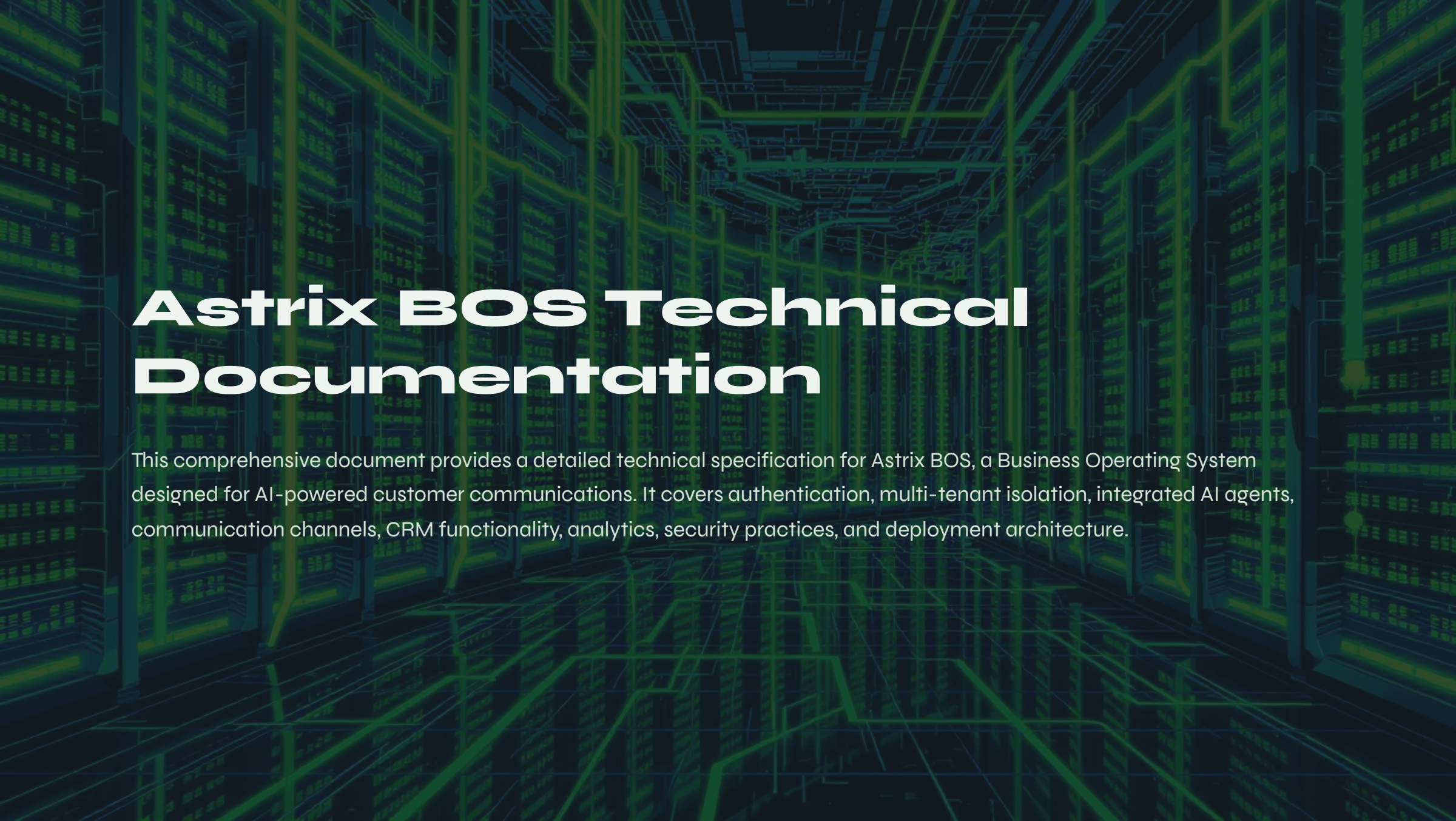


Astrix BOS Technical Documentation

This comprehensive document provides a detailed technical specification for Astrix BOS, a Business Operating System designed for AI-powered customer communications. It covers authentication, multi-tenant isolation, integrated AI agents, communication channels, CRM functionality, analytics, security practices, and deployment architecture.

Authentication, Organizations & Roles

The Astrix BOS authentication system implements a robust multi-tenant architecture with strict isolation between organizations. JWT-based authentication enforces strong security boundaries while providing a flexible role-based permission system.

Authentication Model

The system uses JWT authentication with email/password or SSO options. Tokens carry essential claims including `user_id`, `org_id`, `role`, and `aud:"tenant"`. Tokens have short lifetimes with optional refresh policies to minimize security risks.

Organization Lifecycle

Organizations are self-created by Owners (either as Trial or Paid accounts). Owners can invite additional users with Manager, Developer, or Viewer roles through the Settings → Access interface. Each organization operates in complete isolation from others.

Tenant Isolation

Every tenant route requires both a valid JWT and `X-Org-ID` header. The backend strictly validates that the header matches the token's organization before processing any logic. Mismatches result in a `403 organization_mismatch` error, preventing any cross-tenant data access.

Roles & Permissions

- **Owner:** Full tenant control including billing, keys, and all operations
- **Manager:** Operational capabilities excluding billing/keys
- **Developer:** Build/integration abilities with limited write permissions
- **Viewer:** Read-only access across the platform

HQ Separation

Administrative functions are completely isolated in a separate realm using `aud:"admin"` in JWT tokens. Tenant tokens never work on `/admin/*` routes, maintaining strict separation between customer and administrative operations.

Step-Up Security

Time-based one-time passwords (TOTP) are required for sensitive operations including revealing BYOK credentials, changing user roles, and deleting organizations. These operations enforce step-up authentication with a `401 step_up_required` response when TOTP verification is needed.

Invitations & User Management

New user invitations use short-lived, single-use activation tokens (hashed at rest for security). Accepting an invitation binds the user exclusively to that organization, maintaining tenant boundaries. The system tracks all authorization events in structured JSON logs.

Public Bootstrap

The `/api/auth/me` endpoint is organization-agnostic, allowing the frontend to discover available organizations and roles after login. All other routes enforce the `X-Org-ID` requirement for tenant isolation.

Agents & NOVA (Internal Operations Brain)

Agents are AI workers customized for specific communication channels, while NOVA functions as an internal operations brain that enhances organizational efficiency through automation, strategic insights, and prepared job management.

Agent Model

Each agent belongs to a single organization and operates within a specific channel (voice, SMS, email). Agents follow a lifecycle from Draft → Test → Publish → Pause/Delete. Owners and Managers can test agents in a sandbox environment, publish them for live use, pause, clone, or delete them as needed.

Persona Engine

Each agent has its own persona configuration that includes traits, goals, boundaries, escalation rules, and compliance markers. This ensures appropriate and consistent interactions through each channel.

Knowledge Bases

The system provides two distinct knowledge base types: Brand/NOVA KB (organization-level branding and operational facts) and Agent KB (files, FAQs, and snippets specific to individual agents). This separation prevents NOVA's internal knowledge from affecting external agent behavior.

NOVA Capabilities

NOVA serves multiple roles within the system, acting as:

CRM Primary Editor

NOVA can edit safe CRM fields including stage, tags, owner, and follow-back tasks. Every edit is audited with justification, evidence, and confidence measures.

Strategy & Preparation

NOVA proposes and prepares jobs such as call batches and bulk messaging, providing clear rationales and expected metrics. These prepared jobs require explicit approval before execution.

Draft Agent Studio

NOVA can suggest and create Draft Agents to address observed gaps or improve KPIs. These drafts can be reviewed, modified, and published by authorized users.

Daily Operations Report

Once per day for each organization, NOVA generates a comprehensive operations report that includes KPIs, anomalies, quiet-hours violations to avoid, spend/budget status, and agent optimization proposals. This report provides actionable insights without requiring manual analysis.

Prepared Jobs & BYOK Gate

NOVA creates prepared job objects that live in the organization's data folder. These must be explicitly approved by an Owner or Manager before execution. Additionally, any attempt to send communications (calls, SMS, email) requires the organization to have configured their own BYOK (Bring Your Own Keys) integration. If keys are missing or invalid, the system returns a `400 Configure Integrations first` error.

Role-Based Access Control

The system enforces strict RBAC for agent and NOVA operations:

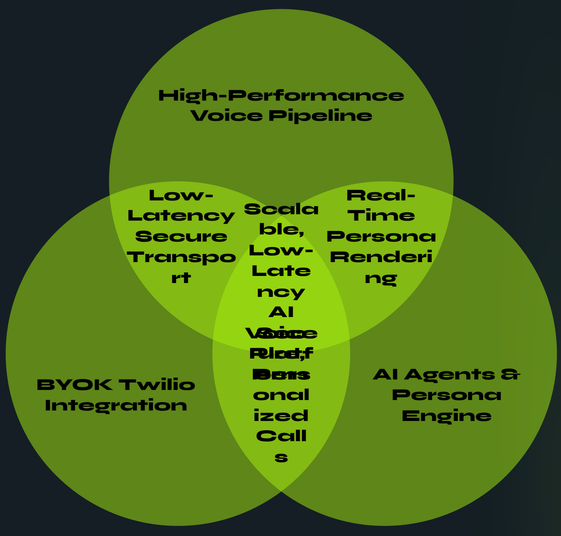
- Owners/Managers: Can create, edit, delete, and publish agents; approve prepared jobs; and manage "one-click" templates
- Developers: Can edit drafts and run tests, but cannot publish agents or authorize communications
- Viewers: Read-only access across the system

Audit & Performance

Every NOVA change and agent lifecycle event is recorded with the acting user. For voice agents, the system targets a p95 first-audio response of ≤2.2 seconds, with streaming turns meeting tier-specific latency targets. NOVA schedules batches to respect per-organization and per-agent concurrency limits.

Voice Calls — Outbound + Inbound

Astrix BOS provides a comprehensive voice calling system for both outbound campaigns and inbound call handling. All voice interactions rely on the organization's own Twilio credentials (BYOK) and can be configured for sequential or parallel calling operations.



Voice Architecture

The voice system combines several key technologies:

- **Transport:** Twilio (tenant BYOK) for dial/signaling and media
- **Brain:** LLM (OpenAI/Anthropic) for conversational turns with per-agent persona
- **TTS:** ElevenLabs (or configured alternative per agent)
- **ASR:** Whisper/Conformer (pluggable speech recognition)
- **Pattern:** Low-latency streaming pipeline with barge-in capability

Outbound Calling Flow

Preparation & Approval

1. NOVA or a user creates a prepared call batch with target leads and selected agent
2. Owner/Manager approves job → system checks BYOK, caps, budgets, quiet hours, suppression, concurrency
3. Runner dials via Twilio Calls API using the organization's number
4. TwiML points to our voice processing endpoint

Call Processing

1. Streaming loop processes audio via WebSocket
2. ASR decodes caller speech → LLM generates responses → TTS streams audio chunks
3. Barge-in (VAD) pauses TTS if callee speaks
4. On hangup: store Opus recording, transcript, optional MP3, and CRM timeline entry
5. No-answer/busy: classify result, schedule follow-up with NOVA advice

Inbound Call Handling

Organizations assign their Twilio voice webhook to our inbound endpoint. The system:

1. Returns TwiML that greets callers with brand-specific prompts
2. Respects configured quiet hours
3. Either records a voicemail or connects to an appropriate agent
4. Stores artifacts (Opus/transcript), creates CRM lead if unknown, attaches voicemail to timeline

Concurrency & Scheduling

10	20	2.2s
Max per agent	**Per org maximum**	**First-audio target**
User-configurable maximum parallel calls from 1–10 per agent, with a default of 3	Plan-tier caps limit total concurrent calls (e.g., Starter ≤10, Growth ≤15, Scale ≤20)	P95 latency from answer to first audible TTS, ensuring natural conversation flow

The runner logic computes allowed concurrency as `min(agent_max, org_max_remaining, global_remaining, budget_remaining, cap_remaining)`.

Dial Strategies:

- **Sequential (1-by-1):** Respects "human paced" calling
- **Burst (parallel up to N):** For speed; NOVA throttles to stay under rate limits
- **Retries:** Configurable (e.g., 2 with exponential backoff) for no_answer/busy/failed
- **Schedule windows:** Users select start time and windows (e.g., "evenings local time"); quiet hours enforced

Voice Quality & Latency Requirements

First-audio target

P95 ≤ 2.2s from answer to first audible TTS

Turn-to-audio target

P95 ≤ 800ms for subsequent replies (chunked TTS with jitter buffer)

Barge-in responsiveness

Voice activity detection pauses TTS within ~150-250ms

Natural prosody

ElevenLabs stability, style, and pause control; longer phrases for natural flow

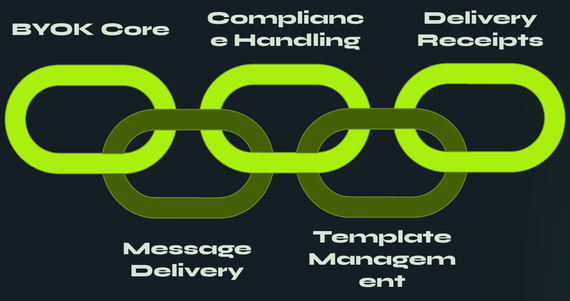
Guardrails

The system enforces strict guardrails for voice operations:

- **BYOK only:** Twilio numbers/credentials must be present & verified
- **Caps & budget:** Trial/Beta caps and Owner budget hard-stop pause jobs safely
- **Quiet hours & consent:** Enforce quiet hours for bulk; check DNC/suppression pre-dial
- **RBAC:** Only Owner/Manager can approve/run/cancel jobs
- **No Astrix numbers:** Every call uses the organization's numbers; Astrix never provides numbers

SMS Communications

The SMS system enables organizations to send individual and bulk text messages using their own Twilio credentials. It provides robust handling of delivery receipts, opt-out management, and compliance with messaging regulations.



Transport & BYOK Requirements

All SMS communications use Twilio Messaging with organization-specific BYOK credentials. The system prefers a Messaging Service SID but can also work with a configured From number. Astrix never provides numbers, even for trial or beta customers.

Sending Modes

Single Send

Allows composing and sending a message to an individual lead. Authorized by role, with configurable permissions for Developers.

Bulk Send

Created as a prepared job requiring Owner/Manager approval. NOVA can prepare these jobs with justification, but nothing sends until explicitly approved.

Templates & Message Handling

SMS agents own templates with merge variables (e.g., {{first_name}}, {{agent_name}}). The system:

- Detects GSM-7 vs Unicode encoding and computes segments and price estimates
- Enforces maximum length per message or auto-splits according to organization policy
- Shapes traffic according to organization tier CPS (messages/second) limits
- Provides soft warm-up for new senders/domains and optional per-carrier pacing

Delivery Receipts & STOP/START Handling

Delivery Receipt Processing

Processes Twilio status callbacks through the full message lifecycle: queued→sent→delivered / undelivered / failed. Updates lead timeline and analytics with current status.

STOP/START & Suppression

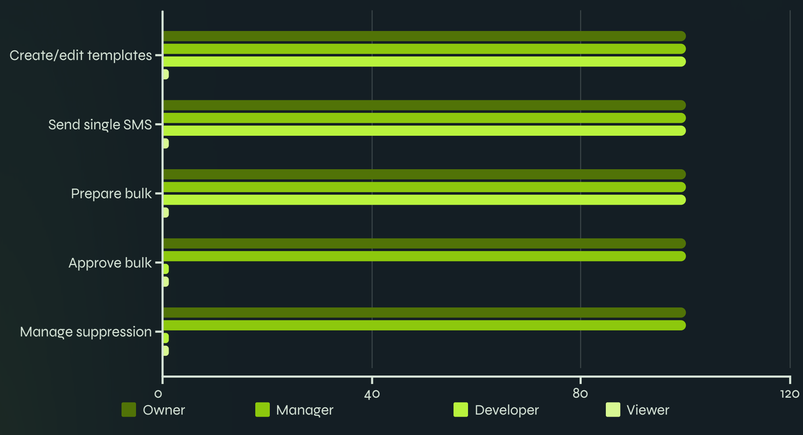
If an inbound message contains STOP keywords, the system adds the number to the suppression list and blocks further SMS until START/UNSTOP is received. Suppression checks occur pre-queue.

Compliance & Protection

The system enforces several protective measures:

- **Quiet hours:** Hard block on bulk messages outside quiet hours; configurable warn/block for one-off messages
- **Geographic restrictions:** Optional per-country allowlist for compliance
- **Rate shaping:** CPS limits by tier (e.g., Trial=0.5, Starter=1, Growth=5, Scale=10)
- **Concurrency control:** Limits outstanding Twilio API calls to avoid 429 errors

Role-Based Access Control



NOVA Integration

NOVA prepares bulk SMS jobs with:

- Rationale and supporting evidence
- Target selection (query or explicit list)
- Template choice
- Schedule window
- Estimates (segments, cost)

NOVA never automatically sends messages; all prepared jobs require explicit approval from an Owner or Manager.

Artifacts & Timeline

The system stores outbound and inbound messages on the lead timeline with DLR state, error codes, and links to the message body. This creates a comprehensive record of all SMS interactions for each lead.

Email Communications

The email system enables organizations to send individual and bulk emails using their own SendGrid credentials. It provides comprehensive tracking, deliverability monitoring, and suppression management.

Transport & BYOK Requirements

All email communications use SendGrid with organization-specific BYOK credentials. The system requires an API key and a verified sender (domain or single sender email). Astrix never sends from its own domain, including for trial or beta customers.

Sending Modes

Single Send

Allows composing and sending an email to an individual lead. Access is controlled by role, with configurable permissions for Developers.

Bulk Send

Created as a prepared job requiring Owner/Manager approval. NOVA can prepare these jobs with justification, but nothing sends until explicitly approved.

Templates & Content Management

Email agents own HTML templates with variables (Handlebars/{{first_name}}), plus subject, from name, reply-to, and preheader. The system provides:

- Template preview rendering with sample lead data
- Support for inline images and attachments
- Sanitized HTML capture for timeline display
- Deliverability guidance including domain auth status (SPF/DKIM/DMARC)

Events & Tracking

Comprehensive Event Tracking

Processes SendGrid Event Webhook for: processed, delivered, open, click, deferred, dropped, bounce, spamreport. Updates per-message state and lead timeline idempotently.

Suppression Management

Maintains a suppression list for unsubscribes, bounces, and spam reports. Bulk emails include List-Unsubscribe headers and footers. Single sends respect suppression pre-queue. Provides START/RESUB flow via tracked links.

Compliance Protection

Enforces quiet hours for bulk emails. Single sends display warnings or blocks per organization policy. Supports optional country allowlists for compliance.

Rate Shaping & Delivery Optimization

The system implements Messages Per Minute (MPM) pacing by tier:

- Trial: 10 MPM
- Starter: 60 MPM
- Growth: 300 MPM
- Scale: 1000 MPM

For new senders, the system provides warmup pacing with gradual ramp-up. If high deferral, bounce, or spam report rates are detected, the system automatically throttles and flags issues in the Ops Report.

Role-Based Access Control

Action	Owner	Manager	Developer	Viewer
Create/edit email templates	✓	✓	✓	▪
Send single email	✓	✓	✓ (configurable)	▪
Prepare bulk	✓	✓	✓	▪
Approve bulk	✓	✓	▪	▪
Manage suppression	✓	✓	▪	▪

NOVA Integration

NOVA prepares bulk email jobs with:

- Rationale and supporting evidence
- Template selection
- Target audience definition
- Schedule planning
- Warmup plan for new senders
- Estimates (messages and retail cost)

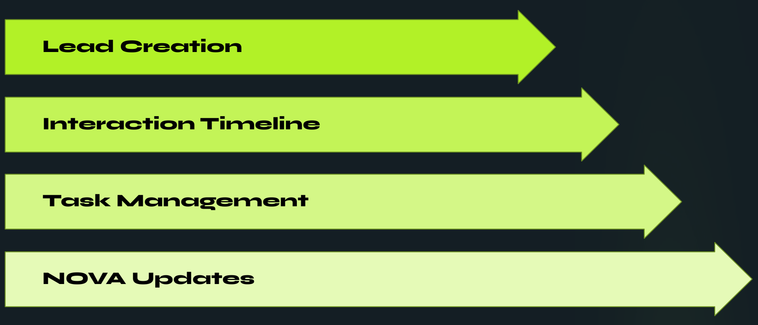
As with all channels, NOVA never sends automatically; all prepared jobs require explicit approval from an Owner or Manager.

Timeline & Analytics Integration

Every outbound email, inbound reply, and tracking event is added to the lead timeline. Owner Analytics consumes aggregated metrics to provide insights on email performance, including delivered percentage, open rate, click rate, and bounce/spam metrics.

CRM System

The CRM system is the central repository for contact information, interaction history, and task management. It integrates deeply with NOVA, which acts as the primary editor for non-contact fields.



Core Lead Management

The system provides complete CRUD operations for leads with strict tenant isolation. It normalizes contact fields (E.164 phone, lowercase email) and assigns a unique lead_id to each contact.

Standard Pipeline

Leads progress through standard pipeline stages:

- New
- Contacted
- Qualified
- Won/Lost/No Answer

Free-form tags complement stages for flexible categorization. NOVA can automatically change stage/tag with full audit trail.

Ownership & Assignment

Each lead can have:

- An owner_user_id (the responsible team member)
- Per-channel preferred agent (voice/sms/email)

Batch assign/unassign operations are available from list views.

Tasks & Follow-ups

The system supports per-lead tasks (todo, follow_back, reminder) with due timestamps and assignee tracking. NOVA automatically creates "Follow-Back" tasks after voicemail/no-answer events or certain types of replies, ensuring consistent follow-up.

Timeline

An append-only timeline combines all interaction types:

- Voice calls (with links to recordings and transcripts)
- SMS (outbound/inbound with delivery status)
- Email (sent/received with tracking events)
- Notes (manual user entries)
- Files (uploaded documents)

Compliance Controls

Suppression & Consent

Per-channel suppression flags (voice/sms/email) track opt-out status. STOP, unsubscribe, spam, and DNC flows update CRM flags and block communications pre-queue.

Quiet Hours & Timezone

The system stores timezone at the lead level (with fallback to organization timezone). Bulk jobs respect quiet hours; one-off communications follow organization policy (warn/block).

Imports & Data Management

The system supports:

- CSV upload with field mapping
- Google Sheets "published CSV URL" import
- Idempotent processing by external_id or normalized contact fields
- Export of any filtered view to CSV

Search & Segmentation

Fast filters enable searching by:

- Stage, owner, tags
- Assigned/unassigned status
- Suppression status
- Activity date ranges
- Country, score

Views can be saved for future use and as sources for prepared jobs.

NOVA as Primary Editor

NOVA automatically applies non-contact edits:

- Stage progression
- Tag application/removal
- Task creation
- Owner/allocation suggestions

Every edit is audited with why/evidence/confidence metrics. NOVA also provides hygiene functions like de-duplication suggestions, invalid contact detection, and timezone inference.

Role-Based Access Control

- **Owner/Manager:** Full CRM edit, batch assign, prepare/approve jobs
- **Developer:** Edit leads, create tasks, prepare jobs (cannot approve)
- **Viewer:** Read-only access

Owner Analytics

Owner Analytics provides organization owners with comprehensive visibility into usage, performance, and costs. This dashboard focuses exclusively on retail metrics without exposing internal supplier costs or markup data.

Scope & Focus

The analytics system is tenant-visible and retail-only, showing metrics for a single organization. It strictly avoids displaying any supplier cost information, markup data, or internal CTO/HQ details.

Usage Metrics

Tracks voice_minutes, sms_segments, email_messages, and optional tts_seconds and llm_tokens. Counts are available at daily, weekly, monthly, and month-to-date intervals.

Quality & Latency

Monitors voice metrics (first-audio p50/p95, call success rate, talk time), SMS delivery rates and latency, and email performance (delivered%, open%, click%, bounce%, spam%).

Conversion Funnels

Visualizes outbound attempts → delivered/answered → positive replies/qualified → meetings/closed-won (when CRM stage mapping exists).

Budget Management

Each organization can configure:

- Monthly budget (budget_monthly_usd)
- Soft alert percentage threshold (e.g., 80%)
- Hard stop percentage (e.g., 100%)
- Auto-resume on cycle toggle

When the hard stop threshold is reached, the system pauses all communications while keeping CRUD operations and analytics available.

Retail Pricing & Time Zones

The system uses immutable per-cycle retail price snapshots for voice_min, sms_segment, email_msg, and optional TTS/token rates. All rollups are computed in the organization's timezone, though the UI allows switching display timezone.

NOVA Strategy Board

The analytics dashboard includes a NOVA Strategy Board containing cards with:

- Title and rationale
- Supporting evidence
- Expected impact
- Quick actions (Preview template, Prepare job, View impacted leads)

Clicking "Prepare job" creates a prepared_job but never automatically sends.

Data Freshness & Exports

Usage data updates near-real-time via emitters/webhooks with idempotency protection to avoid double counting. The system supports CSV exports for usage and KPIs (retail only).

Role-Based Access Control

Owner

Full access to all analytics features and budget controls.

Manager

View access to analytics and export capabilities but cannot modify budget settings.

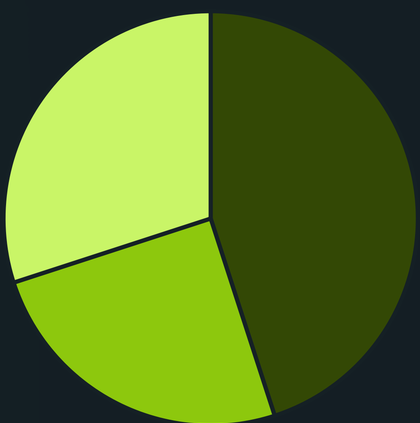
Developer

Read-only access to analytics without billing controls.

Viewer

Read-only access to a limited analytics view without export capabilities.

Budget Enforcement



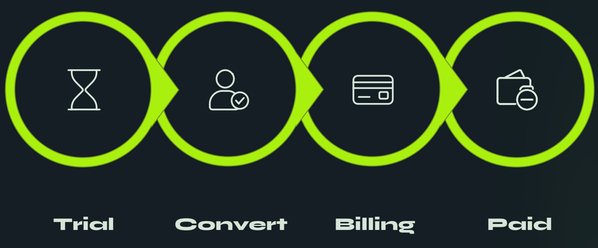
When a budget reaches the hard stop threshold:

- All emission attempts return `402 budget_hard_stop`
- Bulk jobs pause automatically
- Analytics and CRM functionality remain fully available
- The system can auto-resume on the next billing cycle or after manual budget adjustment

For Trial/Beta organizations, the system shows "Est. retail value" rather than actual invoiced amounts, displays caps progress bars, and hides the invoices section.

Plans, Billing & Trial Management

Astrix BOS provides a comprehensive billing and subscription management system with support for trials, beta access, and paid plans. The system maintains strict separation between tenant and administrative realms.



Realms & Access Separation

Tenant Realm (aud:"tenant")

Organization-specific access to plans, trial status, subscription details, budgets, and invoice viewing.

HQ Realm (aud:"admin")

Administrative-only access to plan catalog management, pricing control, global caps, beta organization creation, and supplier/markup configuration.

Plan Catalog

The system defines several tiers (Starter/Growth/Scale) with the following attributes:

- Per-month included usage (voice minutes, SMS segments, email messages)
- Overage rates (retail)
- Concurrency limits (max parallel calls)
- Rate limits (CPS/MPM)
- Feature toggles

Price snapshots are created per billing cycle and remain immutable once set.

Subscription Management

Each organization has one active subscription (Trial/Beta/Paid) with a defined state machine:

trial → active → past_due → unpaid → canceled

The system stores Stripe customer and subscription IDs for ongoing billing management.

Free Trial Structure

Trials are limited by two mechanisms:

1. Time limitation (e.g., 14 days, configurable by CTO)
2. Usage caps (per-channel limits)

When either limit is exceeded, communications are blocked with either `402 trial_cap` or `410 trial_expired`. Owners can upgrade to a paid plan at any time.

Beta Program (CTO-controlled)

Beta organizations can only be created by the CTO through the HQ interface. Beta accounts:

- Show $0 invoices
- Enforce strict usage guardrails
- Operate under a global Beta capacity limit
- Still require BYOK for all communications

The CTO can convert Beta organizations to Trial or Paid status manually with full audit tracking.

Billing Integration (Stripe)

Customer Experience

The system provides hosted checkout and billing portal access through Stripe. Invoice states are mirrored from Stripe webhooks, with email receipts and dunning handled by the Stripe platform.

Webhook Processing

The system processes Stripe events including invoice.created/paid/finalized/payment_failed and customer.subscription.updated to maintain subscription state synchronization.

Budget Controls

As detailed in the Owner Analytics section, organizations can set monthly budgets with soft and hard stop thresholds. When the hard stop is reached, communications pause with `402 budget_hard_stop` until the next cycle or manual adjustment.

Caps Enforcement

The system enforces usage caps at multiple points:

- Prepare time (estimate check)
- Approve time (pre-execution validation)
- Emit time (real-time counter check)

Standard reason codes include `trial_cap`, `trial_expired`, `beta_cap`, `budget_hard_stop`, `plan_cap`, and `payment_required`.

Support Center

Organizations can open support tickets through the tenant interface. HQ administrators can view all tickets, triage issues, and respond. Email notifications use platform mail (support@astrixbos.com) rather than tenant BYOK credentials.

Privacy & Audit

All plan changes, budget edits, trial/beta transitions, Stripe events, and ticket actions are fully audited. Supplier cost and markup data exists only in HQ storage and UI, never appearing in tenant payloads.

Security & Compliance

Astrix BOS implements comprehensive security controls across authentication, data protection, communications, and infrastructure. These measures ensure tenant isolation, regulatory compliance, and robust protection of sensitive information.

Realms & Isolation

The system defines two distinct JWT audiences:

- `tenant` (astrixbos.com)
- `admin` (hq.astrixbos.com)

Every tenant route requires both a valid JWT and `X-Org-ID` header. Mismatches result in `403 organization_mismatch` errors. The system prevents cookie sharing across realms, maintaining strict separation.

Secrets Management

Secrets at Rest

Organization BYOK credentials are stored under `data/{org}/secret/*.json` and encrypted with AES-256-GCM envelope encryption. Each secret uses a random data key, wrapped by the `ENVELOPE_MASTER_KEY` from the environment.

Secrets in Flight

Secrets are never logged or transmitted in clear text. Read operations return masked values (e.g., `AC********789`). Even debug logs cannot unmask sensitive values.

Step-Up Authentication

TOTP verification is required for sensitive operations:

- Revealing secrets
- Rotating keys
- Changing user roles
- Deleting organizations

When TOTP is required but missing, the system returns `401 step_up_required`.

Webhook Security

The system verifies each provider's webhooks using that organization's specific secret:

- Twilio: X-Twilio-Signature validation
- SendGrid: Event Webhook signature verification
- Stripe: Webhook signature check (HQ realm)

Invalid signatures result in `401 invalid_signature` responses. The system processes webhooks idempotently by tracking event IDs.

Transport Security

The system enforces HTTPS-only communication with HSTS preload and A-grade TLS configuration. Security headers include:

- Strict-Transport-Security
- X-Frame-Options: DENY
- X-Content-Type-Options: nosniff
- Content-Security-Policy with default-deny and explicit allowlists

Authentication Controls

Login and sensitive endpoints are rate-limited with IP throttling and user lockouts after repeated failures. JWTs have short TTL with optional refresh policies.

Resource Protection

The system implements per-route throttles (e.g., SMS send CPS, email send MPM) with `429` responses on excess and jittered backoff for retries.

Consent & Policy Enforcement

DNC/STOP/UNSUB preferences are honored at pre-queue time. Quiet hours are strictly enforced. Violations return `409` responses with specific reasons (`quiet_hours`, `suppressed`).

Data Retention

Data Type	Retention Period
Opus recordings	30–90 days (configurable)
MP3 caches	≤7 days
Transcripts	1–2 years
Audit logs	≥1 year

A purger job runs nightly to remove expired artifacts via atomic delete operations.

PII Minimization

The system stores normalized E.164 phone numbers and lowercase emails. PII is masked in logs, and transcripts can be redacted using the `redactions[]` array.

Backup & Disaster Recovery

Nightly backups capture the file tree (`/data/{org}`) and optional database dumps. Backups are stored off-box in encrypted format. The system conducts quarterly restore drills to verify recovery processes.

Monitoring & Alerts

The system provides health probes at `/healthz` (liveness) and `/readyz` (dependencies ready). It implements uptime checks, error rate alerts, quota/budget alerts, and webhook failure notifications.

Audit Trail

An append-only audit captures key events including:

- Authentication events
- Invitations and role changes
- Budget and plan modifications
- NOVA edits
- Job preparation and approval
- Webhook processing
- Retention deletions

Each audit entry includes timestamp, request ID, user/NOVA identifier, organization ID, action, target, and result.

Deployment Architecture

Astrix BOS implements a resilient deployment architecture designed for zero-downtime updates, strong security boundaries, and consistent availability. The system uses a blue-green deployment model behind a secure reverse proxy.

Topology & Realms

Component Arrangement

Component	Hostname/Path
Nginx (reverse proxy)	astrixbos.com, hq.astrixbos.com
Tenant App	astrixbos.com → /api/* & FE
Admin HQ	hq.astrixbos.com → /admin/* & FE
Blue/Green backends	Unix sockets
Shared data	/opt/astrixbos/shared/data

Communication Paths

Nginx routes requests to the active backend (Blue or Green) through Unix sockets:

- /run/astrixbos-blue.sock
- /run/astrixbos-green.sock

Only one backend is active at a time, allowing zero-downtime switching during deployments.

Nginx, TLS & Security Headers

The system uses Let's Encrypt (certbot) for TLS with automatic renewal and HTTP→HTTPS redirects. Security headers are applied consistently across both realms:

- HSTS with preload
- X-Frame-Options: DENY
- X-Content-Type-Options: nosniff
- Strict Content-Security-Policy

Nginx is configured with two upstreams (blue/green) via Unix sockets, health-checked with the `/healthz` endpoint. Rate limits are applied for login attempts and webhook processing.

Blue-Green Release Workflow

1. Build Docker images with version tag vYYYY.MM.DD.N
2. Stage new release under /opt/astrixbos/releases/vX/
3. Bring up as Green: `docker compose -p astrix-green up -d`
4. Health soak: verify `/healthz` & `/readyz` for tenant and HQ
5. Flip Nginx to Green (swap upstream symlink & reload)
6. Soak window (5-15 minutes)
7. Decommission Blue (keep running for fast rollback until confident)

Rollback requires only flipping the symlink back to Blue and reloading Nginx.

Process Management & Health Monitoring

The application runs under Gunicorn with multiple Uvicorn workers (-w 4) and graceful timeout configuration. Health endpoints `/healthz` (liveness) and `/readyz` (dependencies ready) allow Docker to monitor container health and restart unhealthy instances.

All logs are structured JSON sent to stdout/stderr, allowing Docker to act as the collector and maintain stateless containers.

Network Security

The system implements:

- Firewall (UFW) allowing only ports 80/443, blocking application ports
- Optional WireGuard VPN (UDP 51820) for secure HQ access
- SSH restricted to CTO IPs or VPN/jump access
- Non-root containers with least privilege
- Kernel hardening with disabled packet forwarding (except WireGuard) and SYN/ICMP protections

Admin HQ Access Controls

Option A: HQ-over-VPN (Recommended)

HQ listens only on the WireGuard interface (10.8.0.0/24). Nginx restricts access to specific WireGuard IPs including CTO laptop, CTO phone, and optional office gateway. This provides stable access even if public IPs change and establishes cryptographic device identity.

Option B: IP Allowlist

Nginx restricts access to specific allowlisted IPs with a quick-swap configuration via Ansible for emergency access. This approach is simpler but more brittle on IP changes, and is generally recommended as a backup to the VPN approach.

Both options are reinforced with:

- MFA requirement (WebAuthn key and TOTP step-up for sensitive operations)
- Short JWT TTL with optional refresh
- Admin session timeout after inactivity
- Optional device posture verification via mTLS on VPN

CTO Access Continuity Plan

To ensure the CTO is never locked out of the system, the deployment includes:

1. **Primary access:** WireGuard profiles on CTO laptop and phone, stored in a secure vault
2. **Secondary access:** Emergency WireGuard keypair (printed QR + file) with Ansible playbook to temporarily grant access
3. **Identity backups:** Two WebAuthn hardware keys, 10 one-time recovery codes, and encrypted TOTP secret
4. **DNS resilience:** Low TTL for hq.astrixbos.com with pre-provisioned standby domain
5. **Monitoring:** External uptime checks with automated alerts for HQ unavailability

Backups & Disaster Recovery

The system implements:

- Nightly encrypted backups of organization data, configurations, and optional database dumps
- Off-box synchronization to object storage in a different region
- Tiered retention (14 days daily, 8 weeks weekly, 12 months monthly)
- Quarterly restore drills to verify recovery process
- Comprehensive disaster runbook with standby VPS procedures

Integrations & Webhooks Hub

The Integrations & Webhooks Hub processes incoming provider notifications, normalizes them into a consistent format, and distributes them to appropriate system components while maintaining strict tenant isolation.

Provider Support

The system processes webhooks from multiple providers:

- Twilio Voice (status and media events)
- Twilio SMS (delivery receipts and inbound messages)
- SendGrid (email tracking events)
- Stripe (billing webhooks - HQ realm only)

Security & Tenant Mapping

Signature Verification

The system verifies all incoming webhooks using the organization's own secret (BYOK) for Twilio and SendGrid, or the HQ secret for Stripe. Failed verification results in 401 `invalid_signature` without any state changes.

Organization Identification

The system derives the organization ID from verified context: Twilio Account SID / Messaging Service SID mapped to organization; SendGrid signed event and custom args; Stripe customer_id mapping. It never trusts unverified claims in query parameters or request body.

Event Normalization

Raw provider payloads are converted into internal events with a common schema, then distributed to consumers including:

- CRM timeline
- Usage counters
- Quality/latency metrics
- Caps/budgets

Idempotency & Reliability

The system deduplicates events by event_id (provider unique ID or SHA-256 hash of canonical payload). Duplicate receipts are processed as no-ops. For downstream errors, events are queued for retry with exponential backoff (1m, 5m, 15m). After multiple attempts, failed events are moved to a dead-letter queue.

Dead-Letter & Replay Management

The system maintains a file-backed dead-letter queue per organization. Organization owners can replay safe events (delivery receipts, delivered/open status). Restricted events like billing can only be replayed from the HQ interface.

Performance & Back-Pressure

Webhook handlers perform O(1) writes, with heavy work (audio processing, transcripts) deferred to a worker pool. The system applies per-organization concurrent limits to prevent noisy neighbor problems.

Observability & Performance

The system generates structured logs with request_id, org_id, provider, type, status, latency, and deduplication status. It maintains counters for success, error, and retry operations. The webhook processing pipeline targets p95 ≤ 2s (excluding deferred heavy work) with zero event loss thanks to idempotency and the dead-letter queue.

Compliance Integration

The webhook system feeds compliance systems:

- STOP/UNSUB events update suppression lists
- Quiet hours violations are flagged
- Bounce/spam reports trigger email suppression

These updates take effect before future communication attempts.

Schema Versioning

The internal event schema includes a version field (e.g., v1) with migration stubs to ensure backward compatibility. This prevents breaking older events during replay operations.

Real-World Scenarios

Carrier Delivery Without Reply

"Delivered" status is recorded and delivery percentage computed, enabling NOVA to suggest appropriate follow-up windows.

Opt-Out Processing

When a user texts "STOP", the number is instantly added to the suppression list. Future SMS attempts return 409 `suppressed` with a clear reason.

Email Spam Complaint

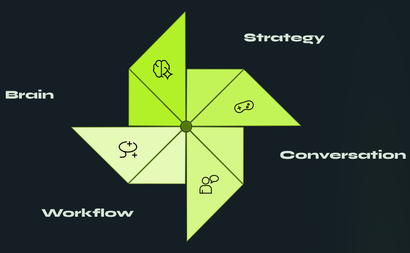
The contact is marked as suppressed with reason "spam". NOVA warns in the Ops Report and bulk email operations exclude the contact.

Provider Retries

The system deduplicates by event_id, preventing duplicate counts in analytics and timeline entries.

NOVA - Internal Ops Brain & Voice Copilot

NOVA serves as both an internal operations intelligence system and a conversational assistant within Astrix BOS. It provides automated insights, prepares actions for approval, and offers an intuitive voice interface optimized for fluency and low latency.



Scope & Capabilities

NOVA focuses on internal operations including CRM edits, analytics explanations, strategy development, prepared job creation, and draft agent suggestions. Customer outreach still requires BYOK credentials and explicit approval.

Access follows standard RBAC patterns:

- Owner/Manager: Full access with approval rights
- Developer: Preparation, drafting, and testing without approval
- Viewer: Ask questions and view information only

NOVA operates exclusively in the Tenant realm (aud:"tenant") with no access to supplier/markup information.

System Capabilities

Primary CRM Editor

Safely automates stage/tag/task/owner allocation and hygiene fixes. Every automated edit includes audit with rationale, evidence, and confidence metrics.

Strategy & Preparation

Creates prepared communication jobs with targets, schedule, and estimates - all requiring explicit approval before execution.

Draft Agent Studio

Suggests new AI agents based on observed KPI gaps, creating drafts with appropriate persona, constraints, and templates.

Ops Report & Strategy

Generates daily KPIs, anomalies, and strategy cards that enable one-click job preparation.

Draggable Blob Interface

NOVA appears as a floating widget ("Blob") on every tenant page:

- Draggable and position-persistent per user
- Controls for push-to-talk, mute, captions, quick-actions
- Context-awareness based on current page
- Privacy toggles for transcript storage (default OFF)
- Accessibility features including captions and keyboard shortcuts

Voice Pipeline Performance

Stage	Target p95	Optimization Techniques
Capture → server	≤40–70ms	20ms PCM frames, small send buffer, Nagle disabled
ASR first partial	≤150–250ms	Streaming decoder, endpointing threshold
LLM first token	≤400–700ms	Prompt cache, compact context, function-calls
TTS first audio	≤300–500ms	Pre-request voice session, SSML punctuation
First audible reply (total)	≤1.8–2.2s	Sum of the above (95th percentile)
Turn-to-turn latency	≤600–800ms	Incremental decoding, barge-in, jitter buffer

Voice Fluency Architecture

Sentence-aware Chunking

Avoids micro-chunks of 2–3 words, instead emitting phrase-sized units (~600–1200ms) for natural speech flow.

Prosody Shaping

Uses SSML or provider parameters for pauses, emphasis, and breathiness with customizable pitch/rate per agent.

Stabilized Punctuation

Applies punctuation restoration with a short stability window (40–70ms) before sending TTS chunks to avoid erratic speech patterns.

Cross-fade Playback

Implements 15–25ms overlaps between chunks to eliminate audible seams in speech output.

Barge-in Support

Voice activity detection with 150–250ms threshold pauses TTS when the user speaks, enabling natural interruptions.

Speech Recognition & LLM Generation

ASR Features

- Streaming decoder with partials every ~100–150ms
- Adaptive end-of-utterance detection (300–600ms silence)
- Intelligent text normalization for numbers/dates
- Noise reduction and beam search optimization
- Hotword boosting for entity recognition

LLM Optimization

- Compact context (≤4–6 KB) with summarized brand information
- System+brand prompt caching by organization and page type
- Function-call schema for deterministic actions
- Optional speculative first-token completion
- Explicit stop sequences and token limits
- Safety guardrails preventing direct communication

Text-to-Speech Quality

The system maintains high-quality speech output through:

- Warm voice sessions per user (30–60s keep-alive)
- Phrase marshaling (1–2 sentences at a time)
- Stutter prevention for re-punctuated content
- Adaptive jitter buffer (250–350ms initial, adjusting with network conditions)
- Configurable voice tuning parameters stored per organization/agent

Security & Guardrails

No Automatic Sending

NOVA only prepares communications; Owner/Manager approval is always required before sending.

BYOK Enforcement

Any communication attempt requires organization BYOK credentials, returning `400 configure_integrations` if missing.

Compliance Adherence

Caps, budgets, quiet hours, and suppression lists are enforced at preparation, approval, and execution stages.

Privacy Protection

Audio is not retained by default; transcripts are optional; PII masking and redaction are applied to saved interactions.

NOVA Dashboard

The NOVA Dashboard provides a centralized interface for:

- Key performance indicators and warnings
- Strategy Board with actionable cards
- Prepared jobs queue with approve/pause/cancel functions
- Draft Agent Studio for agent testing and publishing
- Operations reports with trend analysis
- NOVA configuration settings

Settings & Onboarding

The Settings & Onboarding system provides comprehensive organization management, user administration, integration configuration, and testing capabilities. It guides new organizations through setup while enforcing security and compliance requirements.

Onboarding State Management

Each organization has a checklist tracking setup progress with pass/fail status for essential steps:

- Organization Profile
- BYOK Twilio
- BYOK SendGrid
- Test SMS/Email/Call
- Quiet Hours configuration
- Budget setup
- Users/Roles
- Brand KB
- Webhook Verification

Organization Profile

The profile captures essential organization settings:

- Name
- Timezone
- Quiet-hours window(s)
- Country allowlist
- Legal footer strings for SMS/Email

Integrations (BYOK)

Twilio Configuration

Organizations enter their Account SID, Auth Token, and either Messaging Service SID or From number. Credentials are saved using envelope encryption and displayed in masked form (AC•••789) when read.

SendGrid Configuration

Organizations provide their API key and verified sender information. As with Twilio, these credentials are securely stored and masked on display.

Verification Process

On save, the system performs live checks including credential validation, send permission verification, and webhook signature self-testing to confirm proper organization mapping.

Test Bench

The Test Bench allows organizations to verify their integration setup by sending test communications to their own numbers/emails. Tests honor quiet hours restrictions and create appropriate timeline entries. Voice tests save Opus recordings, transcripts, and short-lived MP3 caches.

Users & Roles Management

Organizations can invite users via email with specific roles (Owner/Manager/Developer/Viewer). The system enforces two-factor authentication (TOTP / WebAuthn) and provides controls to disable or remove users as needed.

Brand KB for NOVA

A Markdown editor allows organizations to define their brand knowledge base, stored as nova/brand_kb.md. This document is separate from agent KBs and influences how NOVA communicates. A preview function shows how NOVA will speak using the defined brand guidelines.

Budgets & Compliance

Budget Controls

Organizations can set monthly budgets with soft/hard stop thresholds. The interface displays Trial/Beta/Plan caps meters and remaining balances.

Compliance Status

The system shows A2P 10DLC status (from Twilio), email domain authentication badges (SPF/DKIM/DMARC), suppression list management, and DNC CSV import capabilities. Bulk tools remain blocked until compliance requirements are met or explicitly overridden by Owners.

Key Rotation

The system supports zero-downtime credential rotation:

1. Add new key
2. Verify functionality
3. Swap active credential
4. Revoke old key

This process requires TOTP step-up authentication and preserves continuous operation.

Danger Zone

Sensitive operations are grouped in a protected "Danger Zone" section:

- Moving agents to Draft state
- Deleting agents
- Exporting organization data
- Closing organization

These functions require multi-step confirmation and respect RBAC permissions.

Role-Based Access Control

100..

Owner Access

Full access to all settings and controls

90%

Manager Access

Can manage users (except Owners), approve campaigns, edit settings (no secrets reveal)

50%

Developer Access

Can prepare/test, edit templates/KB; no approvals or secrets reveal

0%

Viewer Access

Read-only access across the system

Audit Trail

Every configuration change generates an append-only audit entry including:

- BYOK credential updates
- Verification attempts
- Test executions
- User invitations
- Role changes
- Credential rotations
- Data exports

Each entry captures timestamp, request_id, user_id, action, target, and result.

First-Run Experience

New organizations follow a guided wizard flow:

1. Organization Basics → Name, timezone, quiet hours
2. Connect Twilio (BYOK) → Paste keys → Verify
3. Connect SendGrid (BYOK) → Paste key/sender → Verify
4. Test Bench → SMS / Email / Call tests
5. Compliance → View A2P & Domain status
6. Budget & Caps → Set budget; view limits
7. Invite Team → Add users with appropriate roles
8. Brand KB → Define voice and personality

Once all required steps are complete, bulk tools and NOVA action cards become available.

Artifacts & File Storage

The Artifacts & File Storage system manages recordings, transcripts, assets, and exports with secure access controls, efficient storage formats, and automated retention policies.

Voice Artifacts

Call recordings are stored in space-efficient .opus format as the primary source. Optional .mp3 previews are generated on-demand and cached short-term. Transcripts are stored as JSON with speaker segmentation and optional redaction.

Email & SMS Assets

Inline images and attachments are stored per organization and served via short-TTL signed URLs. The system enforces MIME-type allowlists for security.

CRM Files & Exports

File uploads (PDF, images, CSV) are tied to specific leads with validation and scanning. CSV exports for CRM views and analytics are generated on-demand with fast-expiring download links.

Access Control

The system implements two-layer access protection:

1. API authorization requiring JWT (aud:"tenant") + X-Org-ID validation
2. File-specific HMAC signed URLs with claims: {org_id, path, action, exp}

This prevents unauthorized access even if file paths are known.

Performance Optimization

Files are served with several performance enhancements:

- Range request support for audio scrubbing
- Streaming responses to minimize memory usage
- ETag support for caching
- Optional Nginx/X-Accel for zero-copy serving

Retention & Purging

The system enforces configurable retention policies:

- Opus recordings: 30-90 days
- MP3 cache: ≤7 days
- Transcripts: 1-2 years
- Raw webhook inbox: ≤48 hours
- Normalized events: 30-90 days

A nightly purger job performs atomic deletions according to these policies.

Audit & Security

Every artifact operation (create/read/delete) is logged with request_id, org_id, optional lead_id, path, action, and status. The system enforces several security measures:

- Content-type and size limits
- Optional anti-virus scanning
- Image EXIF data scrubbing
- No executable file uploads

Tenant Isolation

All file paths follow the pattern `data/{org_id}/...` ensuring strict separation between organizations. Signed URLs are scoped to specific organizations, preventing cross-tenant access.

API Contract

Signed URL Creation

To access a file, clients first request a signed URL:

```
POST /api/files/sign
{
  "path": "voice/recordings/{call_id}.opus",
  "action": "get",
  "ttl_s": 300
}
```

The system responds with a time-limited access URL.

File Access

The signed URL contains all necessary validation information:

```
GET /api/files/get?
org=...&path=...&action=get&exp=...&sig=...
```

The system validates the HMAC, organization, expiry, and requested path before streaming the file with appropriate headers.

Upload Flow

For file uploads (CRM documents or email assets), the system uses a three-step process:

1. Client requests an upload URL via `POST /api/files/upload-url`
2. Client directly uploads to the returned signed URL
3. Client finalizes with `POST /api/files/finish` including metadata (size/hash/mime)

Signed URL Implementation

Each signed URL token contains:

- org_id: The organization identifier
- path: The file path under the organization's data folder
- action: The permitted operation (get, put)
- exp: Expiration timestamp
- version: For key rotation support

These elements are combined into a canonical string and signed with HMAC using a per-deployment signing secret. For sensitive documents, the system can optionally enforce single-use by tracking token IDs.

Audio Handling Optimizations

The system optimizes audio handling:

- Opus format provides smaller file size with high quality
- MP3 transcoding happens asynchronously without blocking
- Range requests enable fast in-browser scrubbing
- ETags and client caching reduce bandwidth usage

Frontend Architecture

The Astrix BOS frontend implements a clean separation between tenant and administrative interfaces, with comprehensive navigation, role-based UI adaptation, and optimized user experience flows.

Information Architecture

Tenant App (astrixbos.com)

Global elements:

- Topbar: Organization switcher, Search, User menu
- Sidebar navigation
- Floating NOVA Blob for voice/chat assistance

Main sections:

- Dashboard (Owner Analytics)
- CRM (Leads, Views, Imports)
- Agents (Voice/SMS/Email/Drafts)
- Voice/SMS/Email tooling
- NOVA Dashboard
- Billing & Settings
- Support

Admin HQ (hq.astrixbos.com)

CTO-only interface:

- HQ Home (global status)
- Plans & Pricing management
- Beta Organizations
- Supplier & Markup (private)
- Stripe integration
- Webhooks Monitor
- Security & Operations
- Support Center

Routing & Protection

The system differentiates realms by domain:

- astrixbos.com/* → TenantLayout
- hq.astrixbos.com/* → HQLayout

Route protection is implemented through Higher-Order Components:

- RequireTenant({roles?: [...]}) — verifies JWT audience, role, and injects X-Org-ID
- RequireHQ({roles?: []}) — verifies JWT audience for HQ access

UI System & Components

The frontend uses Tailwind + shadcn/ui for consistent design with specialized components:

- Virtualized tables for efficient rendering of large datasets
- Chart components for analytics visualization
- Form handling with validation
- Toast notifications mapped to backend reason codes
- Audio player with range support
- Comprehensive empty states for first-run experience

Key Page Blueprints

Dashboard

MTD spend/usage tiles, budget meter, latency stats, funnel charts, and NOVA strategy cards. Viewable by all roles; budget editing restricted to Owners.

CRM

Lead filters, saved views, preparation actions, and detailed timeline views. Owner/Manager have full access; Developer can edit; Viewer is read-only.

Agents

Agent management with testing bench and template editing. Publishing requires Owner/Manager; Developers limited to drafting/testing.

NOVA Dashboard

Strategy board, prepared jobs queue, draft agent studio, and operations reporting. Approval actions respect RBAC permissions.

UX Flows

The system implements several key user experience flows:

Onboarding

Guided wizard for organization setup, integration, testing, compliance, budget, team, and brand configuration. Must complete to unlock bulk operations.

Prepare → Approve

Workflow for creating and authorizing communication jobs with estimates, compliance checks, and role-based approval gates.

Key Rotation

Zero-downtime credential updates with verification, TOTP step-up, active swap, and old key revocation.

NOVA Interaction

Voice/text interface with natural conversation, action cards, and dashboard integration for deeper exploration.

Performance Optimization

The frontend targets specific performance metrics:

- Largest Contentful Paint ≤ 2.5s (p75) through code-splitting and lazy-loading
- Time to Interactive ≤ 3.0s (p75) by minimizing heavy libraries
- NOVA Blob first-audio ≤ 2.2s p95 using connection optimization

Additional optimizations include table virtualization, route prefetching, and efficient audio streaming.

Accessibility & Internationalization

The system prioritizes accessibility through:

- Keyboard-first navigation with focus management
- ARIA attributes for interactive elements
- Screen reader support
- Color contrast compliance (≥ 4.5:1)
- Motion reduction preferences

The UI is structured for internationalization with ICU-compatible strings and timezone-aware date/time display.

State & Error Handling

The frontend uses React Query / SWR for server state management with minimal local UI state. All API requests automatically include JWT and X-Org-ID headers.

A central error interceptor maps backend reason codes to appropriate UI responses:

- configure_integrations → Settings/Integrations link
- trial_cap/trial_expired/beta_cap → Upgrade CTA
- budget_hard_stop → Budget page direction
- quiet_hours → Schedule suggestion
- suppression → Suppression management link

Role-Based UI Adaptation

The interface dynamically adapts based on user role:

Feature	Owner	Manager	Developer	Viewer
Dashboard view	✓	✓	✓	✓
CRM edit/batch	✓	✓	✓	• (read)
Prepare jobs	✓	✓	✓	•
Approve jobs	✓	✓	•	•
Agent publishing	✓	✓	• (draft/test)	•
Billing controls	✓	• (view)	•	•
Secrets management	✓ (TOTP)	•	•	•

Quality Assurance & Testing

Astrix BOS implements comprehensive quality assurance processes with layered testing approoches to ensure functionality, performance, security, and resilience across all system components.

Test Environments

Testing occurs across three distinct environments:

Staging (stg.astrixbos.com)

Full end-to-end environment with real Twilio/SendGrid test organization keys (BYOK). Mirrors production settings and uses a controlled QA organization.

Sandbox

Automated test environment with provider stubs to avoid external calls. Uses the webhook "stub mode" normalizer and simulated signatures.

Production (Green)

Post-deployment verification with read-only tests and safe Test Bench operations targeting CTO-controlled endpoints.

Test Categories & Release Gates

Testing is organized into a hierarchy of increasingly comprehensive categories:

Level	Description	Timing	Duration
T0: Smoke	Post-deploy/flip verification	After each deployment	15 minutes
T1: Daily Sanity	Early warning on drift	Scheduled daily	45-60 minutes
T2: Pre-release Regression	Comprehensive validation	Before major rollouts	3-4 hours
T3: Performance & Fluency	Latency SLO verification	Pre-release	Varies
T4: Security/Compliance	Security control validation	Pre-release	Varies
T5: DR & Backups	Recovery verification	Quarterly	Half-day

Testing Tools

The QA process employs several specialized tools:

- **API/E2E:** Playwright for frontend flows, PyTest + requests for API testing
- **Load/Latency:** k6 for HTTP scenarios, Locust for job runners, WebSocket latency probes
- **Chaos Testing:** tc/netem for network conditions, process termination, webhook backlog simulation
- **Observability:** Structured JSON log analysis with verification of request_id, org_id, route, latency_ms

Core Test Checklists

Tenant Application Tests

- JWT + X-Org-ID validation on all requests
- BYOK storage, masking, and verification
- Test communication functions and artifact generation
- Quiet hours and budget enforcement
- NOVA Blob performance and functionality
- Signed URL security and expiration
- Analytics accuracy and budget metering

Admin HQ Tests

- VPN/IP-allowlist access controls
- MFA/TOTP enforcement
- Price snapshot management
- Beta organization administration
- Stripe webhook processing
- Webhook DLQ monitoring and replay
- Emergency access procedures

T0: Smoke Testing (Post-Deployment)

Immediately after a blue-green deployment flip, critical functionality is verified:

- Health endpoint checks and security header validation
- Realm and organization boundary enforcement
- Integration verification with BYOK credentials
- Test Bench functionality across all channels
- Prepare → Approve workflow for voice communications
- Signed URL generation and expiration
- NOVA Blob basic interaction

Any failure triggers immediate rollback to the previous color.

T1: Daily Sanity Testing

Automated daily verification covers:

- Onboarding checklist completion
- Compliance badge accuracy
- Budget enforcement
- Quiet hours restrictions
- Suppression list functionality
- Webhook idempotency
- Analytics data accuracy
- HQ access controls

T2: Pre-release Regression Testing

Comprehensive section-by-section validation includes:

- Authentication and RBAC enforcement
- Settings and onboarding flows
- CRM and agent lifecycle
- Communication channel operations
- Artifact handling and retention
- Analytics and billing accuracy
- Webhook processing
- NOVA capabilities
- Security controls

Performance & Security Testing

API Performance

HTTP/API endpoints must maintain p95 < 300ms response time with error rates below 0.1% under load

Voice Response

NOVA voice interactions must achieve first-audio p95 ≤ 2.2s with zero underruns and fast barge-in

Security Controls

All organization isolation, webhook authentication, and access control mechanisms must be 100% effective

Disaster Recovery Testing

Quarterly DR testing verifies:

- Backup creation and encryption
- Restoration to isolated environment
- Authentication and data access verification
- CTO emergency access procedures
- Retention policy enforcement

Release Criteria

Clear go/no-go criteria define release readiness:

- **Smoke (T0):** 100% pass required to continue with deployment
- **Daily (T1):** ≥ 98% pass; failures create tickets and may block risky changes
- **Pre-release (T2):** 100% pass required for functional, security, and core performance
- **NOVA Performance:** Must meet first-audio and barge-in targets with zero underruns
- **Security:** Zero secret leaks, HQ perimeter integrity, signed URL verification
- **DR:** Successful restore drill within last 90 days

Implementation Sequence & Priorities

This section outlines the recommended implementation sequence for Astrix BOS, focusing on delivering core value quickly while building toward the complete system architecture.

Phase 1: Foundation & Security

Authentication & Tenant Isolation

Implement JWT-based authentication with audience separation, organization boundaries, and the X-Org-ID validation mechanism. This forms the security backbone for all subsequent features.

Deployment Architecture

Establish the blue-green deployment pipeline, Nginx configuration, health checks, and basic monitoring. Set up the realm separation between tenant and HQ domains.

Secrets Management

Implement envelope encryption for BYOK credentials with masked reading and secure storage patterns. Add TOTP step-up for sensitive operations.

Phase 2: Core Communication Channels

Build the essential communication capabilities with strict BYOK enforcement:

SMS Implementation

1. BYOK integration for Twilio
2. Single send endpoint with template support
3. Delivery receipt webhook processing
4. STOP/START handling with suppression
5. Quiet hours and compliance enforcement

Email Implementation

1. BYOK integration for SendGrid
2. Template rendering with variables
3. Event webhook processing
4. Unsubscribe and suppression management
5. Basic analytics tracking (delivered/open/click)

Voice capabilities can follow with the streaming pipeline, ASR→LLM→TTS chain, and artifact storage. Focus initially on outbound single calls before implementing bulk functionality.

Phase 3: CRM & Agent Framework

With communication channels established, build the CRM system and agent framework:

CRM Core

Implement lead CRUD, timeline, stage/tag management, and basic search. Focus on a clean file-backed storage model with proper tenant isolation.

Agent Model

Develop the Draft→Test→Publish lifecycle, persona configuration, and channel-specific template management. Establish the RBAC model for agent operations.

Import/Export

Add CSV import with mapping, deduplication, and validation. Implement CSV exports with signed URLs.

Phase 4: NOVA & Prepared Jobs

Introduce NOVA's core capabilities:

- CRM primary editor functionality (non-contact field automation)
- Prepared job creation for all channels
- Approval workflow with RBAC enforcement
- Basic strategy suggestions
- Conversation interface foundation

Focus initially on the prepared→approve workflow before implementing the full voice interface. Ensure all operations maintain the never-auto-send principle.

Phase 5: Analytics & Billing

Build the business management features:

Owner Analytics

Implement usage tracking, quality metrics, spend calculation, and budget visualization. Ensure strict retail-only focus without supplier cost leakage.

Billing Integration

Add Stripe integration, plan catalog, subscription lifecycle, and invoice visualization. Implement trial/beta mechanics and caps enforcement.

Budget Controls

Complete the budget management system with soft/hard stops and enforcement throughout the communication pipeline.

Support Center

Add ticket creation, management, and notification system for customer assistance.

Phase 6: Settings & Onboarding

Develop the comprehensive settings and onboarding experience:

- Checklist-driven setup wizard
- Integration verification workflow
- Test bench for all channels
- User/role management
- Brand KB editor and preview
- Key rotation with zero downtime

Phase 7: Advanced Features

With the core system established, implement more sophisticated capabilities:

NOVA Voice Copilot

Complete the draggable Blob interface with optimized voice pipeline, fluency enhancements, and action card framework. Fine-tune latency to meet performance targets.

Enhanced Webhooks

Extend the webhook system with comprehensive DLQ management, replay capabilities, and tenant-safe event normalization.

Advanced Security

Implement comprehensive audit trails, retention policies, and security monitoring. Complete the DR procedures and emergency access mechanisms.

Continuous Improvement

Throughout all phases, maintain focus on:

- **Test Coverage:** Build automated tests for each component
- **Performance Optimization:** Monitor and improve latency metrics
- **Security Hardening:** Regular review of security controls
- **Documentation:** Keep technical documentation current with implementation

Prioritize features that deliver immediate customer value while maintaining the architectural integrity needed for long-term scalability and security.

API Reference & Integration Points

This section provides a consolidated reference for Astrix BOS APIs, organized by functional area with authentication requirements, expected parameters, and response formats.

Authentication & Headers

All API requests require the following:

Header	Description	Example
Authorization	JWT Bearer token with appropriate audience	Bearer eyJhbGci...
X-Org-ID	Organization identifier (tenant routes only)	org_abc123
Content-Type	Media type for request body	application/json

Realm-specific requirements:

- Tenant routes (/api/*): JWT must have aud:"tenant"
- HQ routes (/admin/*): JWT must have aud:"admin"

Authentication APIs

```
# Login
POST /api/auth/login
Request: { email, password }
Response: { access_token, exp }

# Current User Info
GET /api/auth/me
Response: { user_id, email, orgs:[{org_id,name,roles:[...]}] }

# Organization Management
POST /api/org - Create organization
GET /api/org - Get current organization profile

# User Management
POST /api/users/invite { email, role }
POST /api/invite/accept { token, password|sso_assertion }
PUT /api/users/{id}/role - Requires TOTP
DELETE /api/users/{id} - Requires TOTP
```

Agent Management APIs

```
# CRUD Operations
POST /api/agents - Create Draft Agent
GET /api/agents - List (filter by channel, state)
GET /api/agents/{agent_id} - Fetch specific agent
PUT /api/agents/{agent_id} - Edit agent

# Lifecycle Management
POST /api/agents/{agent_id}/test - Sandbox test
POST /api/agents/{agent_id}/publish - Owner/Manager only
POST /api/agents/{agent_id}/pause - Owner/Manager
POST /api/agents/{agent_id}/clone - Returns new draft
DELETE /api/agents/{agent_id} - Soft delete (Owner/Manager)
POST /api/agents/{agent_id}/revert - Revert to Draft
```

NOVA APIs

```
# Operations Report
POST /api/nova/ops-report/run - Generate ad-hoc report
GET /api/nova/ops-report/latest - Get latest report

# Draft Agents
POST /api/nova/drafts/suggest { goal?, evidence? } - Create Draft Agents

# Prepared Jobs
POST /api/nova/prep/call-batch { agent_id?, query|lead_ids, schedule?, constraints? }
POST /api/nova/prep/bulk-sms { agent_id, template_id, query|lead_ids, schedule?, constraints? }
POST /api/nova/prep/bulk-email { agent_id, template_id, query|lead_ids, schedule?, constraints? }

# Job Control
POST /api/nova/job/{job_id}/approve - Owner/Manager only
POST /api/nova/job/{job_id}/pause
POST /api/nova/job/{job_id}/resume
```

Voice Communication APIs

```
# TwiML & Control
POST /api/voice/twiml/outbound?job_id=...&lead_id=... - Returns TwiML
POST /api/voice/twiml/inbound - Inbound call handling
POST /api/voice/events - Call status webhook

# Artifacts
GET /api/voice/recordings/{call_sid}.opus|.mp3 - Returns signed URL
GET /api/voice/transcripts/{call_sid}.json
```

SMS Communication APIs

```
# Sending
POST /api/sms/send { lead_id, agent_id, template_id?, body?, media_urls?[] }

# Webhooks
POST /api/sms/webhook/dlr - Delivery receipt processing
POST /api/sms/webhook/inbound - Inbound message handling

# Templates & Suppression
GET /api/sms/templates?agent_id=...
POST /api/sms/templates
PUT /api/sms/templates/{id}
DELETE /api/sms/templates/{id}
GET /api/suppression/sms
POST /api/suppression/sms/unsuppress { phone }
```

Email Communication APIs

```
# Sending
POST /api/email/send { lead_id, agent_id, template_id?, subject?, html?, reply_to?, attachments?[] }

# Templates & Assets
GET /api/email/templates?agent_id=...
POST /api/email/templates
PUT /api/email/templates/{id}
DELETE /api/email/templates/{id}
POST /api/email/assets - Upload asset
GET /api/email/assets/{id} - Get signed URL

# Webhooks
POST /api/email/webhook/events - SendGrid event processing
```

CRM APIs

```
# Lead Management
POST /api/crm/leads - Create lead
GET /api/crm/leads?stage=&owner=&assigned=&suppressed=&q=&tag=&tz=... - List/filter
GET /api/crm/leads/{lead_id} - Get single lead
PUT /api/crm/leads/{lead_id} - Update lead
DELETE /api/crm/leads/{lead_id} - Soft delete

# Batch Operations
POST /api/crm/leads/batch-assign { lead_ids[], owner_user_id?, preferred_agents? }
POST /api/crm/leads/batch-stage { lead_ids[], stage }
POST /api/crm/leads/batch-tags { lead_ids[], add?:[], remove?:[] }

# Tasks
POST /api/crm/leads/{lead_id}/tasks
PUT /api/crm/leads/{lead_id}/tasks/{task_id}
POST /api/crm/leads/{lead_id}/tasks/{task_id}/complete

# Timeline & Notes
GET /api/crm/leads/{lead_id}/timeline
POST /api/crm/leads/{lead_id}/notes { text }

# Imports & Exports
POST /api/crm/imports/csv - Upload CSV
POST /api/crm/imports/google-sheets { csv_url, mapping }
GET /api/crm/exports/csv?view_id=... - Generate export
```

Analytics & Billing APIs

```
# Usage & Analytics
GET /api/usage/summary?cycle=YYYY-MM&tz=... - Dashboard tiles
GET /api/usage/timeseries?
metric=voice_minutes|sms_segments|email_messages|retail_spend|first_audio_p95_ms&from=&to=&tz=... -
Time series data
GET /api/analytics/funnels?from=&to=&tz=... - Funnel data
GET /api/analytics/leaderboards?bucket=agent|template&from=&to= - Rankings

# Budget Management
GET /api/billing/budget - Current budget
PUT /api/billing/budget { budget_monthly_usd, soft_alert_pct, hard_stop_pct, auto_resume_on_cycle }

# Billing & Subscription
GET /api/billing/plan - Current plan details
GET /api/billing/subscription - Subscription status
POST /api/billing/checkout-session - Get Stripe checkout URL
POST /api/billing/portal-session - Get Stripe billing portal URL
GET /api/billing/caps - Current usage vs caps
GET /api/billing/invoices - List invoices
GET /api/billing/invoices/{id} - Invoice details

# NOVA Strategy
GET /api/nova/strategy/cards - Get strategy cards
POST /api/nova/strategy/cards/{id}/prepare - Create prepared job
POST /api/nova/strategy/cards/{id}/dismiss - Archive card

# Exports
GET /api/usage/export.csv?from=&to=&tz=...
GET /api/analytics/export.csv?from=&to=&tz=...
```

Settings & Onboarding APIs

```
# Organization Settings
GET /api/settings/org
PUT /api/settings/org
GET /api/settings/onboarding - Checklist status

# Integrations (BYOK)
POST /api/settings/integrations/twilio { account_sid, auth_token, messaging_service_sid?, from_number? }
POST /api/settings/integrations/sendgrid { api_key, sender_email? }
POST /api/settings/integrations/verify - Test connections

# Key Rotation (requires TOTP)
POST /api/settings/rotate/twilio { new_creds, totp_code }
POST /api/settings/rotate/sendgrid { new_key, totp_code }

# Test Bench
POST /api/settings/test/sms { to_phone }
POST /api/settings/test/email { to_email }
POST /api/settings/test/call { to_phone }

# Brand KB (NOVA)
GET /api/settings/nova/brand-kb
PUT /api/settings/nova/brand-kb
POST /api/settings/nova/brand-kb/preview - Sample in org tone

# Compliance & Suppression
GET /api/settings/compliance - A2P/SPF/DKIM/DMARC status
POST /api/settings/dnc/import - CSV upload
```

File & Artifact APIs

```
# Signed URLs
POST /api/files/sign { path, action, ttl_s }
GET /api/files/get?org=...&path=...&action=get&exp=...&sig=... - File download

# Upload Flow
POST /api/files/upload-url - Get upload URL
PUT [signed-url] - Direct upload
POST /api/files/finish - Finalize metadata

# Event Management
GET /api/events/dead-letter - List failed events
POST /api/events/replay { event_ids[] } - Retry processing
```

HQ (Admin) APIs

```
# User Management
GET /admin/users
POST /admin/users - CTO only

# Plans & Pricing
GET /admin/plans
PUT /admin/plans
GET /admin/price-snapshots
POST /admin/price-snapshots - Close/open cycle

# Beta Management
POST /admin/beta/orgs { name, owner_email, plan:"beta", caps{...} }
GET /admin/beta/orgs
POST /admin/beta/orgs/{id}/deactivate

# Support
GET /admin/support/tickets
POST /admin/support/tickets/{id}/reply

# Webhooks
POST /admin/stripe/webhook - Stripe event handling
```

Standard Response Codes

APIs use consistent HTTP status codes:

Code	Meaning	Common Scenarios
200	Success	Standard successful response
400	Bad Request	missing_org_header, invalid params
401	Unauthorized	invalid_token, step_up_required, invalid_signature
402	Payment Required	trial_cap, beta_cap, budget_hard_stop, plan_cap
403	Forbidden	organization_mismatch, insufficient_role, admin_realm_required
409	Conflict	quiet_hours, suppressed, country_blocked
410	Gone	trial_expired, invalid_or_expired_token
429	Too Many Requests	Rate limiting, includes retry_after header

Error responses include standardized format:

```
{
 "error": {
 "code": "organization_mismatch",
 "message": "The organization ID in the token does not match the X-Org-ID header",
 "request_id": "req_abc123"
 }
}
```

Maintenance & Operations Guide

This guide provides detailed information for ongoing maintenance, troubleshooting, and operational procedures for Astrix BOS. It's designed to