

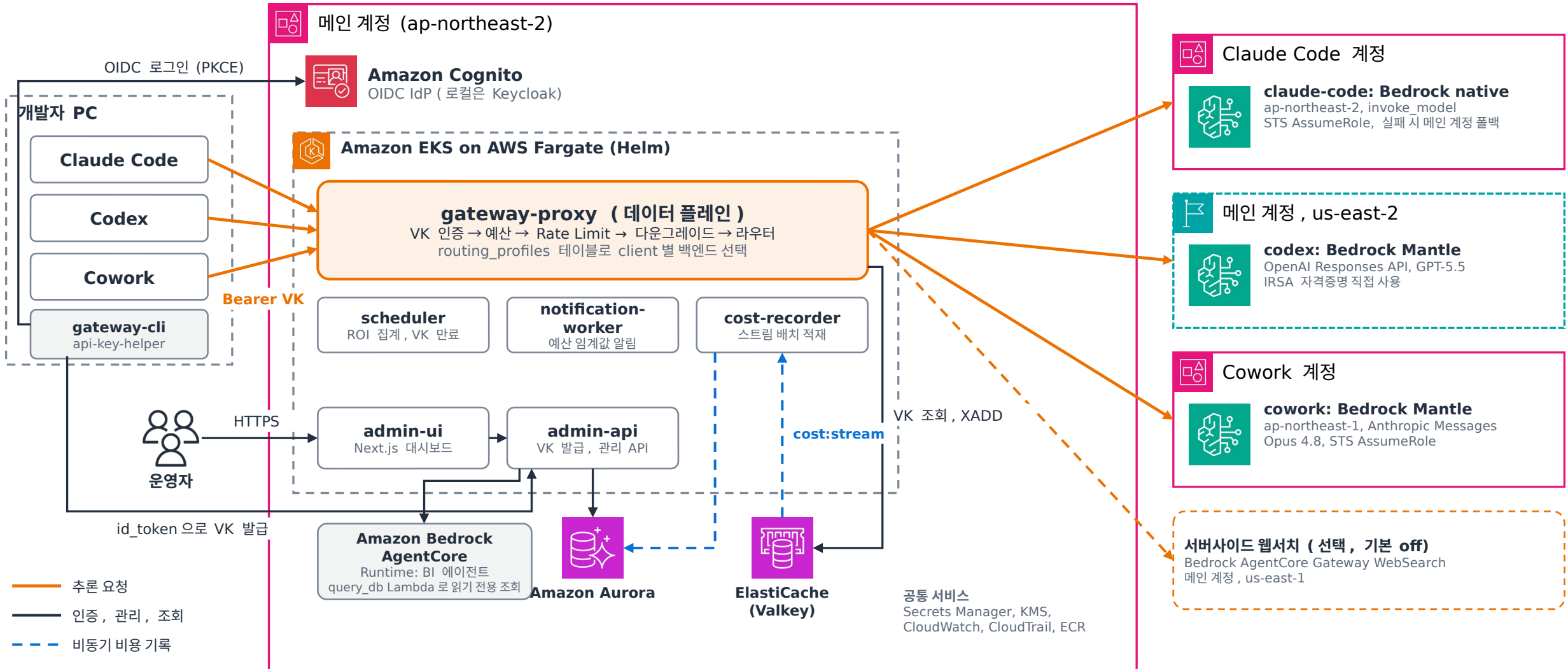


AWSome AI Gateway 아키텍처

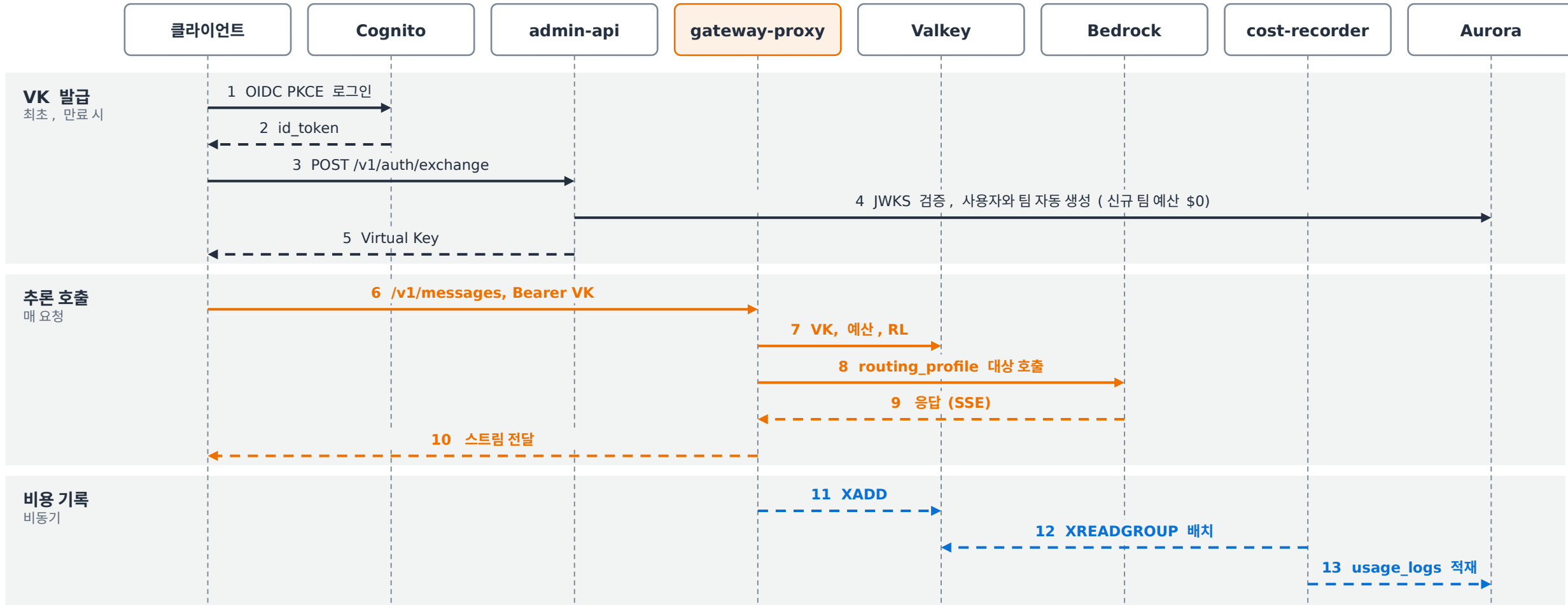
Claude Code, Codex, Cowork 를 위한 단일 인증 ,
거버넌스와 멀티계정 Bedrock 라우팅



하나의 게이트웨이 , 클라이언트별 Bedrock 분기



VK 발급은 한 번 , 매 요청은 캐시만 확인합니다



라우팅은 코드가 아니라 `routing_profiles` 데이터로 정합니다

client	진입 API	백엔드	계정 , 리전	자격증명
claude-code	/v1/messages	Bedrock native (invoke_model)	Claude Code 계정 , ap-northeast-2	STS AssumeRole, 실패 시 메인 계정 폴백
codex	/v1/responses	Bedrock Mantle (GPT-5.5)	메인 계정 , us-east-2	IRSA 직접 (assume 없음)
cowork	/v1/messages	Bedrock Mantle (Opus 4.8)	Cowork 계정 , ap-northeast-1	STS AssumeRole (ExternalId)



RoutingProfileLoader가 **client** 별 **row** 를 **Redis** 에 **300 초** 캐시합니다 .
미등록이거나 비활성인 client 는 기본 Bedrock native 경로로 폴백합니다 .
cross-account 자격증명은 (role, region, ExternalId) 기준으로 캐시하고 , Mantle bearer 토큰은 MantleCredentialBroker 가 발급합니다 .

로컬 docker-compose 는 관리형 서비스만 컨테이너로 바꿉니다

구성 요소	AWS 배포	로컬 (docker-compose)
실행 환경	EKS Fargate, Helm (6 Deployment + migration Job)	docker compose 컨테이너
IdP	Amazon Cognito	Keycloak :8088 (realm dslg)
DB	Amazon Aurora PostgreSQL	PostgreSQL 16 + pgvector :5432
캐시 , 스트림	Amazon ElastiCache (Valkey)	Redis :6379
Bedrock 자격증명	워크로드별 IRSA + STS AssumeRole	EC2 인스턴스 역할 , cross-account 실패 시 in-account 풀백
관측	Amazon CloudWatch, AWS CloudTrail	OTel Collector, Prometheus, Loki, Grafana :3001
BI 에이전트	Amazon Bedrock AgentCore Runtime	구성하지 않음

Q&A

참고 : ARCHITECTURE.md,
README.md