WiFi 音箱 - 量产版 AI 接入方案研发实施版 v1

用于研发、测试、运维与架构评审的实施版本

版本：1.0

日期：2026 年 3 月 21 日

目录

一、实施目标与范围

1.1 目标

目标	说明
统一身份	设备必须通过证书或设备密钥登录，不允许匿名运行。
统一入口	文本与高价值任务统一经 APIG / AI Gateway。
统一会话	所有任务都挂在 session_id、lease_id 下运行。
统一结算	平台 authoritative usage 是最终结算真相。

1.2 非目标

第一版不追求：多地域双活、所有场景都做临时 API Key、用户侧可见精细 token 明细。

1.3 关键约束

约束	说明
单设备单活	一台设备同一时刻只允许一个主会话。
JWT TTL	推荐 10-30 分钟，支持提前刷新。
租约低水位	文本剩余 20%-30%，语音剩余 30%-40% 时后台续租。
grace quota	只用于收完当前句 / 当前段 / 当前片，不可长期透支。

二、分层架构与状态机

2.1 分层架构

```
设备
  -> device login
  -> session open / renew / close / resume
  -> media play / stream

APIG / AI Gateway
  -> JWT 验证
  -> session affinity
  -> rate limit
  -> route / fallback

业务服务
  -> device registry
  -> session & lease
  -> quota ledger
```

```
  -> risk engine
  -> ops backend

火山云侧
  -> endpoint / RTC / usage / bill / callback
```

2.2 会话状态机

```
IDLE
  -> OPENING
  -> ACTIVE
  -> RENEWING
  -> DRAINING
  -> CLOSED

INTERRUPTED -> RESUMING -> ACTIVE
ACTIVE -> TERMINATED_BY_RISK
```

2.3 租约状态机

```
ISSUED
  -> CONSUMING
  -> LOW_WATERMARK
  -> RENEW_REQUESTED
  -> EXHAUSTING
  -> GRACE_CONSUMING
  -> CLOSED
```

三、接口与对象设计

3.1 设备登录

字段	类型	说明
device_id	string	设备唯一 ID
device_nonce	string	防重放随机串
device_signature	string	设备签名
firmware_version	string	设备固件版本

```
POST /api/v1/device/login

response:
{
  "access_token": "<JWT>",
  "expires_at": 1774053000,
  "refresh_before_seconds": 120
}
```

3.2 开启会话

字段	类型	说明
user_id	string	当前用户 ID
task_type	string	任务类型，如 TALK / STORY / CHAT
device_state	string	设备当前运行状态
audio_codec	string	语音场景音频编码

```
POST /api/v1/session/open

response:
{
  "session_id": "sess_xxx",
  "lease_id": "lease_xxx",
  "granted_units": 12000,
  "soft_threshold_units": 3600,
  "grace_units": 1200,
  "gateway_route": "story-primary",
  "rtc_token": "<RTC_TOKEN>"
}
```

在这个接口里必须完成：entitlement 校验、单设备单活校验、风险分级、预算预占。

3.3 续签租约

```
POST /api/v1/lease/renew

request:
{
  "session_id": "sess_xxx",
  "lease_id": "lease_xxx",
  "estimated_consumed_units": 8600,
  "current_segment": "story_part_03"
}

response:
{
  "next_lease_id": "lease_next",
  "granted_units": 10000,
  "soft_threshold_units": 3000,
  "grace_units": 1200
}
```

设备端要在后台静默触发，不允许当前流式链路同步阻塞等待。

3.4 恢复与关闭

```
POST /api/v1/session/resume
```

```
POST /api/v1/session/close
```

场景	要求
断网恢复	必须重新校验会话是否有效、租约是否有效、是否允许 grace。
主动关闭	可立即释放未使用预算，但最终结算要等平台 usage。
风控终止	服务端必须能强制把会话置为 TERMINATED_BY_RISK。

3.5 平台回调

```
POST /api/v1/vendor/usage/callback
POST /api/v1/vendor/task/callback
POST /api/v1/vendor/bill/callback
```

回调处理服务必须做到幂等写入，避免重复通知造成账本重复结算。

四、数据表与账本实现

4.1 核心表

表名	关键字段	说明
device_registry	device_id, product_key, risk_level	设备身份与风控元数据
user_subscription	user_id, plan_id, entitlement_set	套餐与权益
device_binding	user_id, device_id	设备绑定关系
session	session_id, user_id, device_id, status	会话主表
lease	lease_id, session_id, granted_units, grace_units	租约表
quota_ledger	ledger_id, session_id, delta_units, action	账本流水
vendor_usage_raw	task_id, input_tokens, output_tokens	平台原始用量
usage_settlement	session_id, settlement_delta	最终纠偏结果

4.2 账本规则

```
RESERVE -> CONSUME_ESTIMATE -> RELEASE
                      -> GRACE_CONSUME
vendor usage -> SETTLEMENT_DELTA
```

```
ops adjust -> MANUAL_ADJUST
```

实时账本用于运行期控制；最终账本以 settlement delta 为准。

4.3 settlement delta

场景	处理方式
实时估算 < 平台权威值	补记 SETTLEMENT_DELTA 扣减
实时估算 > 平台权威值	释放多占用额度
回调缺失或延迟	进入待结算状态，异步重试拉取 usage

五、部署与测试建议

5.1 部署拓扑

规模	建议部署
1000 台	2 台业务服务 + 1 套托管 MySQL + 1 套托管 Redis + 托管 APIG
10000 台	4 台 8 核 16G 业务服务 + 独立 worker + 高可用 MySQL + 托管 Redis + 托管 APIG

5.2 压测重点

压测点	说明
session/open	检查并发开启会话时的 entitlement、风控与预占性能。
lease/renew	检查低水位续权高峰下的热点和锁冲突。
usage callback	检查回调幂等写入和账本处理吞吐。
JWT verify	检查 APIG / 网关在高并发下的鉴权开销。

5.3 联调清单

清单	说明
弱网恢复	断网、重连、resume、grace 逻辑联调。
额度耗尽	低水位续租失败时是否自然收口。
风控终止	黑名单、速率限制、异常续签熔断是否生效。

账本对账	实时估算与平台权威 usage 是否能正确纠偏。

文档版本：1.0

最后更新：2026 年 3 月 21 日