基于 5G-MEC 的半导体设备预测性维护系统

回文刚

海宁跨界国际半导体制造有限公司，浙江嘉兴，314400；

摘要：随着半导体制造工艺复杂度的提升，设备稳定性成为影响良率与生产效率的关键因素。传统的预防性或事后性维护模式已无法满足高精密设备对连续性与稳定性的需求。本文提出一种基于 5G 移动边缘计算（MEC）的半导体设备预测性维护系统，通过融合高频实时感知、低延迟通信、边缘智能推理与云协同建模，实现对关键设备状态的持续监测、故障趋势的提前识别和运维策略的自动闭环调控。文章系统构建了该系统的网络架构、边缘 AI 分析模型与业务集成路径，并结合晶圆机、刻蚀台等典型设备进行仿真验证与实际应用评估，结果表明本系统在降低非计划停机率、优化备件调度与提升维护效率方面具备显著优势。

关键词：5G；MEC；半导体设备；预测性维护；边缘智能；工业物联网

DOI：10.69979/3041-0673.25.10.046

1 系统总体架构与边缘部署模式

1.1 基于 5G 专网的通信网络设计

系统在构建之时，选用 5G 工业专网来充当通信方面的骨干部分，其能够将整个洁净生产车间都涵盖在内。在部署基站的时候，会去考虑采用高频段（就像 n79 这样的）小基站进行较为密集的覆盖策略，如此一来，便能保证毫米波信号在设备分布极为密集的区域不存在任何的盲点情况。借助独立核心网（也就是 5GC），可以让数据路径完完全全实现本地化，进而使得业务数据在厂区内部能够以高速且安全的方式完成传输。每一台核心设备都装配有 5G 工业终端模块，通过运用网络切片技术，从中划分出那种具备低时延同时又高可靠特性的通道，专门用于运维数据的传输工作。经过实际的测试可以发现，这样的网络架构是能够达成 10ms 以内的端到端传输延迟的，也能够为 MEC 推理模块给予其所需要的实时数据方面的有力保障。

1.2 MEC 边缘节点的软硬件构建模式

在半导体制造那种对洁净程度要求颇高、实时性要求很高且可用性也要求较高的生产环境当中，MEC（也就是多接入边缘计算）节点在进行部署的时候，务必要同时考虑到物理方面的安全性、对环境的适应能力以及算力上的冗余能力等诸多方面。本系统运用的是分布式边缘节点部署架构，会把 MEC 设备放置在厂房的通信弱电间里面，或者是工艺间隙的舱室当中，又或者是特定设备集成模块之内，通过这样的方式尽可能地去贴近数据的源头所在之处，从而有效减少数据回传所产生的延迟情况。

边缘计算节点选取的是工业级 GPU 边缘服务器，此服务器搭载了高性能的 ARM 架构处理器以及 Tensor 加速单元。它拥有抗电磁干扰的能力，在高湿度环境下可防腐蚀，并且具备高温散热的能力，这些特性使其能够满足长期运行所需要的物理耐受方面的需求。其操作系统是依据实时 Linux 内核裁剪而成的版本，该系统能够支持在硬件中断级别对数据进行优先响应，如此便可保证预测模型在处于高负载的环境之下，不会因为系统资源的争用而出现性能方面的波动情况。

在软件架构这块领域，MEC 节点负责运行 Docker 容器化服务，与此同时，它还会与 K8s 轻量边缘编排（就像 K3s 这样的）相互结合起来，以此来对模型微服务依照需求进行部署，并且能够实现策略模块的热更新操作。数据采集层呢，是通过工业 MQTT 协议总线来接入相关数据的，而且它还能够兼容像 SECS/GEM、Modbus、OPC UA 等等这些主流的工业协议，进而可以自动地去完成协议解码的工作，同时也能完成信号校准以及时间戳校正等一系列任务。此外，系统内部还嵌入了数据质量校验模块，这个模块能够对传感器出现的异常情况、数据发生的漂移现象以及突变信号等进行有效识别，从而避免因为模型输入出现失真的状况而对判断的准确性产生不良影响。

为了让服务能够一直保持运行状态，MEC 节点引入了主备热切换机制，搭建起一主一备或者一主多备这样带有冗余性质的架构。当主节点出现诸如处理能力达到瓶颈、出现热失控状况又或者发生服务故障等情况的时候，系统就会凭借健康检测以及心跳机制，自动去触发负载迁移这一操作，进而切换到备份节点，让备份节点能够毫无缝隙地接管相关任务。所有的模型镜像以及配

置文件，都能够借助边云协同管理平台来完成统一的更新与推送操作，以此保证各个节点在策略方面能够保持统一，在版本方面也能达成一致，切实构建起具备‘就近智能’以及‘高可用性’特点的边缘 AI 运维关键支撑点。

MEC 节点可不只负责预测模型的推理，它还拥有设备本地报警响应、维护策略执行这些功能，而且能进行结构化工单的生成，还可下发智能运维建议等，如此一来便达成了感知—计算—响应的闭环执行能力。其具备本地化以及自主决策的特性，这对于那种对延迟极为敏感、对通信安全有着特殊要求的晶圆制造场景来说，是相当适用的。

1.3 云边协同机制与数据生命周期管理

系统在整体架构方面达成了以边缘为主、云端为辅的协同格局。云平台重点担负起模型训练相关事宜，同时还负责设备全生命周期的数据归档工作，以及运维知识图谱的管理任务。每一台 MEC 节点在完成本地模型推理这一工作的同时，会把关键诊断摘要信息传送给云端，以此来对模型参数进行持续的优化，并对全局设备健康趋势予以更新。云平台依照业务规则，周期性地向下发派模型权重、风险等级策略以及操作指令，从而保证边缘节点的推理逻辑能够和整体策略相契合。此外，系统还引入了数据生命周期管理方面的策略，以此达成对设备数据从最初的采集，到后续的清洗、计算、归档，直至最终销毁这一全过程的治理，进而满足半导体行业针对数据安全与质量所设定的严格标准。

2 设备健康建模与智能维护推理机制

2.1 多维信号融合的状态感知模型构建

在半导体设备运行期间，传统的监测手段往往只是以单一的物理量作为基础来开展相关监测工作，这样的话就很难全方位地反映出设备实际所处的真实状态。为了能够提升对于设备运行状态感知的维度，同时也为了提高对故障前兆进行识别时的灵敏度，此次所研发的本系统特意构建起了一个多源传感器融合模型，依靠这个模型来对设备的运行状态展开全面且细致的感知。该模型会综合起来采集设备在运行过程当中所产生的诸如温度、电流、电压、振动、声音以及光谱等各种各样的信号，然后借助多模态融合技术针对这些采集到的信号来进行统一的建模处理。在这个过程当中，对于振动信号以及电参数而言，会通过小波变换的方式分解出频率与时间这两个维度的特征，如此一来便能够有效地将那些周期性的异常趋势给揭示出来；而针对声音信号与光谱波形呢，则是经过傅里叶分析来提取出主频以及能量分布的相关情况，通过这样的方式去识别设备内部是否存在异常摩擦或者是激光器老化之类的情况。

随后呢，把各类信号特征都送进卷积神经网络（CNN）当中，去提取图谱级的空间特征，找出局部变化模式和传感器之间存在的关联性。接着，凭借长短期记忆网络（LSTM）来建立起时序预测的路径，针对设备状态的演化趋势展开建模工作。到最后，系统会输出状态向量，按照多维评估指标的标准，把设备状态划分成三个不同的等级，也就是稳定运行（Normal）、轻微异常（Warning）、高风险预警（Critical）这三类。并且每一种状态都和不同的维护策略触发逻辑紧紧绑定在一起，将其作为边缘MEC节点决策引擎的输入内容，这样一来，便能够达成“看得全、判断准、响应快”的现场智能诊断能力了。

2.2 异常趋势识别与失效预测模型设计

设备故障通常并非一下子就出现的，而是在长时间不断运行以及部件慢慢老化的这个过程里逐步呈现出来的。本系统依据历史维护记录还有故障演化相关的数据来搭建了一套多任务学习模型，这套模型有着不错的识别精度，而且在预测方面也具备一定的前瞻性。该模型把多通道残差神经网络（ResNet）和注意力机制（Attention）融合到了一起，其中多通道残差神经网络（ResNet）能够让在特征提取过程里对深层细节的保持能力得以增强，而注意力机制（Attention）则着重关注那些高风险信号通道，以此来使预测准确率得到提升。模型的训练目标可不单单是对当前状态做分类判断，同时还要输出故障类型分类概率、故障触发窗口、关键部件位置以及 RUL（剩余寿命）曲线这些内容。

系统在边缘 MEC 节点完成部署之后，能够凭借实时采集到的信号，并与过往所建立的历史学习模型相互对照，以此来判定设备当下所处的风险程度，同时针对未来 48 至 96 小时之内有可能出现的异常状况展开概率方面的建模工作。一旦设备的剩余使用寿命（RUL）估算值低于该设备预先设定的标准维护阈值，又或者是出现故障突发可能性急剧上升这样的情况，那么系统便会即刻触发预警相关的机制，而且还会自动与部件级故障预测模板建立关联，进而实现精准定位。就拿离子注入设备来说吧，通过对其电源模块的波形畸变程度以及温升梯度变化加以分析，该模型能够提前足足 48 小时就识别出电源板存在老化的问题，其预测的准确率更是超过了 92%，这样一来就为设备的维护预留出了十分充裕的响应时间，从而有效地防止了计划外停机情况的发生，

也避免了物料出现浪费的现象。

2.3 智能维护策略匹配与工单生成机制

状态识别以及故障预测，其最终所追求的目标在于，促使设备运维行为发生转变，也就是从以往单纯的‘响应修复’模式逐步过渡到‘预判干预’模式。在达成这一目标的过程中，系统精心打造了一个智能维护策略匹配引擎，该引擎是以知识图谱作为基础构建起来的。它能够把设备型号、风险等级、使用年限、已经执行过的策略、当前所承担的任务批次等诸多方面的多维特征融合到一起，进而自动地生成那些可以切实执行的维护操作方面的建议。在策略库当中，涵盖了远程微调、在设备所在地自行校准、激活冷备部件、替换预备件、协同开展诊断等十多种不同的干预路径。并且，还会借助规则引擎来针对各项维护任务进行风险方面的评估以及资源方面的匹配操作，以此来保证每一项维护任务在实际开展的时候，不但具备切实可行的特性，而且还拥有经济划算的特点。

工单生成模块会在边缘 AI 推理模块的触发之下运作起来，并且结合相关的预测结果，自动去编制那种结构化的数字工单。这工单里面涵盖了诊断结论、推荐的操作步骤、对操作耗时的预估情况、所需物料的编号，还有责任部门以及执行窗口方面的建议等诸多内容。系统会借助 5G 专网把工单推送至一线工程师的移动终端，或者推送至 AR 眼镜的界面之上，以此达成设备、人以及云这三端的联动效果。当存在多个设备都有维护任务的情况时，系统就会运用图优化算法来对任务点加以聚类处理，同时结合工程师的工位以及设备所处的位置，动态地规划出一条巡检路线，且这条路线是最短路径的，进而提升工时的利用率以及运维执行的效率。

依据实际部署所反馈的情况来看，该工单机制在时间方面有着不错的表现，其平均能够提前生成的时间达到了故障前的 36 小时。并且，它还能够对超过 85%的现场维护任务予以有力支持，让这些任务可以达成‘提前预约’，接着‘分时执行’，最后实现‘任务闭环’这样完整的流程自动化操作。如此一来，便有效地减轻了人工调度方面所承受的压力，进而构建起了一个以边缘智能作为核心要点的预测性运维生态体系。

3 系统部署实践与应用成效评估

3.1 系统在先进制程产线中的落地情况

此系统已然在某家有着 12 英寸先进制程的晶圆厂之中完成了试点方面的部署事宜，涉及到诸如光刻机、等离子刻蚀机、沉积设备以及测试机等多达 27 台的核心设备。其部署周期仅仅耗费了 5 周时间，在此期间，不仅完成了 8 座 5G 基站的部署工作，还完成了 6 组 MEC 边缘节点的设置，并且开展了 3 轮模型训练与测试相关的工作，就生产线设备接入情况来看，覆盖率已然达到了 95%之高。该系统同原有的 SECS/GEM 接口进行对接的时候十分顺利，切实成功达成了 5G 数据通道和边缘智能模块两者之间的无缝集成这样的效果。在长达 30 天的运行测试阶段里，此系统累计发出的预测性维护工单数量达到了 21 张，所具备的准确率更是高达 90.4%，而且未曾出现过任何一起因为突发故障进而致使整线停机的情况。

3.2 非计划停机率与维护成本变化分析

把系统部署前后的设备维护数据拿来作对比，就会发现预测性维护系统发挥了不小的作用。它使得月平均非计划停机率从 3.2%一下子降到了 0.9%，之所以能有这样的成效，主要是因为它能够提前察觉到潜在的故障发展趋向，从而可以合理地去安排维护的时间段。以前那种因为着急抢修而造成的人力方面以及配件方面的浪费情况现在已经明显减少了很多，整体的维保成本也随之下降了 22.5%。与此同时呢，由于维护流程实现了数字化，再加上有远程指导提供支持，单次平均维修响应时间就从 4 小时大幅缩短到了 1.6 小时，这无疑极大地提升了产线的运营效率，也让设备的可用性变得更高了，如此一来，便充分验证了该系统在控制成本以及提升效率这两个方面所具备的综合价值。

4 结语

随着半导体设备对可靠性、智能性与预测性的不断提升要求，传统维护模式面临前所未有的变革需求。未来，伴随 AI 算力增强、工业大模型发展与人机融合交互的进一步深化，该系统将在推动半导体设备“智维管控”方向上持续发挥核心作用。

参考文献

[1]张尉.基于物联网的机电一体化设备实时监测与维护系统设计[J].家电维修,2025,(05):83-85.
[2]刘万平,马经黎,谢蓉.基于数字化盐田采卤泵的设备预测性维护系统[J].盐科学与化工,2025,54(04):46-49+54.
[3]李君,田丰,徐皓,等.基于客户精准画像的卷烟零售客户服务体系重构[J].商场现代化,2025,(07):17-19.