PR #43688 完整报告

vllm-project/vllm

[Feature] SSL support for dp supervisor

合并时间：2026-05-30 03:28

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43688

执行摘要

- 一句话：为 DP supervisor 添加 SSL 支持
- 推荐动作：值得精读。该 PR 展示了在分布式推理框架中为管理服务添加 SSL 支持的正确姿势：区分外部用户流量和内部节点流量，合理跳过不必要的验证，同时保持代码简洁。对于需要 HTTPS 的生产部署场景是必要变更。

功能与动机

作为 PR #40841 的后续，之前 --data-parallel-multi-port-external-lb 不支持 HTTPS（会抛出错误），此 PR 移除了该限制，使 DP supervisor 可以与 SSL 证书一起使用，满足生产环境的安全需求。

实现拆解

1. 修改参数验证（validate_multi_port_external_lb_args）：将禁止 SSL 的检查替换为 ssl_keyfile 和 ssl_certfile 必须同时提供的校验；移除了对 ssl_ca_certs 的限制，允许用户传递 CA 证书路径。
2. 调整 URL 生成（_child_base_url）：根据是否提供了 ssl_keyfile 和 ssl_certfile，将 scheme 从 http 改为 https。
3. 健康探测适配 SSL（_probe_endpoint）：当启用 SSL 时，发起探测请求时设置 ssl=False，跳过证书验证——因为探测目标为节点内环回地址，安全模型假设节点内部可信。
4. Supervisor 启动时注入 SSL 配置：在 DPSupervisor.run 中将 ssl_keyfile、ssl_certfile 等参数传递给 uvicorn.Config，使其以 HTTPS 模式运行。
5. 配套测试：
   - 新增 _generate_self_signed_cert 辅助函数利用 openssl 生成临时自签名证书。
   - 添加 test_validate_multi_port_external_lb_args_allows_ssl 单元测试，验证 SSL 参数通过校验。
   - 增强 MockVLLMServer 支持 SSL 启动，并在生命周期集成测试中启用 SSL，测试 supervisor 与子服务器的 HTTPS 通信。

关键文件：

- vllm/entrypoints/openai/dp_supervisor.py（模块 入口层；类别 source；类型 core-logic；符号 validate_multi_port_external_lb_args, _child_base_url, _probe_endpoint,

DPSupervisor.run）：核心入口模块，关键参数验证、URL 生成和健康探测函数均在此修改，是 SSL 支持的主逻辑所在。

- tests/entrypoints/openai/test_dp_supervisor.py（模块 测试；类别 test；类型 test-coverage；符号 _generate_self_signed_cert, test_validate_multi_port_external_lb_args_allows_ssl, init, _poll_supervisor_health）：测试文件，新增自签名证书生成函数和 SSL 集成测试，确保功能正确。

关键符号：validate_multi_port_external_lb_args, _child_base_url, _probe_endpoint, DPSupervisor.run, _generate_self_signed_cert, test_validate_multi_port_external_lb_args_allows_ssl

关键源码片段

vllm/entrypoints/openai/dp_supervisor.py

核心入口模块，关键参数验证、URL 生成和健康探测函数均在此修改，是 SSL 支持的主逻辑所在。

```
def _child_base_url(args: argparse.Namespace, port: int) -> str:
    host = args.host or "127.0.0.1"
    if host == "0.0.0.0":
        host = "127.0.0.1"
    elif host == "::":
        host = "::1"
    # 根据 SSL 配置决定 scheme
    scheme = "https" if args.ssl_keyfile and args.ssl_certfile else "http"
    return f"{scheme}://{host}:{port}"


async def _probe_endpoint(session, args, port, path, ...):
    for iteration in range(conn_err_failure_threshold):
        try:
            probe_ssl = None
            if args.ssl_keyfile and args.ssl_certfile:
                # 内部探测走环回，跳过证书验证
                probe_ssl = False
            async with session.get(
                _child_base_url(args, port) + path, ssl=probe_ssl
            ) as response:
                return response.status == HTTPStatus.OK
        except (aiohttp.ClientError, asyncio.TimeoutError) as e:
            ...
```

评论区精华

1. 缺失参数处理（@tessapham）：建议显式处理 ssl_keyfile 或 ssl_certfile 缺失的情况。作者在 validate 中增加了两者必须同时提供的检查，已解决。

2. 新增测试覆盖（@jperezdealgaba）：建议添加 SSL 参数验证测试和实际 SSL 连接测试。作者添加了相应的单元测试和生命周期测试，已解决。
3. 证书链验证（@jperezdealgaba）：提议创建 ssl.SSLContext 验证证书链。作者回应解释：内部通信（supervisor 到子服务器）发生在单节点内，跳过验证是安全的；外部用户到 supervisor 和子服务器的 SSL 已由 uvicorn 完整处理。该设计接受，未修改。

- 缺失 ssl_keyfile 或 ssl_certfile 处理 (correctness): 作者在 validate 函数中添加了两者必须同时提供的检查。
- 添加 SSL 参数验证和连接测试 (testing): 作者添加了 test_validate 单元测试，并在生命周期测试中使用了 SSL。
- 证书链验证设计讨论 (design): 设计被接受，未修改代码。

风险与影响

- 风险：
  - 行为变更：之前 SSL 参数会导致异常，现在必须同时提供 ssl_keyfile 和 ssl_certfile，缺少任一将报错。依赖旧行为的部署需调整。
  - 测试依赖：_generate_self_signed_cert 依赖 openssl 命令行工具，若测试环境未安装则跳过，需在 CI 中确保可用。
  - 内部通信安全：健康探测跳过了证书验证（ssl=False），虽然设计假设节点内部可信，但若 supervisor 与子服务器不在同一节点（如跨主机部署），存在中间人攻击风险。当前仅支持单节点多进程部署，此风险可控。
  - 文件权限：证书文件路径需对 vLLM 进程可读，权限不足可能导致启动失败。
- 影响：
  - 用户：现在可以在使用 --data-parallel-multi-port-external-lb 时启用 HTTPS，增强了通信安全性。需要同时提供 SSL 证书和密钥文件，提升了部署准备成本。
  - 系统：supervisor 和子服务器均以 HTTPS 运行，健康探测调用相应调整。不影响不启用 SSL 的场景。
  - 团队：维护了从入口到子服务器的全链路加密，且保持了内部探测的简洁性。测试覆盖 SSL 场景，减少了回归风险。
  - 风险标记：行为变更：SSL 参数必须同时提供，测试依赖 openssl 命令，内部健康探测跳过证书验证

关联脉络

- PR #40841 Data parallel multi-port external load balancing supervisor: 此 PR 的先行工作，为 DP supervisor 添加了基础功能，本 PR 在此基础上添加 SSL 支持。