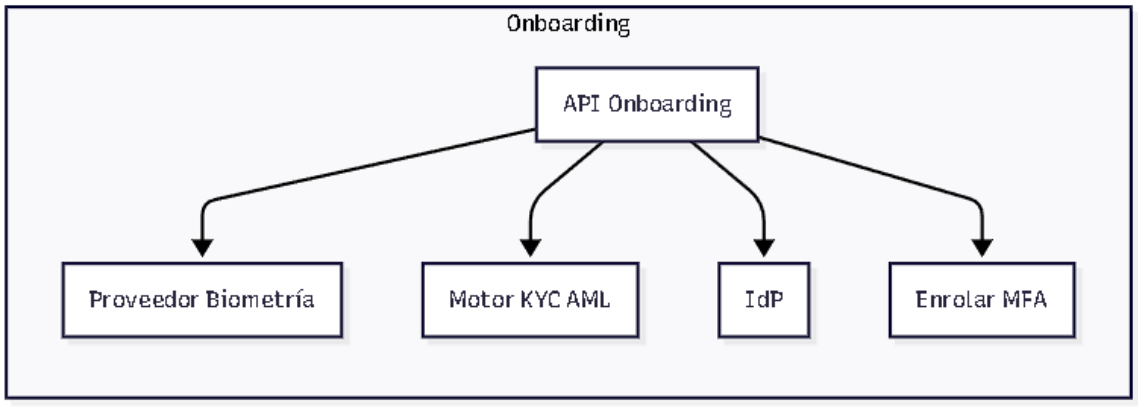
(Redis) para lecturas calientes y DBA/DBM/DBT como bases por servicio (autonomía de esquema, fallo contenido, escalado específico). AUD es el almacén WORM para registros inmutables con retención regulatoria.

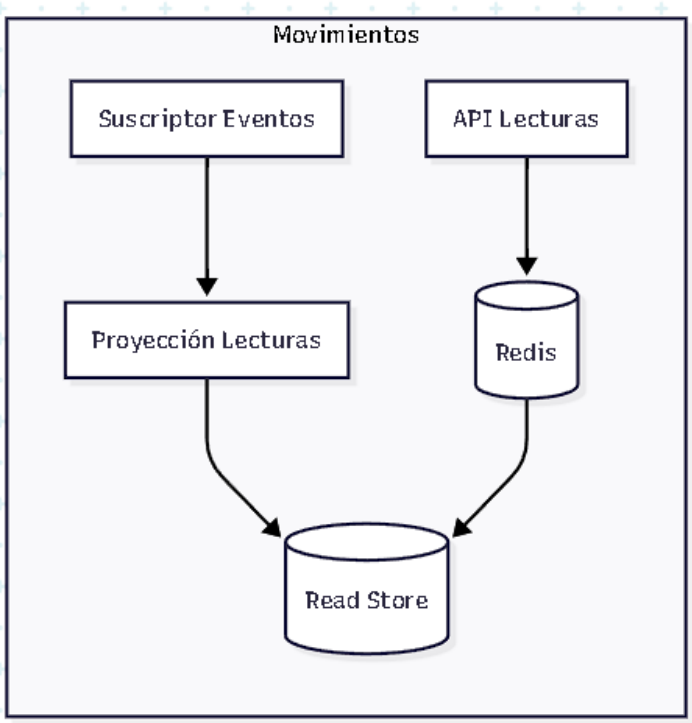
Las flechas reflejan flujos típicos: clientes → GW → BFF → servicios; lecturas pasan por CACHE para optimizar TTFB y costos; eventos de dominio van a MQ para notificaciones y proyecciones; los servicios envían trazas y logs a la plataforma de observabilidad (no dibujada aquí). Este layout favorece escalabilidad horizontal por bloque, resiliencia (circuit breakers, retries), seguridad (mTLS este-oeste, OPA para políticas, secretos en Vault/KMS), y operabilidad (despliegues canary/blue-green por contenedor) con un blast radius acotado ante fallos.

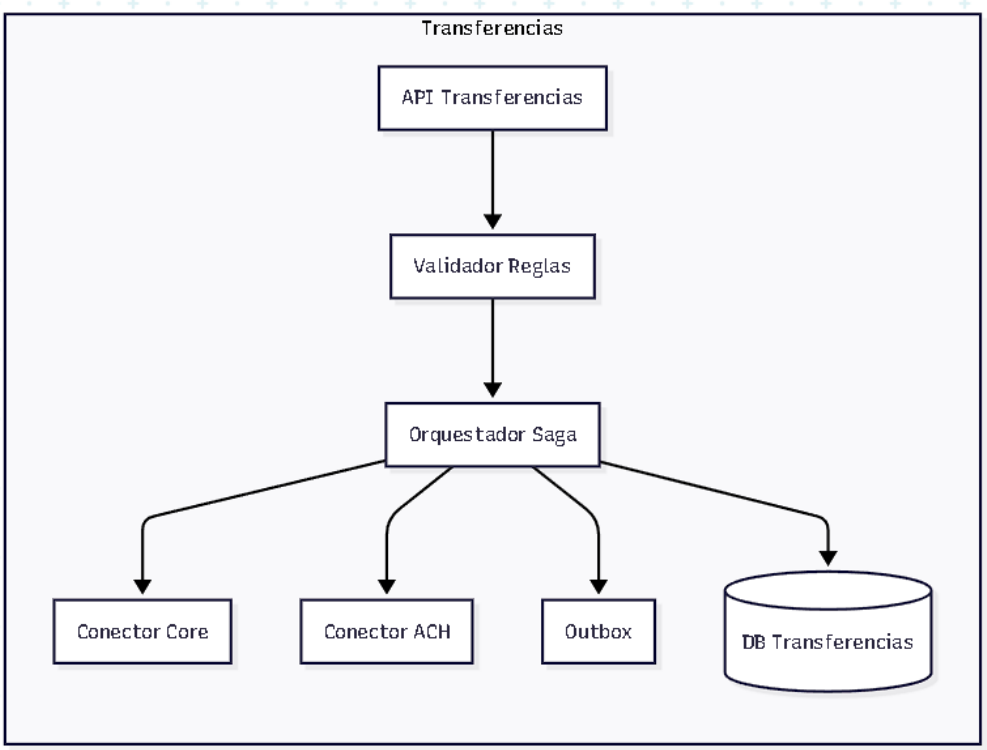
Nivel 3: Componentes

En el servicio de transferencias, el orquestador de Sagas coordina validaciones, reservas de fondos, comunicación con ACH y confirmaciones, asegurando idempotencia con claves únicas. El servicio de movimientos implementa CQRS para manejar cargas de lectura altas, alimentando un modelo de lectura optimizado desde eventos del core. El servicio de onboarding conecta con un proveedor biométrico para realizar pruebas de liveness y coincidencia facial, valida contra listas PEP/sanciones y registra al usuario en el IdP, enrolando factores de autenticación.

A continuación se desglosan los contenedores clave: Transferencias, Movimientos y Onboarding. Transferencias incorpora reglas de negocio y orquestación de Sagas con idempotencia, cumpliendo límites regulatorios diarios y verificaciones AML/KYC. Movimientos implementa CQRS para escalabilidad y trazabilidad, asegurando auditoría de transacciones. Onboarding integra biometría y validaciones AML, cumpliendo con estándares de identificación digital y prevención de fraude (FATF, NIST SP 800-63).







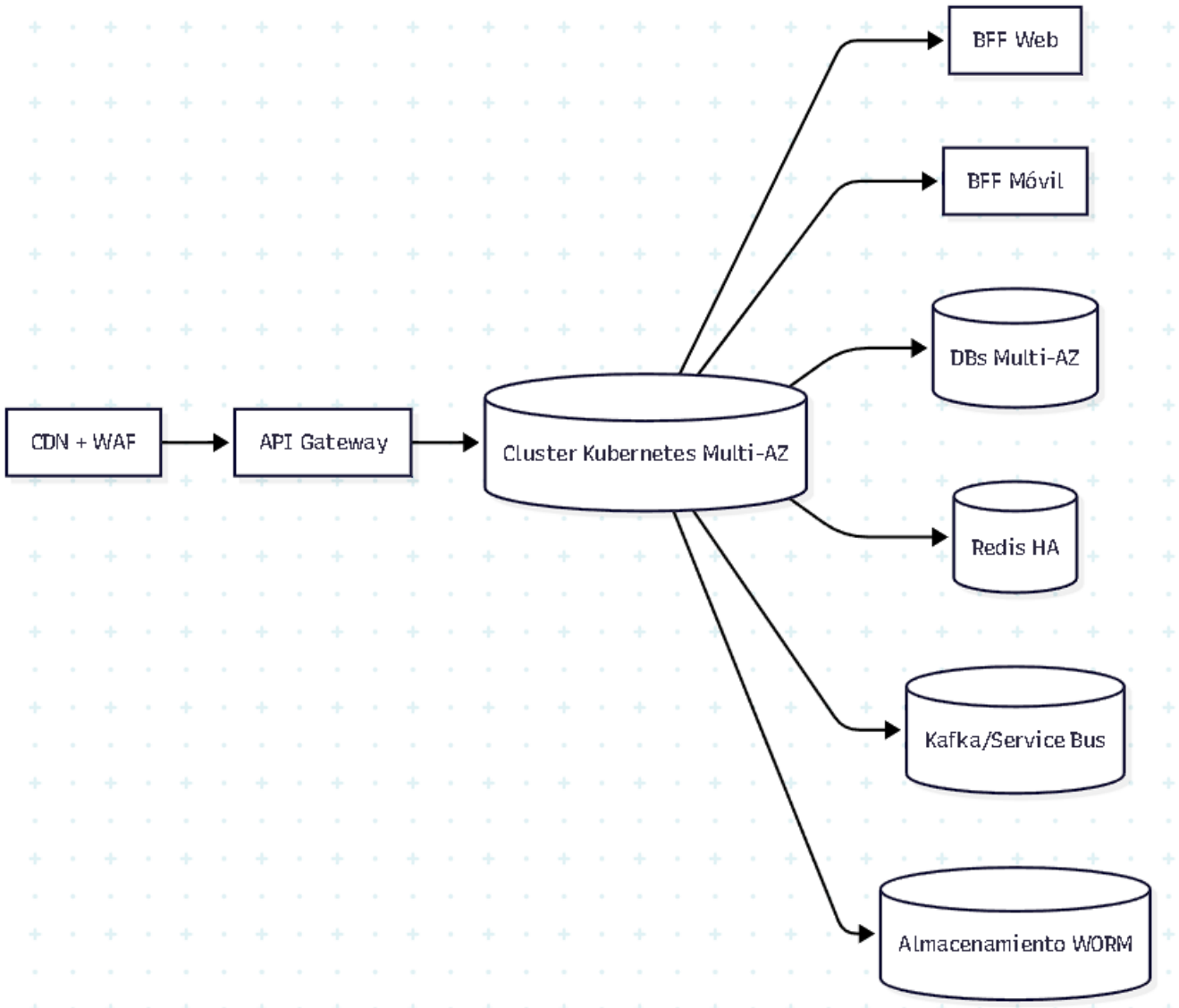
En Transferencias, API_T expone endpoints autenticados con scopes específicos (lectura/creación/seguimiento). Cada petición porta una Idempotency-Key única para asegurar exactly-once effect en presencia de reintentos por caídas de red o latencia. VAL_T centraliza reglas: límites por usuario/día, listas internas, validación de horario ACH, verificación de titulares/alias. ORQ implementa la Saga: 1) reserva fondos en el Core (vía CCORE), 2) envía la orden a ACH (vía CACH), 3) confirma o compensa según el resultado, 4) publica eventos con Outbox para notificaciones, proyecciones y auditoría. DBT persiste estado, histórico y claves de idempotencia; índices por trackingId, status y createdAt facilitan consultas y conciliaciones.

En Movimientos, SUB consume CDC/eventos del core bancario, PROJ denormaliza y construye el READ store orientado a consultas (índices por fecha, tipo, monto; paginación por cursor para listas largas). CACHE acelera lecturas frecuentes (últimos 90 días) con invalidación por evento nuevo. API_M sirve estas vistas garantizando latencias predecibles, reduciendo presión sobre sistemas transaccionales. El patrón CQRS separa claramente escritura y lectura, habilitando escalado independiente y resiliencia frente a picos.

En Onboarding, API_O coordina la verificación biométrica con BIO (liveness + face match), ejecuta reglas KYC/AML (listas PEP, sanciones), y, si pasa, crea identidad en el IdP y realiza ENR (enrolamiento de MFA/passkeys y device binding). La separación en conectores evita acoplamiento al proveedor, simplifica pruebas y cumplimiento (auditorías pueden revisar cada paso con correlationId). En conjunto, este nivel expone fronteras transaccionales, puntos de fallo (Core/ACH/BIO), y las mitigaciones (timeouts, retries, circuit breakers, Outbox) que sostienen consistencia y experiencia.

C4 — Nivel 4: Despliegue

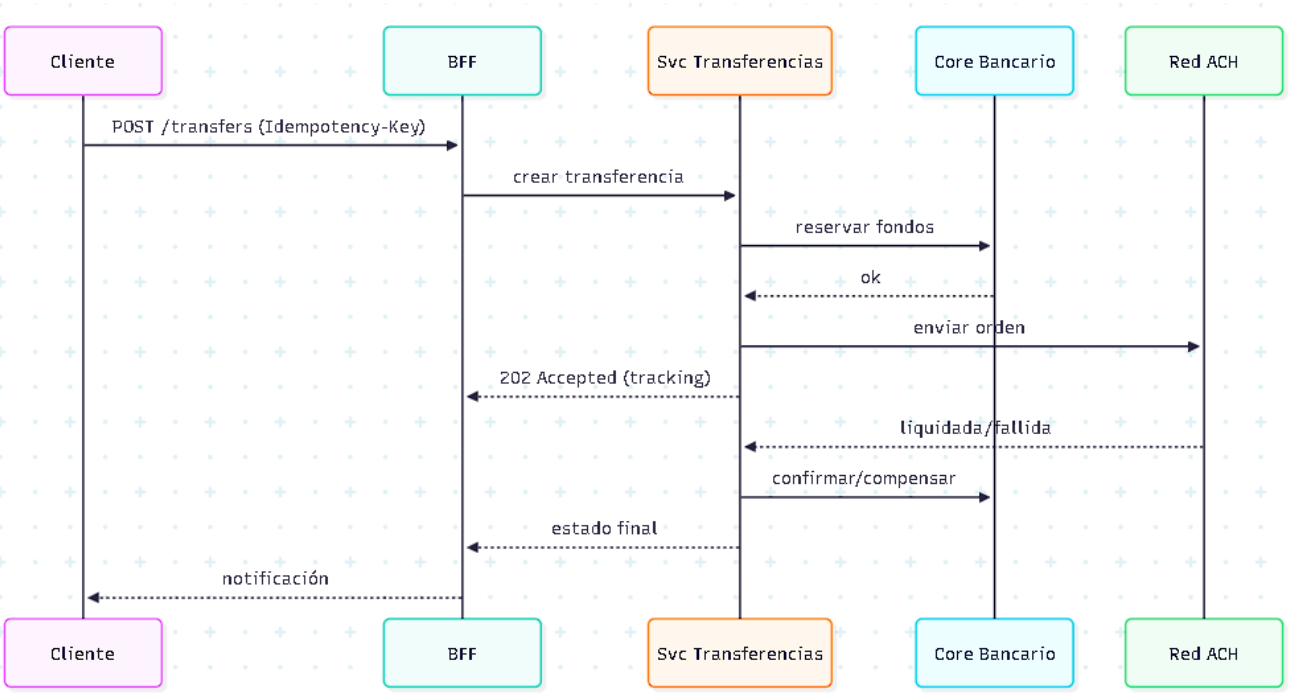
El despliegue se realiza sobre un clúster Kubernetes multi-AZ, con bases de datos y cache en alta disponibilidad, mensajería gestionada, y almacenamiento inmutable para auditoría. Un CDN/WAF distribuye y protege el contenido estático y las llamadas API. El sistema contempla planes de recuperación ante desastres con RPO ≤ 5 minutos y RTO ≤ 30 minutos.



La API seguirá el estándar OpenAPI, con seguridad basada en OAuth2 y PKCE, y uso de claves de idempotencia para evitar duplicados. Los modelos de datos incluyen tablas para transferencias, proyecciones de movimientos y un registro de auditoría encadenado con hashes.

En seguridad, se aplicarán controles OWASP ASVS/MASVS, cifrado TLS1.3 y AES-256, mTLS interno, DPoP/MTLS para vincular tokens, atestación de dispositivos, y gestión de secretos centralizada. El análisis STRIDE identifica amenazas como spoofing, manipulación de datos y denegación de servicio, proponiendo mitigaciones específicas.

En operación, se implementarán métricas, trazas distribuidas y alertas para cumplir SLOs exigentes. La estrategia de pruebas abarcará desde unitarias y de integración hasta rendimiento, seguridad y aceptación. El roadmap contempla un MVP con login, movimientos, transferencias internas, notificaciones básicas y auditoría, evolucionando hacia transferencias interbancarias, onboarding biométrico, y recuperación multi-región.

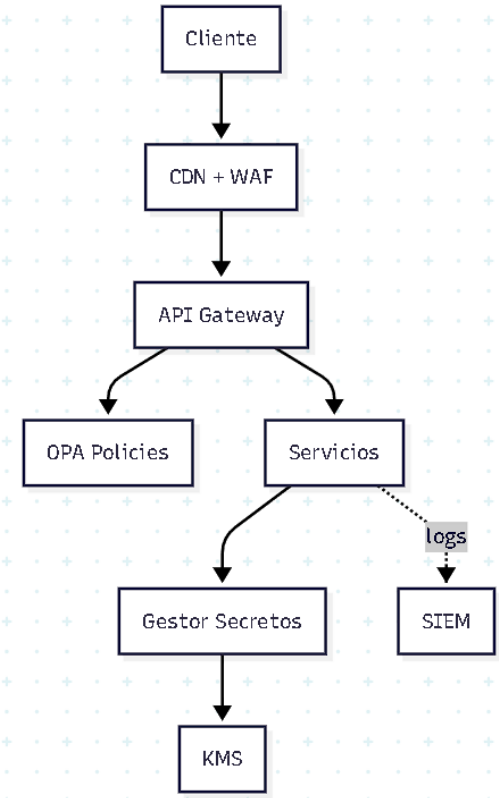


Esta secuencia describe el ciclo de vida de una transferencia interbancaria con controles de idempotencia, resiliencia y trazabilidad en cada paso. El Cliente inicia la operación mediante el BFF, que valida la autenticación (token OIDC), aplica políticas (montos, límites) y agrega la Idempotency-Key. TRF intenta reservar fondos en CORE; si fallan comunicaciones, se aplican reintentos con backoff, timeouts razonables y circuit breaker para evitar cascadas. Luego envía la orden a ACH y devuelve 202 Accepted al cliente con trackingId, desacoplando la UX del tiempo de liquidación. Cuando ACH responde, TRF confirma en el Core o compensa liberando la reserva. Todos los pasos emiten eventos (para auditoría, notificaciones y proyecciones) y comparten un correlationId que permite reconstruir el rastro end-to-end. Se contemplan fallos parciales: si no llega la respuesta de ACH, se consulta por estado; si el cliente reintenta, la Idempotency-Key evita duplicados; si expira la ventana, la operación se marca pendiente y se notifica al usuario. Las pruebas E2E deben medir p95/p99, tasas de error por paso y validar que nunca se produzcan dobles débitos.

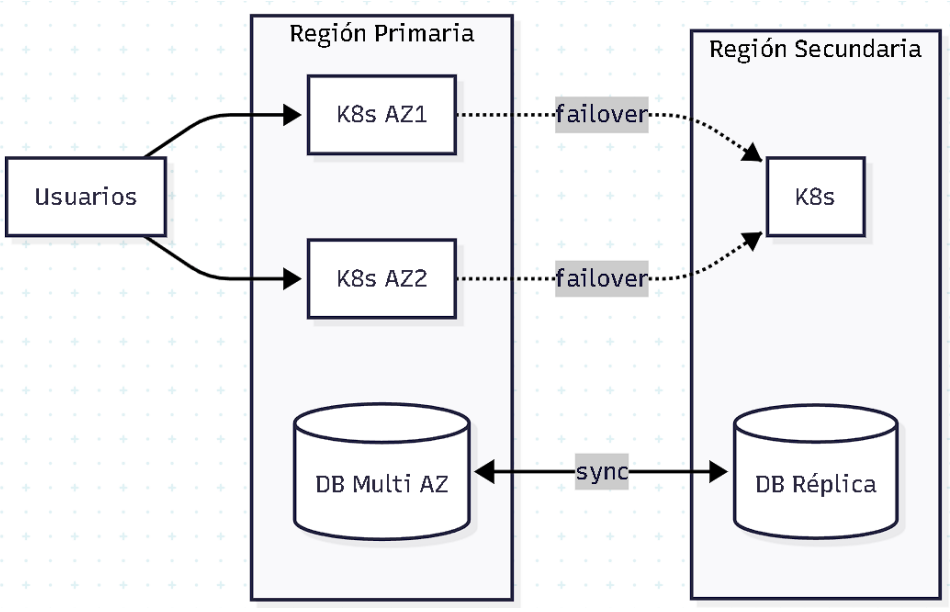
El flujo de auditoría garantiza integridad, trazabilidad y no repudio. Cada servicio (A) genera eventos con campos mínimos (actor, acción, entidad, correlationId, deviceId, IP, ubicación aproximada, marca de tiempo) y los publica en la cola Q con política de entrega al menos una vez y particionado por entidad para conservar orden. Un procesador persiste en L, un log append-only con hash encadenado y sellos de tiempo confiables; periódico anclaje criptográfico (hash maestro) impide alteraciones invisibles. El log se replica a W, almacén WORM con retención legal (7–10 años según normativa), bloqueo contra borrado y cifrado at-rest.

R ofrece vistas segmentadas por rol (cumplimiento, auditores, SOC) y exportaciones seguras; las consultas de auditoría no afectan a las cargas transaccionales. Se aplican medidas de privacidad (minimización de PII, seudonimización cuando sea viable, redacción de secretos) y controles de acceso con SoD y RBAC/ABAC. Alarmas automáticas detectan patrones anómalos (picos de 409 por idempotencia, fallos de MFA, geolocalizaciones atípicas). Incluso ante fallas parciales, los eventos quedan persistidos o encolados,

asegurando que la historia operacional pueda reconstruirse con precisión temporal.

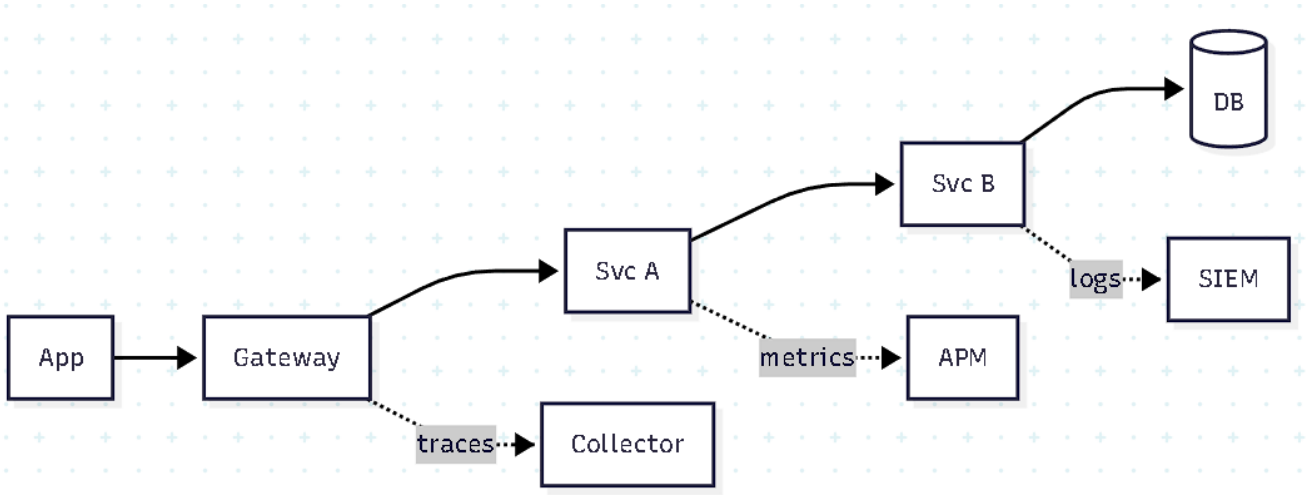


La defensa en profundidad combina controles perimetrales, políticas centralizadas y seguridad de datos. CDN + WAF mitigan ataques de capa 7 y aplican reglas OWASP CRS; GW valida tokens, impone rate limits por cuenta/IP/dispositivo y unifica telemetría. Las políticas OPA (POL) centralizan autorización contextual (scopes, atributos, riesgo), dejando trazabilidad de decisiones. Entre SVC, el tráfico es mTLS, con certificados rotados; las credenciales viven en VAULT y se cifran/rotan con KMS (claves con doble control, auditoría de uso). Los logs enriquecidos fluyen al SIEM para detección y respuesta (detección de anomalías, correlaciones multi-fuente). Complementan esta capa el hardening de contenedores, SAST/DAST/SCA, firmas de artefactos y validación en CI/CD, passkeys/MFA obligatorios para operaciones sensibles y token binding (DPoP/MTLS) para reducir robo/uso indebido de tokens. Todo ello disminuye MTTD/MTTR y sostiene cumplimiento (OWASP ASVS/MASVS, LOPDP).



El plan de alta disponibilidad y recuperación ante desastres garantiza continuidad con RPO ≤ 5 min y RTO ≤ 30 min. En la región primaria, el clúster de Kubernetes se distribuye en dos AZs: balanceadores envían tráfico a pods saludables y el orquestador recrea réplicas ante fallos; la DB Multi-AZ replica de forma síncrona y ofrece PITR y backups cifrados. La región secundaria mantiene una réplica de la base (asincrónica o casi síncrona) y recursos listos para failover. La conmutación puede ser manual controlada o automatizada vía DNS con TTL bajos y health checks. Se emplean tokens stateless para sesiones y warm-up de caches tras conmutar.

Los runbooks definen procedimientos y responsables; se realizan simulacros trimestrales (pérdida de AZ, caída de IdP, latencias hacia ACH) con métricas de tiempo de conmutación y degradación funcional controlada (p. ej., priorizar operaciones internas mientras se restablece ACH). Los pipelines de despliegue soportan blue/green o canary con rollback automático, y la integridad de artefactos se verifica con firmas. Este enfoque reduce el blast radius, evita puntos únicos de fallo y asegura que fondos y registros permanezcan consistentes y recuperables..



La observabilidad une trazas, métricas y logs para diagnosticar y mejorar el sistema. Todas las llamadas propagan traceId/spanId (W3C) desde APP hasta DB; el Gateway inicia o continúa trazas y un collector OTEL exporta a la plataforma de trazas. Detectamos cuellos de botella por servicio/dependencia y medimos latencias p50/p95/p99 por operación. Las métricas siguen RED (Rate, Errors, Duration) para endpoints y USE (Utilization, Saturation, Errors) para recursos: el APM muestra tasas de peticiones, errores 4xx/5xx, duraciones, uso de CPU/memoria y lag de colas, con alertas sintomáticas (por ejemplo, subida de p95 con incremento de errores). Los logs son estructurados (JSON), con campos de negocio y técnicos (usuario, cuenta, endpoint, correlationId, deviceId, IP); se envían al SIEM para correlación y detección de anomalías (picos de 409 por idempotencia, fallos de MFA, patrones de abuso).

Este modelo evita puntos ciegos y facilita post-mortems: cada incidente se reconstruye con evidencias (trazas, métricas, logs) y acciones de mejora (ajustar límites, escalar servicios, optimizar consultas). Las SLOs se calculan desde esta telemetría (p. ej., p95 < 250 ms en lecturas, disponibilidad 99.9%) y las alertas enlazan a paneles y trazas para reducir MTTR. Así, la operación se vuelve proactiva y alineada con objetivos de negocio.