

广东证券期货业协会 2024 年 重点课题研究报告

课题名称：行情+AI 策略+交易的 FPGA 一体化平台研究

课题研究单位： 广发证券股份有限公司

课题负责人： 辛治运

课题组成员： 王蓁，陈春来，张展，张肖，王建岗

课题研究时间： 2024 年 9 月—2024 年 12 月

摘 要

中国证券做市业务正迎来从机制完善到规模扩张的双重发展机遇，商做市业务向上空间逐步打开。随着做市制度的不断深化，券商在期权、债券、权益类、衍生品、基金以及公募 REITs 等多个市场的做市能力预计将得到显著提升。目前，做市业务已成为头部券商的重要盈利点和加杠杆的重要途径，其带来的业务收入有效缓解了券商整体业绩的波动性。

广发证券是国内首批获得基金主做市商、期权主做市商业务资格的证券公司，做市业务连续多年处于第一梯队，并为 30 多家基金管理公司的 500 多只基金产品提供做市服务，为沪深交易所全部期权品种提供主做市服务。2024 年 11 月 22 日，广发证券收到中国证监会《关于核准广发证券股份有限公司上市证券做市交易业务资格的批复》，正式获得上市证券做市交易业务牌照。广发证券的做市业务，凭借精准且稳定的报价能力，致力于市场投资者提供充足的流动性支持，有效削减交易成本，缓解市场剧烈波动，从而满足投资者的多样化投资需求。

做市业务报价的精准度与实时性，高度依赖于业务系统的卓越运行效率以及先进的 AI 量化投研技术。为此，广发证券打造了集行情分析、AI 策略制定与交易执行于一体的 FPGA 平台，该平台能够实时处理海量市场数据，快速训练并推理 AI 模型参数及网络架构，智能捕捉市场细微波动，并以更高的精准度和运行效率，高效执行做市业务交易功能。在 AI 模型研究方面，本课题提出利用 LSTM 深度学习网络模型作为做市业务的基础策略，LSTM 网络模型通过门控机制能够有效解决传统 RNN 在处理长序列数据时遇到的问题，可以对长序列数据中的长期依赖关系进行建模，在金融市场中感知行情数据变化和生成更加准确的推理结果方面有极大优势。在 FPGA 一体化平台研究方面，本课题针对做市业务系统中适合 FPGA 加速的功能模块进行分析，总结通过对 Level-2 行情解码功能模块，因子计算与策略前置计算，AI 推理功能模块和交易功能模块的加速，可以使系统整体运行效率明显提升。当前，行情解码模块的硬件加速方案非常丰富，券商、机构和金融科技公司推出了多种硬件加速方案，本课题可以直接利用行业成熟方案作为选型。在因子计算与策略前置计算模块设计时，采用多标的独立子模块的设计方法，一方面可以防止标的之间参数的互相影响，另一方面可以提升系统的整体可维护性。AI 推理功能模块则采用当前较为主流的 Vitis-AI 加速库技术，该技术可以在 FPGA 硬件

平台上实现自定义深度学习网络模型，支持精度为双精度浮点数，模型参数和网络结构可调，在本课题中具有良好的适配性。此外，FPGA 加速卡上还集成了部分交易功能、系统监控功能和运维功能等，进一步提升了硬件系统的稳定性和可靠性，为做市业务的连续性要求提供技术保障。

本课题的主要成果及特色创新包括以下几个方面，第一，做市业务信息系统创新，引入 AI 技术至做市业务场景，显著增强了原本依赖传统策略的定价效能；第二，硬件 AI 加速库的应用和落地，通过对行业主流 AI 加速库进行技术调研，引入可在 FPGA 上实现的硬件加速库，实现 AI 算法的高效运行；第三，FPGA 技术创新，本课题将交易所行情解码，AI 推理和交易策略等功能集成到 FPGA 加速卡中，实现行情+AI 策略+交易的一体化方案，显著降低了系统整体延时；第四，针对高集成度硬件方案，我们提出了有效的监控与运维策略，确保了相关交易业务系统的稳定可靠运行。通过本课题的研究和系统研制，为广发证券做市业务系统的技术演进和发展提供了坚实基础。同时，为证券行业同类信息系统的运行效率提升和系统智能化转型提供了参考和借鉴。

关键词：做市业务；精准定价；AI；FPGA；一体化

目 录

摘 要	2
第一章 课题研究背景	1
第一节 证券公司做市业务简介	1
第二节 精准定价能力是证券公司做市业务的关键	2
第三节 本课题的研究意义	3
第二章 证券公司做市业务需求	4
第一节 证券公司做市业务体系	4
第二节 场内做市业务典型业务场景	5
第三节 当前做市业务系统的问题分析	7
第三章 课题引入的 AI 和 FPGA 技术分析	10
第一节 LSTM 网络模型	10
第二节 FPGA Vitis-AI 加速库	13
第四章 行情+AI 策略+交易的一体化平台设计	18
第一节 系统与功能分析	18
第二节 一体化平台的设计方案	20
第五章 系统上线运行及测试数据	26
第一节 系统上线部署方案	26
第二节 测试数据对比	28
第三节 生产环境整体效能分析	28
第六章 总结与展望	30
参考文献	32

第一章 课题研究背景

第一节 证券公司做市业务简介

证券市场交易制度^[1]是资本市场价格形成的重要机制，主要分为报价驱动制度（做市商制度^[2]）和指令驱动制度（竞价制度）。做市商制度是利用做市商自有资金、证券和投资者进行交易的市场交易制度。中国的做市商制度^[3]在 20 世纪 90 年代就已出现，最早由全国证券交易自动报价系统（STAQ）和 NET 系统分别在 1991 年和 1993 年尝试实行做市商制度，但由于市场环境等多种因素，于 1999 年 9 月停止交易。2000 年，中国人民银行发布《全国银行间债券市场债券交易管理办法》，首次提出“双边报价商”概念。2015 年，上海证券交易所推出了上证 50ETF 期权，引入做市商制度，以提高期权市场的流动性和稳定性，做市商在提供连续报价、促进交易等方面发挥了重要作用。

2022 年 1 月，沪深交易所分别发布债券做市指引，为债券做市提供制度基础，并于 2023 年 2 月正式启动沪深交易所债券市场做市业务。做市商制度的加速推进，为券商机构加速业务转型提供了制度基础。2022 年 5 月，证监会制定了《证券公司科创板股票做市业务交易业务试点规定》（以下简称《做市规定》），通过试点方式在科创板引入做市商机制，积累经验并稳步推进做市商制度。《做市规定》提出中国证监会核准取得上市证券交易业务资格的，可以试点从事科创板股票做市交易业务；经中国证监会核准取得上市证券做市交易业务资格的，可以在其他交易所按照规定开展证券做市交易业务。截至 2024 年 12 月，共有 20 家券商获批上市证券做市交易业务资格¹，18 家券商获批股票期权做市业务资格。《做市规定》落地以来，上市券商公告获取做市业务资格明显加快，当前，券商做市业务正处于机制完善与规模扩张的双重机遇期，众多券商纷纷加速布局，然而，日益激烈的竞争压力也随之而来。

2024 年 11 月 22 日，广发证券收到中国证监会《关于核准广发证券股份有限公司上市证券做市交易业务资格的批复》²，正式获得上市证券做市交易业务牌照。据统计，2024 年仅有 3 家证券公司获批上市证券做市交易业务牌照，仅有广发证券获得上市证券

¹ 证监会官网，网址为 www.csrc.gov.cn，最后访问日期：2024 年 12 月 12 日。

² 关于核准广发证券股份有限公司上市证券做市业务交易资格的批复，网址为 <http://www.csrc.gov.cn/csrc/c101857/c7520195/content.shtml>，最后访问日期：2024 年 11 月 21 日。

做市交易项下的全资格牌照，包括科创板股票做市、北交所股票做市、沪深交易所债券做市等业务资格。做市业务作为广发证券的重点发展领域，对业务团队的交易能力、风险管理能力及系统处理能力提出了更高要求，同时，也充分展现了公司在投资交易、风险管理及科技金融等方面的综合实力。

第二节 精准定价能力是证券公司做市业务的关键

2023 年 11 月 3 日，中国证监会就修订《证券公司风险控制指标计算标准规定》（以下简称《计算标准》）向社会公开征求意见。对新一轮《计算标准》的修订，证监会指出对证券公司开展做市、资产管理、参与公募 REITs（Real Estate Investment Trust，不动产投资信托基金）等业务的风险控制指标计算标准予以优化，进一步引导证券公司在投资端、融资端、交易端发力，充分发挥长期价值投资、服务实体经济融资、服务居民财富管理、活跃资本市场等作用^[4]。在做市业务方面，具体调整内容有三个方面：第一，降低做市业务风险资本占用。证券公司做市账户中持有的金融资产及衍生品的市场风险资本准备按各项计算标准的 90% 计算；中证 500 指数成分股风险资本准备比例由 30% 下调至 10%。第二，降低头部券商风险资本折算系数。分母端风险资本准备方面，连续三年 A 类 AA 级以上（含）由 0.5 调整为 0.4，连续三年 A 类由 0.7 调整为 0.63，降低大宗商品衍生品资金流出比例，大宗商品衍生品资金流出比例从 12% 降低至 8%。第三，降低中证 500 指数成分股及其他资产所需稳定资金。根据最新风控新规，中证 500 指数成分股所需净稳资金由一般上市股票标准的 50% 降至 30%；其他所有资产按到期日期重新规划所需稳定资金折算系数，由原来的 100% 改为 50%、75%、100% 三个区间。从风控指标计算新规可以看出监管放宽相关风控指标的政策意图，以支持做市业务发展。

券商做市业务盈利^[5]主要来源于交易中介收入（Facilitation Revenues）和库存收入（Inventory Revenues）。交易中介收入为做市买卖价差，主要来源于做市商定价能力和交易份额，与投资能力无关，是一种非方向性的稳定收入。库存账户收入是由于做市商无法立刻匹配抵消双边订单，采用自营库存股作为对手方交易而形成，其收益主要包括持有期间头寸价值变动损益和股利、利息等孳息收益，成本包括融资成本、对冲成本和资本成本。相较于传统方向性自营业务，非方向性的做市业务盈利稳定性更强，有助于提升自营收入质量。

当前，影响非方向性做市业务盈利的主要因素包括：市场波动性，高波动性的市场环境可能导致更大的买卖价差，但同时也增加了库存管理和风险管理的难度；交易成本，主要包括经纪佣金和清算费用在内的交易成本会直接影响到最终利润空间；对冲成本，为了减轻因持有证券而面临的风险，做市商可能会进行对冲操作，这部分成本也会影响净收益；融资成本，当需要借入资金来支持头寸时，利息支持也会成为一项重要的成本考虑因素。做市商的核心盈利点在于他们能够持续提供买卖报价，利用买卖之间的价差来获取利润，这个价差虽然通常很小，但由于做市商每日可能执行成千上万笔交易，因此积累起来可以产生显著的收入。做市商的收入与交易频率和规模直接相关，更高的交易量意味着更多的买卖价差收入机会，活跃的市场条件对做市商更加有利，而精准的定价能力使得做市商能够在最小化风险的同时最大化收益，如果定价过高或过低，可能导致库存风险增加或失去交易机会。

尽管做市业务可以带来稳定的收入流，但其同样面临诸多挑战，比如市场竞争激烈、监管要求严格以及技术快速变化等。此外，在极端市场条件下，如行情波动剧烈期间，证券公司做市业务可能会遇到难以预料的风险暴露情况，这对其资本实力、风险管理能力水平和精准定价能力均提出了更为严苛的要求。

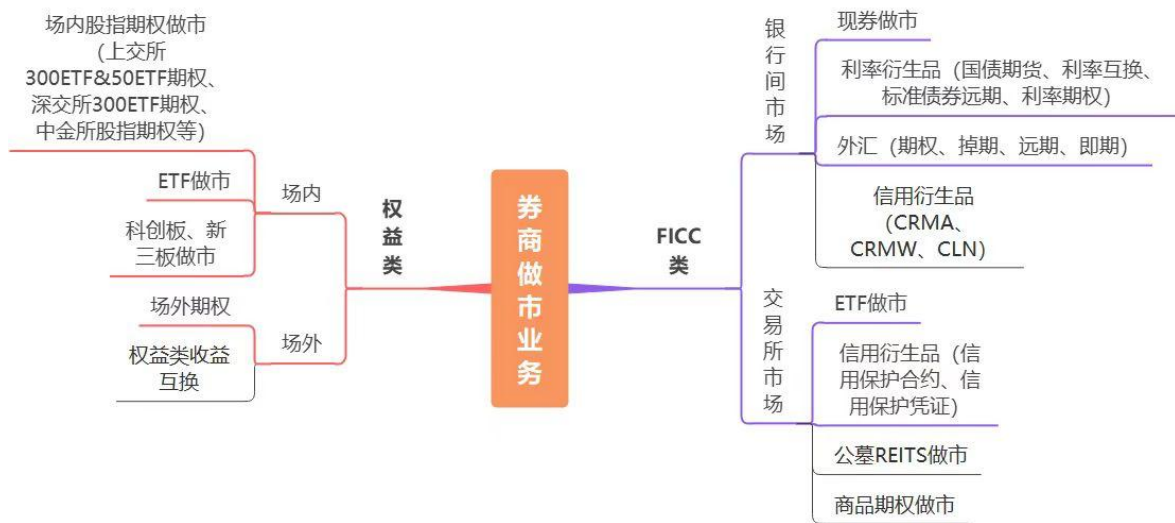
第三节 本课题的研究意义

本课题聚焦于证券公司的做市业务，针对该业务在精准定价和风险管控方面的强烈需求，对公司做市业务系统展开深入研究。通过引入当前金融行业热门的人工智能技术、FPGA 硬件开发技术等手段，旨在进一步提升做市业务系统的定价精度和风险管理能力，推动做市业务向更加智能化、精准化的服务方向发展，为证券公司做市业务的平稳运行提供坚实的技术支撑。

第二章 证券公司做市业务需求

第一节 证券公司做市业务体系

广义的做市是一种业务模式，在各类市场中参与交易的投资撮合行为，由 Market Maker 延伸为 Dealer 的角色，即交易商，盈利点从股票买卖的差价扩展到各类品种的差时、差场、差价、差额等。



根据上图可以对券商做市业务分为两大类^[6]，一类是场内做市商业务，例如权益类中的场内期权做市、科创板做市和 FICC（Fixed Income, Currencies & Commodities，即固定收益凭证、货币及商品期货）类的国债期货做市等，场内价格透明，这类做市业务券商不需要直接面对客户，以提供流动性为主要业务，同时成熟的市场中可借定价优势持续获利。但场内 ETF（Exchange Trade Fund，即交易型开放式指数基金）做市有一点不一样，是服务于具体公募的一些 ETF 产品提升流动性，这类场内做市之前，券商会与相应的公募客户直接对接，其收入来源也不仅限于买卖价差。第二类是场外交易商业务，例如权益类场外衍生品业务，FICC 的销售交易（票据市场、债券做市等），以及利率、外汇、信用类的衍生品业务等。这几类业务会直接面对机构客户，以客户需求为主导，一对一提供资本中介或产品创设服务，收入来源主要包括资本中介上的利息收入以及创

设产品上的消费收入。相较于场内做市，场外交易相对更能沉淀具体的机构客户资源。

综上所述，场内做市方面的能力注重定价策略、系统优势、流动性方面的资本实力，场外交易的能力注重客户资源，细化来说是看销售、产品创设及风控水平。

第二节 场内做市业务典型业务场景

一、沪深期权做市业务分析

A 股场内做市业务品种主要包括上交所 50ETF 期权和 300ETF 期权做市，深交所 300ETF 期权做市和中金所 300 股指期权做市。



图 2-2 A 股场内衍生品市场工具

针对上交所、深交所和中金所 4 类场内股指期权的做市，主要包括券商和期货子公司在其中提供流动性服务。期权做市业务能否平稳发展，一是交易所必须选择有资金实力的做市商，因为报价单会占用自有资金，需要机构有一定的资金实力，二是交易所前期须给予做市商较大优惠，至少要覆盖做市商的资金成本。

做市商需要有定价能力才能做得好，才能在该业务上盈利。其定价能力的作用就是承托整个市场，例如现货市场行情没有明显波动，但期权市场瞬间出现大幅跳涨或跳跌，这时就需要有定价能力的做市商提供买入或卖出服务，使现货市场行情与期权市场预期一致。

做市商义务做买卖动作，会在市场中被动成交，其盈利点只能在买卖价差的空间上

获得。理想情况是瞬间买入和卖出，可以赚到价格差，但实际情况一般是间隔一段时间才会卖掉，比如买了一个看涨期权，没有卖出且该期权下跌，就会产生相应的亏损，这时就需要 Delta 对冲掉，再过一会儿有机会卖出这个看涨期权后，再把 Delta 对冲回来，这就是赚得波动率^[7]（如下公式）的差。

$$\sigma = \sqrt{\sum \frac{1}{n} (p - \bar{p})^2}$$

即在 Delta 上已对冲干净的情况下，用低的波动率买入一个期权，再用高的波动率卖出这个期权，赚的是波动率差。因此做市商 Delta 上基本不赚钱，但能赚一个波动率的差，这就是做市商盈利的基本逻辑。如果波动率上也亏，还可以利用期货，赚隐含波动率（如下公式中的 σ ）和实际波动率的差。

$$C = S_0 N(d_1) - X e^{-rT} N(d_2)$$

$$d_1 = \frac{\ln(\frac{S_0}{X}) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

期权非常好的一点就是和标的是联动的，可以用现货去复制期权，如果复制出来的成本比期权成本低，就会产生盈利。整体来看，做市商盈利在于定价的精准度和系统上的速度，瞬间买卖价差上的效能，以及各类字母上的风控监测等。

二、ETF 做市业务分析

ETF 做市业务是做市商在证券市场上，按照规定的交易规则和市场情况，提供买卖对手盘，提升 ETF 成交量和活跃度，让投资者能够在希望的报价上买入卖出。券商对公募客户的服务，继券结代销、跟投、投研后又多一项 ETF 做市服务，合作面更广，而且

ETF 做市带来相对稳定的收入，公募一般会覆盖自营做市的资金成本，做市损益等。

做市商需要持有一定数量的 ETF 份额库存，根据当前市场需求、ETF 的净值变化和市场趋势等因素来调整库存水平。在实际报价中，做市商需要综合考虑多种因素，包括市场的买卖盘力量对比、交易成本和自身的风险承受能力等，通常，做市商的卖出报价会高于买入报价，这个价差就是做市商的利润来源之一，但价差不能过大，否则会失去市场竞争力。例如，在一个竞争激烈的 ETF 市场，做市商可能会将价差控制在较小的范围内，如 0.1%~0.3% 左右，以吸引更多的投资者进行交易。做市商在 ETF 做市业务中面临多种风险，其中包括市场风险，即 ETF 价格波动带来的风险，如果做市商持有大量 ETF 份额，而市场价格下跌，会导致其库存价值缩水。为了管理风险，做市商可以采用多种方法，如通过期货、期权等衍生品进行对冲，做市商可以卖出 ETF 期货合约，当 ETF 价格下跌时，期货合约的盈利可以部分抵消库存 ETF 份额的损失。

监管机构对 ETF 做市商的资质有严格的规定，做市商一般需要具有一定的资本实力、良好的信誉、专业的交易团队和完善的风险控制体系。在我国金融行业中，证券公司要成为 ETF 做市商，需要满足净资本等一系列财务指标要求，并且要有相关的交易和做市经验。此外监管机构会对做市商的报价进行监督，要求其报价要有合理的依据，如参考 ETF 的净值、市场的交易活跃度等因素，如果做市商出现异常报价，故意拉高或压低价格，会受到相应的监管处罚，这对做市商的做市业务系统也提出了严格的要求。

第三节 当前做市业务系统的问题分析

一、当前做市业务系统存在的问题

1. 软件做市业务系统运行效率不足，导致做市业务定价能力受限

软件做市业务系统运行效率不足可能是由于多种原因造成的。首先，可能是硬件基础设施的问题，包括服务器的配置较低，CPU 处理能力，系统内存大小或存储 I/O 速度不能满足做市业务高频率交易数据处理的要求。其次，软件系统本身的架构设计也是主要影响的因素，例如使用运行效率较低的 Python 语言作为策略运行脚本，或者代码没有经过足够的优化，存在一些复杂的嵌套循环、低效的算法或者过多的数据库查询操作等。做市业务的定价模型通常是比较复杂的数学模型，涉及大量的计算，如果系统不能快速处理这些计算，定价的及时性就会受到影响，导致定价能力下降。

此外，网络因素也不能忽视，如果系统所处的网络环境不稳定，带宽不足或网络延迟过高，会影响数据在不同组件的传输速度。例如，做市业务系统可能需要从外部数据源获取实时的市场行情数据来进行定价，网络延迟可能导致数据更新缓慢，不能及时更新证券的最新成交价、成交量等信息，基于过时数据计算出来的价格会偏离市场实际情况，使做市商在交易中面临更大的风险，或者错过一些有利的交易机会。

2. 采用性能优化的算法对系统整体穿透延时降低效果有限

做市业务系统本身往往具有高度的复杂性，涵盖了多个子系统、众多功能模块以及大量的数据交互链路。各个模块之间的耦合度可能较高，即便优化了部分算法，在实际运行中，由于模块间的相互依赖、调用顺序等因素，优化效果可能被其他未优化或难以优化的部分所抵消。例如，前端交易界面的操作请求需要经过多个中间层的处理、验证，再传递到后端定价与交易执行模块，而这些中间层的协调配合不佳或者存在冗余的处理流程，就会使得算法在局部的优化难以对系统整体穿透延时产生显著的影响。

做市业务系统还可能和外部多个数据源、合作机构的系统进行对接，外部系统的响应速度、接口稳定性等不可控因素也会影响到整体穿透延时，即使内部算法进行专门优化，但在等待外部系统响应的过程中依然会产生较大延时，进而削弱了整体的优化效果。

3. AI 技术引入金融行业后，传统做市策略面临着失效的风险

AI 技术具备强大的数据处理和分析能力，能够快速处理海量的金融市场数据，包括历史交易数据、实时行情数据、宏观经济数据等。它可以挖掘出数据中深层次的规律和关联性，而传统做市策略往往基于相对有限的历史经验和简单的统计分析方法来制定。例如，传统做市策略可能只是参考过去一段时间内某几种金融产品的价格波动范围来设定买卖价差，但 AI 技术通过机器学习算法能分析数千个变量之间的复杂关系，对价格走势作出更精准的预测，这使得传统做市策略所依据的简单价格区间判断等方法显得过于粗糙，容易出现偏差，面临失效风险。

二、做市业务系统中引入 FPGA 和 AI 技术的必要性

1. FPGA 硬件技术在金融行业已得到广泛应用

当前随着金融科技不断发展，FPGA 技术已逐步引入并在证券行业推广开来。FPGA 能够提供极低的延时，因为其可以直接在硬件层面进行交易策略的执行，无需像软件系

统那样经过操作系统的调度和多层软件中间件的处理。例如，从市场数据接收到交易信号发出，FPGA 可以在纳秒级别的时间内完成处理，相比传统软件交易系统，延时可能降低几个数量级。

FPGA 技术在金融行业的典型应用场景包括 FPGA 极速行情、极速柜台和实时风控等，这些产品在竞争激烈的现货交易、期货交易和衍生品交易中得到了广泛的应用，特别是在当前波动剧烈的行情下，其稳定和低延时的处理能力，使相关业务系统的平稳运行得到保障。

2. 业务系统复杂功能卸载，进一步提升了软件可维护性

当业务系统把复杂功能卸载到硬件后，软件系统中与之对应的那部分复杂代码可以被移除或者简化。原本充斥大量复杂数学公式、嵌套循环以及多分支判断的代码段，用于实现如复杂金融衍生品定价、风险计量等功能，现在无需在软件层面维护。这使得软件代码的整体结构变得更加清晰简洁，开发人员在阅读和理解代码时无需梳理复杂功能的逻辑，从而更易于定位代码中的问题，进行修改和扩展，有效降低了维护的难度。

3. AI 技术引入做市业务策略系统中，进一步增强做市业务的定价能力

金融市场中各因素之间的关系往往是非线性的，传统的定价模型大多基于线性假设，难以准确刻画复杂的市场情况。AI 技术中的神经网络等模型可以很好地拟合这些非线性关系。例如，在对金融衍生品进行定价时，其价格受到标的资产价格、波动率、利率、剩余期限等多个因素的交互影响，神经网络模型能够通过大量的训练数据学习到这些因素之间复杂的非线性作用机制，进而构建出更贴合实际的定价模型，相较于传统定价模型能更精准地计算出合理的价格范围，提升做市业务的定价准确性。

AI 模型具备动态学习与自我优化的能力，在做市业务实践中，随着市场数据的持续涌入，模型能实时更新参数、优化结构，以适应市场变化。比如，当市场出现突发的重大事件导致股价异常波动时，基于 AI 的定价模型能够迅速捕捉到这种变化，并通过对新数据的学习及时调整定价策略，使做市商的报价始终能紧跟市场动态，保持定价的合理性和竞争力。

第三章 课题引入的 AI 和 FPGA 技术分析

本课题针对当前券商做市业务系统存在的问题和短板进行研究，引入当前应用较为广泛的 AI 技术和 FPGA 技术，将原有传统软件和线性策略完成的做市业务进行技术升级，提升券商做市业务的精准定价能力和风险管理能力。以下对本课题引入的 AI 技术和 FPGA 技术进行详细分析。

第一节 LSTM 网络模型

一、LSTM 网络模型介绍

循环神经网络（RNN）在处理序列数据方面有很好的应用，如自然语言处理和时间序列预测。但是，RNN 存在一个严重的问题，即梯度消失或梯度爆炸。当处理长序列数据时，随着时间进一步的增加，在方向传播过程中，梯度可能会趋近于零（梯度消失）或者变得非常大（梯度爆炸），导致模型难以有效地学习长序列中的长期依赖关系，为了解决这个问题，长短期记忆网络（Long-Short Term Memory，简称 LSTM）^[8, 9]被提出。

LSTM 网络模型的基本结构如下图所示。

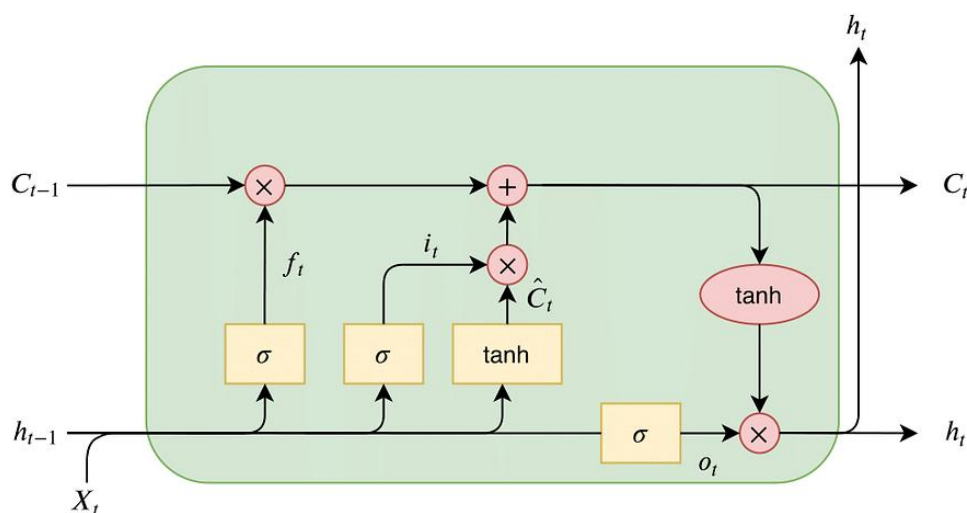


图 3-1 LSTM 网络模型基本结构

LSTM 网络模型核心点在于引入了记忆单元（Cell State）和一系列门控机制。该模

型通过引入一个长时间的记忆单元来存储重要的信息，并通过门控机制来控制信息的存储、遗忘和输出。这一设计使得 LSTM 能够在序列中“记住”关键信息，而非依赖单一的隐藏状态。LSTM 的结构包含三个关键的门：遗忘门、输入门和输出门，每个门通过特定的数学公式来控制信息流动。

1. **遗忘门**：控制记忆单元中哪些信息需要保留或遗忘。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

2. **输入门**：决定当前输入信息写入到记忆单元的程度。计算过程包括两个部分：首先使用 sigmoid 函数生成写入比例：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

然后生成一个新的候选值，通过 tanh 函数进行激活：

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

最后，将输入门和候选值相乘，得到更新后的输入值。

3. **更新记忆单元状态**：将遗忘门和输入门的结果结合，更新记忆单元状态。

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

4. **输出门**：输出门决定当前时间同步的输出信息。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

然后将记忆单元状态通过 tanh 激活函数处理，再于输出门的值相乘，得到最终的隐藏状态输出。

$$h_t = o_t \cdot \tanh(C_t)$$

这些门控机制使得 LSTM 可以“选择性地”记忆和遗忘信息，从而有效避免了梯度消失和梯度爆炸的问题，能够更好地捕捉序列中的长时间依赖关系。因此 LSTM 相较于普通 RNN 在处理长序列任务（如文本生成、语音识别、时间序列预测等）中表现更为出色。

二、LSTM 网络模型在金融行业的应用案例

1. 股票价格预测——AQR Capital Management

AQR Capital Management 是一家全球知名的量化投资管理公司。他们在投资策略中运用了包括 LSTM 在内的多种机器学习技术来预测股票价格走势。该公司收集了大量的股票市场数据，包括个股价格、市场指数、财务报表数据以及宏观经济数据等。这些数据经过清洗、归一化等预处理步骤后，被用于构建 LSTM 模型。该公司的 LSTM 模型结构较为复杂，包含多个隐藏层，以更好地捕捉股票价格时间序列中的长期和短期依赖关系。

通过对模型的训练和优化，AQR 能够利用该模型预测股票价格在未来一段时间（如未来一周或一个月）内的大致走势。这种预测帮助他们调整投资组合，例如，当模型预测某股票价格可能上涨时，适当增加该股票在投资组合中的权重；当预测价格可能下跌时，减少权重。这使得他们在投资决策过程中能够更有前瞻性，提高投资组合的绩效。

2. 信贷风险评估——蚂蚁金服

蚂蚁金服作为一家大型的金融科技公司，在信贷业务（如花呗、借呗）中面临着庞大的信贷风险评估任务。传统的信贷风险评估方法主要用于静态的信用评分模型，难以全面、动态地评估用户的信贷风险。

蚂蚁金服利用用户的消费记录、还款记录、芝麻信用分的历史变化等时间序列数据作为 LSTM 模型的输入。这些数据包含了用户的消费行为模式和还款能力的动态变化信息。LSTM 模型通过学习这些数据中的时间序列特征，能够预测用户未来的还款能力和违约概率。在信贷审批阶段，该模型能够更准确地筛选出高风险和低风险用户。对于低风险用户，可以提供更便捷的信贷服务和更优惠的利率；对于高风险用户，可以加强风险监控或者拒绝信贷申请。这大大提高了蚂蚁金服信贷业务的风险管理水平，降低了坏

账率。

3. 金融市场情绪分析——汤森路透

汤森路透是全球领先的金融信息和数据分析提供商。在金融市场中，新闻和社交媒体上的信息会对市场情绪产生重大影响，进而影响资产价格。汤森路透收集了来自各大新闻媒体、社交媒体平台的金融相关文本信息。通过自然语言处理技术，将文本转化为向量表示，然后利用 LSTM 模型对这些向量序列进行分析，以识别文本中所蕴含的情绪（如乐观、悲观、中性）的变化趋势。例如，在重大经济事件（如央行加息、企业并购）发布前后，观察市场情绪的动态变化。金融机构可以根据汤森路透提供的市场情绪分析报告来调整投资策略。例如，当市场情绪转向悲观时，投资者可能会减少风险资产的投资；而当市场情绪乐观时，可能会加大投资力度。这种基于市场情绪的分析为投资者提供了额外的决策参考，有助于更好地把握市场时机。

4. 证券期货行业智能分析与决策——国金证券

国金证券成功取得“证券期货行业基于长短期记忆模块的大语言模型多 Agent 系统”的专利。该系统运用 LSTM 和大语言模型构建多 Agent 系统，通过对历史数据的深度挖掘和理解，更好地预测未来的市场走向，优化投资策略，提升交易决策的精准度，多 Agent 系统还能模拟多种市场情景，通过协作机制达到信息快速处理和反馈。

通过分析多个案例，可见 LSTM 网络模型在金融行业应用广泛，其特性与本课题高度契合，适宜引入并作为 AI 做市策略的核心模型架构。

第二节 FPGA Vitis-AI 加速库

一、Vitis-AI 加速库介绍

AMD 公司的 Vitis-AI³是一款集成开发环境，可用于加速 AMD 自适应平台的人工智能推理。Vitis AI 提供了优化的知识产权（IP）、工具、库、模型以及多种示例设计和教材等资源，在整个开发过程中为用户提供帮助。它在设计时充分考虑了高效性和易用性，能够在 AMD 自适应片上系统（SoC）以及 Alveo 数据中心 FPGA 加速卡上充分释放人工智能加速的潜力。

³ AMD 公司 Vitis-AI 官网，网址为 <https://www.amd.com/en/products/software/vitis-ai.html>，最后访问日期为 2024-12-20

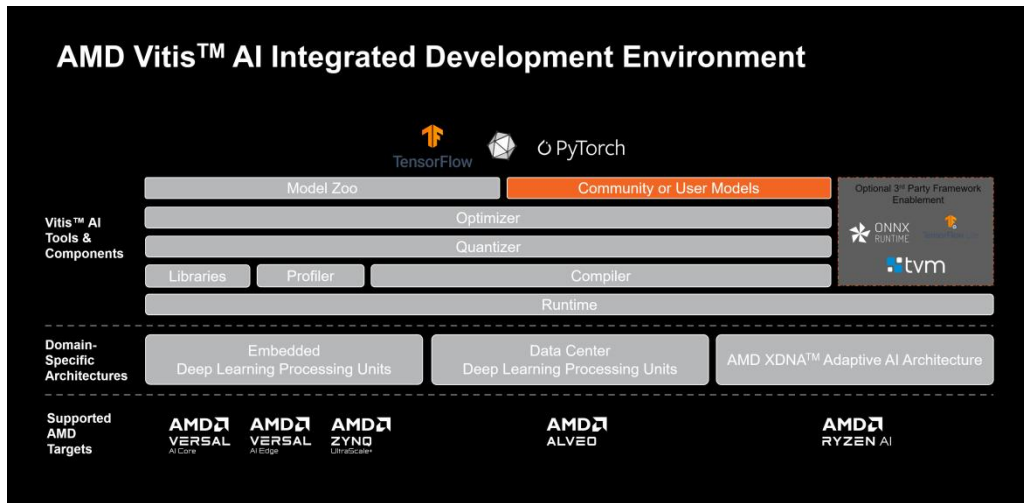


图 3-2 Vitis-AI 加速库

该加速库支持主流 AI 模块框架，能够进行多样化深度学习推理、卷积神经网络（CNN）和生成式大语言模型的最新模型；同时具备强大的量化工具，可实现最佳的模型精度和处理效率；此外，该模型编译流程简便，且拥有高级应用程序编译接口（API），能实现自定义模型的快速部署。除以上特点外，该加速库具备高效且可配置的神经网络处理器（NPU）核心，可满足边缘、数据中心以及云端在吞吐量、延迟和功耗方面的不同需求。

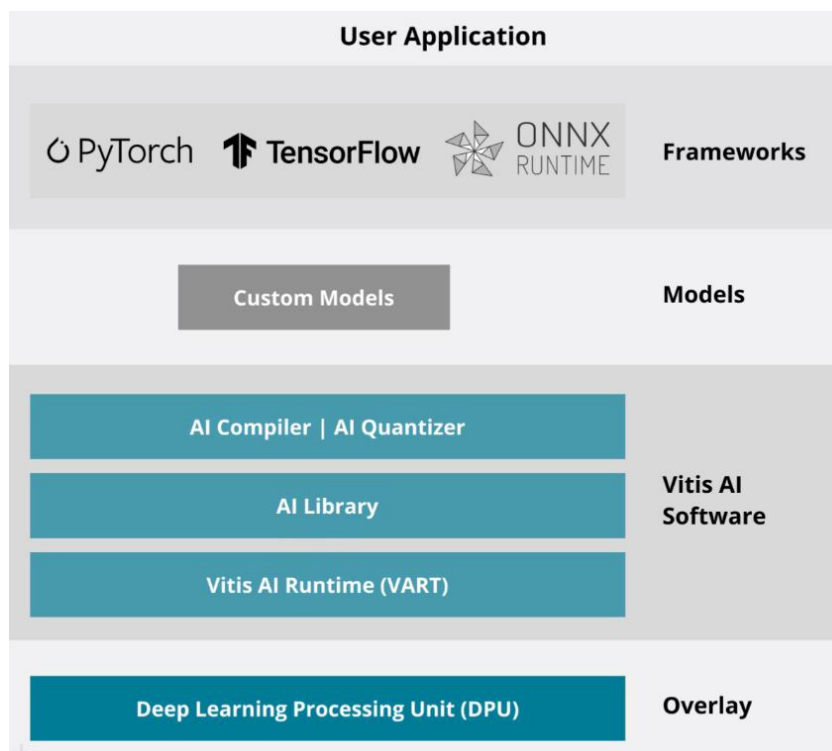


图 3-3 Vitis-AI 开发框架

Vitis-AI 开发框架^[10, 11]在应用层支持 Pytorch、TensorFlow 和 ONNX Runtime 主流开发框架，具有良好的兼容性。在标准开发框架支持下，用户可以自定义各种 AI 网络模型，并进行前期模型验证工作。Vitis-AI 可以将用户自定义模型通过 Vitis-AI Runtime（VART）转换成硬件可识别的程序，并在目标硬件（DPU）中运行和使用。在标准流程的支持下，用户仅需要关注所需研制的网络模型结构、参数和接口等，具体硬件加速方式由加速库统一处理，极大地降低了 FPGA 开发 AI 网络模型的技术难度。

二、Vitis-AI 加速库支持的模型

Vitis-AI 加速库提供了一系列 AI 模型库，其中包含了经过优化的深度学习模型，旨在加快在 FPGA 平台上部署深度学习模型的速度^[12]。这些模型涵盖了不同的应用领域，包括自动驾驶、医疗、金融、机器人技术、数据中心等，开发者可以用这些预训练的模型开始，快速开发和迭代 AI 模型和结构。

表 3-1 Vitis-AI 加速库支持的部分模型⁴

功能	应用	Vitis-AI 模型名称
Semantic Segmentation	Medical Segmentation	pt_3D-UNET_3.5
NLP	Question and Answering	pt_bert-base_3.5
Object Detection	Face Mask Detection	pt_face-mask-detection_3.5
Depth Estimation	Binocular depth estimation	pt_fadnet_0.65_3.5
Semantic Segmentation	ADAS 2D Segmentation	pt_HRNet_3.5
Image Classification	General	pt_inceptionv3_0.3_3.5
Pose Detection	Pose Estimation	pt_movenet_3.5
Image Classification	General	pt_OFA-depthwise-res50_3.5
Super Resolution	Super Resolution	pt_OFA-ran_3.5
Image Classification	General	pt_OFA-resnet50_0.88_3.5
Object Detection	General	pt_OFA-yolo_0.3_3.5
Semantic Segmentation	ADAS 3D Detection	pt_pointpillars_3.5
Industrial Vision / Robotics	Stereo Depth Estimation	pt_psmnet_0.68_3.5
Image Classification	General	pt_resnet50_0.3_3.5
Super Resolution	Super Resolution	pt_SESR-S_3.5
Super Resolution	Super Resolution	pt_xilinxSR_3.5
Object Detection	General	pt_yolov4csp_3.5
NLP	Question and Answering	tf_bert-base_3.5
Object Detection	General	tf_efficientdet-d2_3.5
Image Classification	General	tf_efficientnet-edgetpu-L_3.5
Industrial Vision / Robotics	Hierarchical Localization	tf_HFNet_3.5
Super Resolution	Super Resolution	tf_ran_0.98_3.5
Object Detection	Medical Detection	tf_RefineDet-Medical_0.5_3.5
Image Classification	General	tf_resnetv1-101_0.35_3.5

⁴ Vitis-AI 3.5 Model Zoo, 网址为 https://xilinx.github.io/Vitis-AI/3.5/html/docs/reference/ModelZoo_Github_web.htm

Object Detection	General	tf_ssdmobilenetv1_3.5
Semantic Segmentation	Medical Segmentation	tf2_2D-UNET_3.5

Vitis-AI 加速库中的模型经过专门的优化，在深度学习的推理阶段，这些模型能够利用硬件的特性实现高效计算。例如，对于卷积神经网络（CNN）中的卷积层操作，加速库可以通过对硬件底层的并行计算单元进行合理调度，与未优化的模型相比，在处理图像分类任务时，推理速度可能会提升数倍甚至数十倍，从而更快地对输入数据做出响应。

本课题在实际研究和系统研制中，将 FPGA 的 AI 功能模块采用 Vitis-AI 加速库的底层基础模型，用其搭建适合做市业务场景的 LSTM 模型结构，并导出至 FPGA 可运行的模块，与其他硬件功能模块进行对接，最终完成一体化平台的设计和上线验证工作。

第四章 行情+AI 策略+交易的一体化平台设计

第一节 系统与功能分析

一、券商做市业务系统整体链路分析

证券公司做市业务系统整体链路包含多个环节和组件，大致包括交易前链路，交易链路和交易后链路，业务系统上游对接交易所行情，下游对接交易所报盘网关，形成完整的业务功能链路。以广发证券上海期权/ETF 做市业务系统链路为例，给出做市业务系统的整体结构。

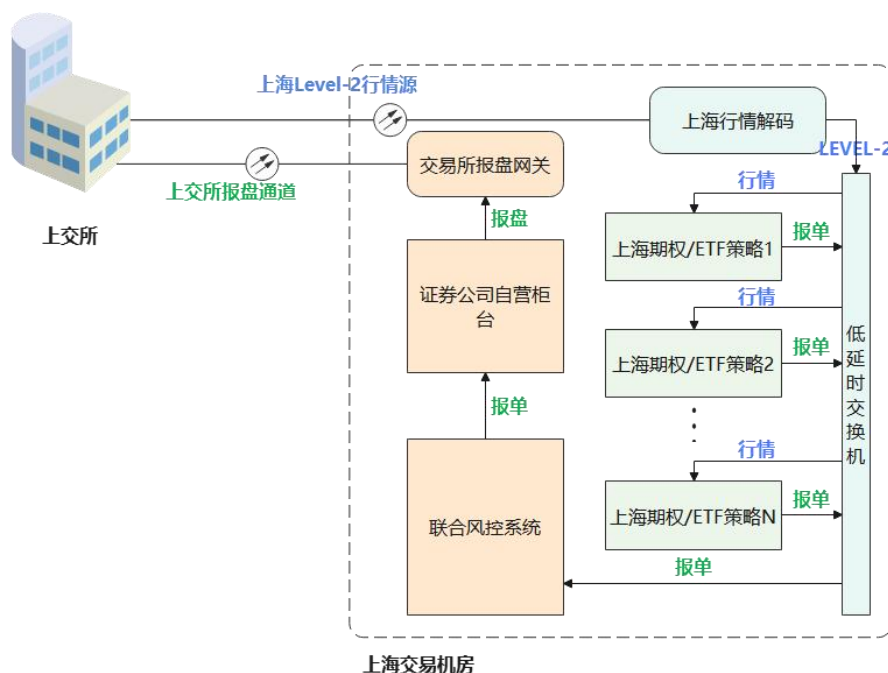


图 4-1 上海期权/ETF 做市系统交易链路

做市商从上海证券交易所获取 Level-2 原始行情并进行解码，得到标的 ETF 的价格、成交量、IOPV（Indicative Optimized Portfolio Value，基金份额参考净值）等数据，以及期权合约的 greeks 值等风险指标，将相关数据送入做市策略系统。同时，收集宏观经济数据、行业动态等信息，结合自身的风险偏好和库存情况，制定做市策略，确定买卖报价和交易数量等。

做市商通过做市系统将根据策略制定的双边买卖报价发送到上海证券交易所的交易系统中，报价包含期权合约的价格、数量和到期日等关键信息。上海证券交易所的交易系统根据“价格优先、时间优先”的原则，将做市商的报价与投资者的买卖申报进行撮合。当投资者的买入申报价格高于做市商的卖出报价，或者投资者的卖出申报价格低于做市商的买入报价时，交易系统自动撮合成交。

上海证券交易所的清算系统负责对交易进行清算，计算出做市商和投资者的资金和证券收付情况。做市商根据清算结果，与登记结算机构进行资金和证券的交收，确保交易的最终完成。

二、一体化方案中选取在 FPGA 上实现的功能

在以上交易链路的分析中，可以将交易链路大致拆解为几个部分。第一，交易所 Level-2 行情接入与解码；第二，解码完成后行情快照推送至策略服务器；第三，策略服务器对输入的行情进行相关策略计算，并由交易员设置相应的仓位、风控和规则等，策略服务器完成策略计算后，根据相关规则生产对应的报单规则；第四，策略服务器向交易所报单，其间需要经过联合风控系统检查和柜台报单操作。针对交易链路的分解，项目组分析可以在 FPGA 上实现的功能有以下几点。

1. 交易所 Level-2 行情接入与解码

交易所 Level-2 行情利用 FPGA 接入和解码已经在证券行业得到广泛的应用^[13, 14]，其技术可行性已得到充分论证。例如，中信建投证券公司对期交所五档组播行情利用 FPGA 进行解码加速，盛立科技公司推出自研的 FPGA 行情系统支持上交所逐笔合并数据通道，恒生电子也推出了 FPGA 极速行情系统以满足客户需求。这些案例的实现方式都是将交易所原始行情接入加速卡网络接口，并在内部通过硬件逻辑对 TCP 行情进行解码，对 FIX 协议进行解析，获得解码后的逐笔、快照和指数等行情，最后通过网络输出至业务系统。

2. 解码完成的行情转发功能

当前 FPGA 行情解码系统，大多数采用网络部署和本地部署的方案，网络部署方式需要解码完的行情通过低延时交换机进行转发，同时其中涉及以太网协议的转换和光纤传输，均会对行情转发延时产生较大影响。而采用本地部署方式同样存在该问题，本地行情推送需要经过 PCIe 接口，而 Level-2 行情的数据量很大，通过接口协议封装后，延

时损耗较大。基于以上分析，将解码完成的行情转发功能集成到 FPGA 卡内后，可以省去网络传输、协议封装的延时，提高数据传输效率，这样 FPGA 内部就会有专门的接口推送行情至其他功能模块，有效降低了系统复杂度。

3. 行情因子与策略计算

做市业务的策略可由行情触发的因子来计算，这种因子计算可以看成行情的深度加工，其中涉及数学运算、数据存储和逻辑设计等，当前 FPGA 技术已支持双精度浮点数的数学运算，支持丰富的金融加速库，实现行情因子与策略计算已不是复杂的工作，并且这些计算逻辑位于行情解码后，使用 FPGA 计算可最大限度发挥系统处理能力。

4. AI 模型计算

由上一章基于 FPGA 的 Vitis-AI 加速库研究可知，利用 FPGA 实现 AI 模型算法的技术已经成熟，通过将 AI 模型集成到加速卡上，可以极大地提高系统运行效率。在 AI 模型计算方面，系统需要考虑模型训练的问题，FPGA 适合在交易链路中进行 AI 模型推理，但是其并不适合模型的训练，在系统建设时，需要增加相应的 GPU 服务器完成模型的训练工作，并将训练结果及时推送给 FPGA 模型，并由其进行交易链路中的实时推理工作。

5. 定价结果报送与调参

定价结果报送与调参属于系统基础功能，具有较强的逻辑性，由硬件完成相应功能可以简化系统软件的相应功能负载，同时不会损耗较多的逻辑资源，对系统延时优化也有一定的作用。

通过以上功能的卸载，已经将做市业务系统中耗时较长的部分进行加速，可以使系统的整体穿透延时显著降低，而剩下的其余报盘和控制功能，由于其涉及标准协议交互和监控，并且其延时损坏并不明显，因此可以将系统相关功能继续由软件实现。

第二节 一体化平台的设计方案

一、系统整体结构

一体化平台的系统整体结构包含行情接入功能模块，策略服务器和 FPGA 加速卡等，为了保证系统运行的稳定可靠，在行情输入端增加主备切换逻辑，保障业务的连续性。系统整体结构如下图所示。

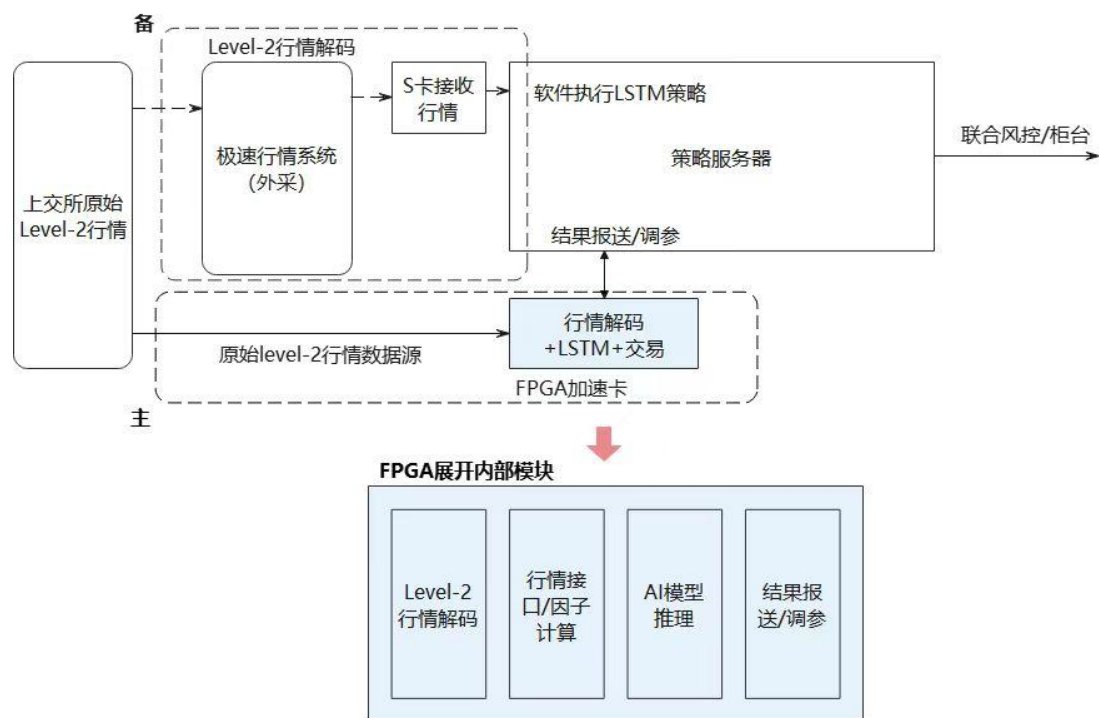


图 4-2 一体化方案系统整体结构

在行情输入端，上交所原始 Level-2 行情源通过网络转发至两个目标系统，分别为原有极速行情系统（外采）和 FPGA 加速卡网口。一体化平台的主业务链路是 FPGA 加速卡进行行情源的解码，行情因子提取与计算，AI 模型推理与结果报送调参。相关结果推送至策略服务器，策略服务器对接收到的 FPGA 计算结果进行实时读取，结合前端参数配置和持仓情况，生成相应的报单信息。

二、FPGA 详细设计

一体化方案中的核心功能在 FPGA 加速卡上实现，结合 AI 技术在硬件上的实现方式和行情、策略和交易等功能的设计需求，对 FPGA 的内部功能模块进行了详细的设计和研制工作。

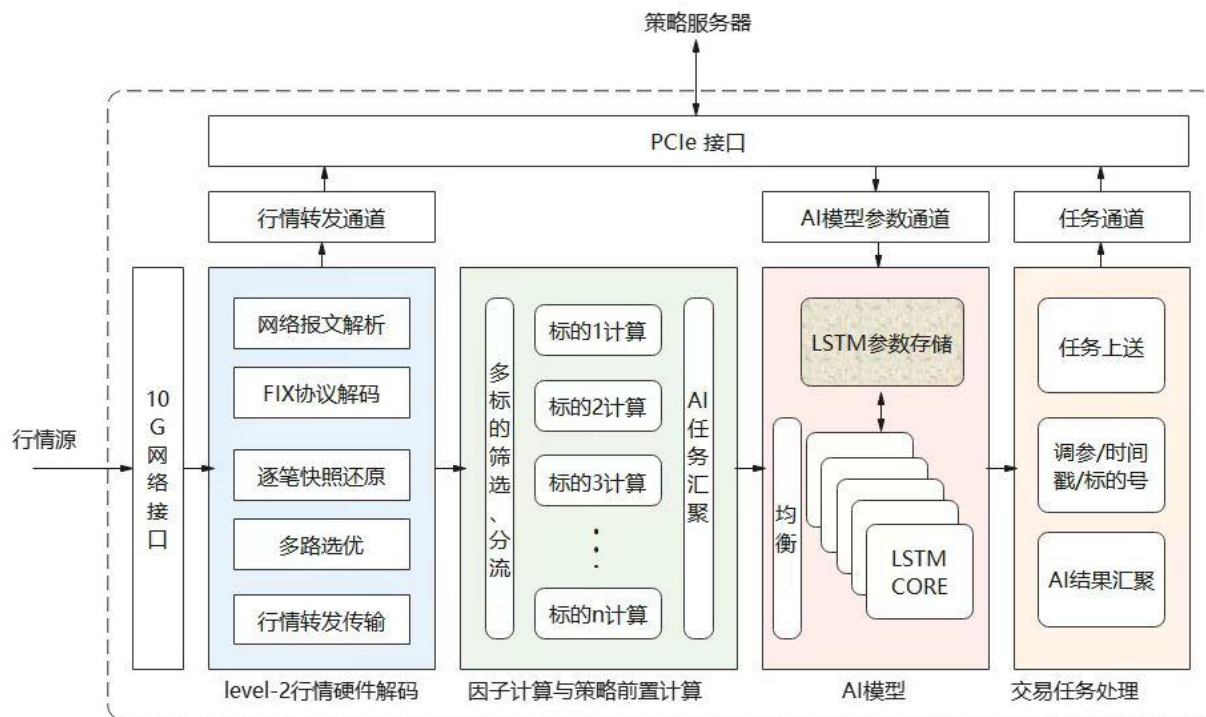


图 4-3 FPGA 详细设计

FPGA 内部的基础接口包括 10G 网络接口和 PCIe 传输接口，当前上交所和深交所的 Level-2 行情源流量均在 10G 以内，上交所采用 TCP 协议传输，深交所采用组播方式传输。由于广发证券期权做市主要业务集中在上交所标的，因此本课题主要考虑上交所 TCP 协议格式。上交所行情格式为 FIX（Financial Information eXchange，金融信息交换协议）消息格式^[15]，该消息格式是一种用于金融交易领域的电子通信协议，它定义了金融机构之间交换交易相关信息的标准格式。在 Level-2 行情硬件解码模块的设计中，广发证券直接采购供应商提供的标准行情解码模块，这种方式既能保证模块的性能指标，又可以将项目组主要精力用于业务逻辑的开发。Level-2 行情硬件解码模块主要支持的功能包括：网络报文解析，FIX 协议解码功能，逐笔快照还原功能，多路选优（支持 2 路行情源输入，模块内部选择延时低的行情源进行解码）以及行情转发传输等功能。除了行情网络接口外，PCIe 接口用于 FPGA 加速卡与策略服务器进行通信，策略服务器一方面可以直接获取加速卡内部解码后的行情数据作为系统监控使用，另一方面，策略服务器需要向 FPGA 加速卡策略参数和 AI 模型参数，用于加速卡内部 AI 网络模型的实时推理，除此之外，策略服务器还会通过 PCIe 通道接收 FPGA 加速卡上送的 AI 推理结果和策略运算结果等。PCIe 传输接口在 FPGA 加速卡内部有三个对应的功能模块，分

别为行情转发通道、AI 模型参数通道和任务通道。

因子计算与策略前置计算模块是用于对行情进行深度处理和生成 AI 推理任务的功能模块。根据课题论文对 LSTM 网络模型的分析，模型主要的推理方法是对序列数据的推理和分析，这就要求该模块内要可以对标的行情数据进行一定量的存储，而不同标的的买卖价格、成交量、IOPV 不尽相同。为了解决不同标的之间行情数据可能相互干扰的问题，本模块的设计采用独立标的占用独立子模块的方法，即每一个标的均有独立的子模块来完成行情数据存储，标的因子计算和标的 AI 推理任务的生成。该模块在入口处带有多标的筛选和分流模块，将不同标的的行情数据分流到对应子模块中，而在标的的计算子模块之后，AI 任务汇聚功能会将各个子模块的任务进行集中整合，利用 FIFO（First Input First Output，先入先出）机制推送至 AI 模型推理模块。

AI 模型推理模块主要是完成 LSTM 网络模型的推理，本模块在设计时有两个方面的考量。第一方面，LSTM 网络模型参数规模较大，为几十 KB 到上百 KB 之间，而一体化方案又涉及多个标的实时推理，每个标的均带有一组独立的 LSTM 网络模型参数，这就要求 AI 模型推理模块需要配置专门的存储空间来存储和读取这些参数。在本设计中，项目组使用 FPGA 内部的 URAM（Ultra RAM，大型块状缓存单元）来完成模型参数存储，URAM 提供 64bit 的数据存储位宽，不少于 4096 个存储地址，并且可以支持多 URAM 并联，最终本项目采用 URAM 存储 LSTM 模型参数，可支持最大 64 个标的，每个标的 100KB 的模型参数存储，并且模型参数读取延迟不超过 5 个时钟。另一方面，为了应对多个标的同时进行 LSTM 网络模型推理的任务，对 LSTM 模型进行多模块部署与任务均衡，考虑 LSTM 模型仅需要相应的任务输入和参数输入，无需与标的绑定就可以进行推理，这里仅需增加均衡子模块，并对多标的推理任务进行编排，就可以使 AI 模型推理模块顺利运行。

交易任务处理模块则是对 AI 模型推理后的结果进行汇聚和上报。各个标的的推理结果需要根据交易员设置的参数进行相应调参，然后将结果增加上标的号和时间戳，并按照相应的协议格式推送给策略服务器，最终上送的协议格式如下图所示。

标的号	时间戳	理论价
510050	9:30:00	3.124653
510500	9:30:00	0.798734
510050	9:30:03	3.156832
510500	9:30:03	0.799985
510050	9:30:06	3.224530
510500	9:30:06	0.765392

图 4-4 协议格式示例

通过 FPGA 内部各个模块的协同工作，可以由交易所 Level-2 行情驱动，在加速卡内部完成做市业务主要的因子计算和 AI 模型策略，然后将各个标的的计算结果推送至策略服务器，由策略服务器完成剩余的报盘工作。此外，FPGA 内部模块之间采用标准接口传输，不同模块之间添加了接口隔离功能，并且相关逻辑做了一定的冗余设计，使得加速卡运行的稳定性得到保障。FPGA 在实现主要业务功能的同时，还额外增加了专门的监控运维通道，该通道不对系统开放，研发人员可以在加速卡运行过程中，对加速卡出现的问题进行排查和测试，同时提供底层控制逻辑，帮助加速卡在故障模式下快速恢复，增加业务系统运行的稳定性和可靠性。

三、系统灾备与运维管理

系统在主业务链路运行时，还需考虑系统的灾备和运维管理，保证做市业务系统平时连续不中断，其中主要考虑的问题有以下几点。

1. 系统冗余

在系统中的关键模块，例如“极速行情系统（外采）”和“软件执行 LSTM 策略”模块，需要采用冗余设计。当主交易链路的 FPGA 加速卡出现异常时，系统可以立即启用备用系统，保证交易链路的保持持续运行。在策略服务器方面，也可以采用双机热备或多机集群的方式。双机热备模式下，两台服务器同时运行相同的策略软件，一台为主服务器，一台为备用。主服务器出现故障时，备用服务器能够在极短时间内接管业务，确保策略执行的连续性。

2. 数据备份

为了防止数据丢失，应建立数据备份机制。例如，对原始 Level-2 行情源数据采用

定期全量备份和实时增量备份相结合的方式；在交易时段。对新产生的业务数据进行增量备份，确保数据的完整性。

3. 系统监控

由于做市业务系统对可靠性和连续性有极高的要求，因此需要对整个系统进行全方位监控，包括硬件状态监控和软件性能监控。对于硬件方面，监控 FPGA 加速卡的温度、功耗、负载等参数，当出现异常时及时报警并采取应急措施。同时，对软件运行情况进行监控，如策略服务器上软件的执行效率，资源占用情况等。通过性能监控工具，实时获取策略执行的时间、内存占用等数据，及时发现性能瓶颈并进行优化。

4. 灾备恢复

本系统还需制定详细的灾备恢复计划，明确在不同场景（如硬件故障、软件崩溃、网络中断等）的恢复步骤和责任人员。例如，当 FPGA 加速卡出现硬件故障，运维人员应能够迅速定位故障位置，按照预先制定的流程更换加速卡，并从备份数据中恢复到故障前的运行状态。同时，还应定期进行灾备恢复演练，检验灾备恢复计划的有效性，演练可以模拟在不同的情况下，系统的恢复能力和业务连续性。

5. 系统升级与维护

定期对交易系统进行软件升级和硬件维护。对于硬件维护，需要定期对 FPGA 加速卡的功能、性能等情况进行分析，并针对发现的问题进行改正，发布和升级新的硬件版本。同时，对于软件部分也需进行定期升级维护，特别是当策略参数和配置项更新后，软件系统对应的功能也同步更新升级，确保系统始终与业务运行需求保持一致。

第五章 系统上线运行及测试数据

第一节 系统上线部署方案

本项目研制的一体化做市业务系统，旨在提供流动性支持和价格发现，确保市场的有效运行。系统主要包含行情网络的接入，FPGA 和服务器的连接，GPU 训练平台部署，以及 AI 模型的存储与转发等关键组件。此外，系统架构需具备高并发、低延迟的特点，同时保证系统的稳定性和安全性。

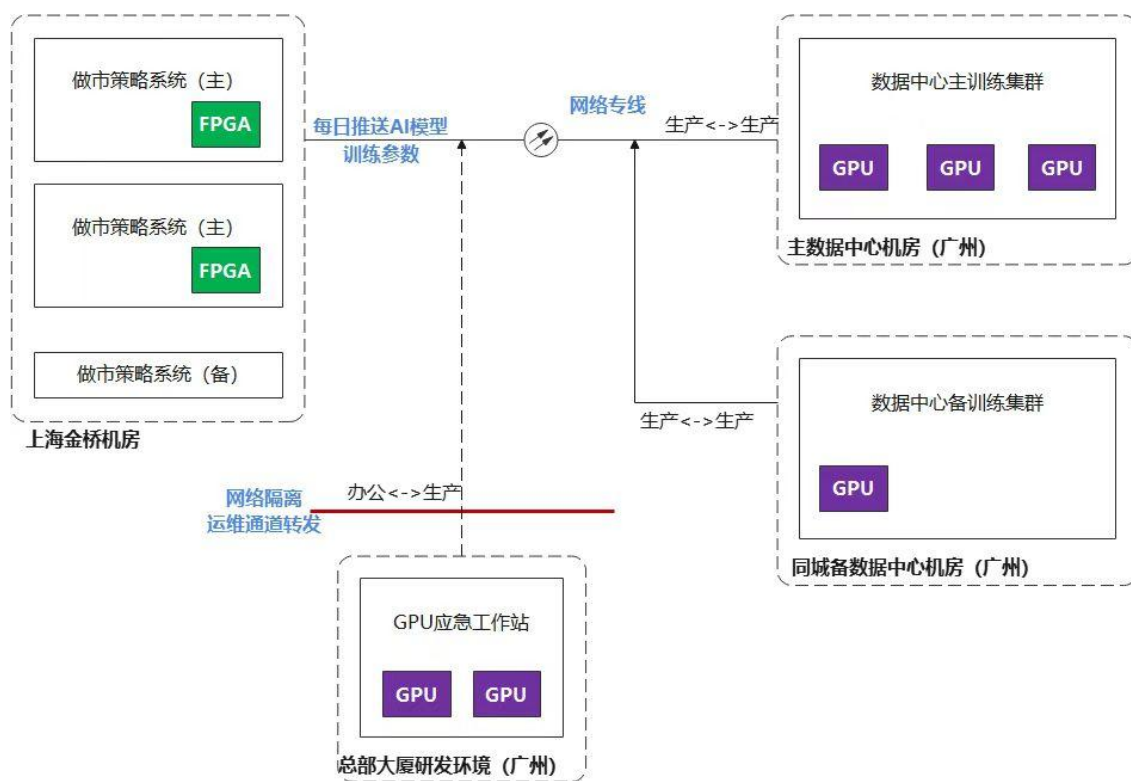


图 5-1 系统上线部署方案

下面分别对上线部署方案中涉及各个部门进行分析和论述。

1. 上海金桥机房

上海金桥机房为做市策略系统的主机房，包含带有 FPGA 加速卡的策略服务器，采用主备部署方式。在每个交易日开盘前，FPGA 会从 GPU 数据中心的训练模型中拿到当日策略的 AI 模型参数，将参数和配置数据下载至加速卡中完成初始化。交易时间开

始后，FPGA 获取交易所行情数据并进行实时策略计算，响应结果推送至策略服务器中进行汇集。策略服务器根据交易员预设参数和配置，生成对应的报单信息，向联合风控和柜台系统报送交易信息，并等待交易所的成交信息应答。

2. 广州的数据中心机房（主、备）

广州数据中心机房主要部署 GPU 集群训练平台，完成做市策略 AI 模型的每日训练。训练平台采用主备部署方案，在平时系统运行时，依托主数据中心的训练平台进行模型训练，其训练数据来源为公司每日落盘的行情数据。由于主数据中心的 GPU 集群为公司基础算力平台，上面运行着公司多个业务系统，为了更好地利用相关算力，采用动态调节算力的方案，当集群算力负荷较满时，可以将训练任务调度至夜间进行。除了主数据中心机房，该系统还将备用数据中心的 GPU 集群纳入其中，当主数据中心进行集中升级时，可将训练任务调整至备用数据中心继续完成训练任务。这样的系统结构也满足公司关键业务系统具有同城灾备能力的要求。

3. 总部大厦研发环境

考虑到 AI 模型结构和参数会经常变化，广发证券在建设该系统时，在总部大厦研发环境增加部署了一套应急 GPU 环境。这套环境的主要作用是防止主备 GPU 数据中心在训练方法更新，数据清洗，以及模型结构调整等无法及时变更的问题，可以在总部大厦环境由研发人员进行手动模型训练和调整，生产当日生产环境使用的 AI 模型参数。由于总部办公网络和生产交易网络严格隔离，数据的跨网段交互需要运维人员手工处理。当相关调整结束后，可以将相关信息更新至生成数据中心 GPU 集群，恢复业务系统的正常运行模式。

本课题系统的上线方案采用分布式架构部署，系统分布在上海和广州两地，通过网络专线进行连接，实现了异地的协同工作和数据交互。这种架构可以提高系统的可靠性和容错性，避免单点故障对整个系统造成影响。系统中无论是做市策略系统还是数据中心训练集群，都有主备设计，确保在任何一个组件出现问题时，都有备用方案可以快速切换，保障业务的连续性。此外，系统中大量使用 FPGA 和 GPU 等高性能计算设备，分别用于快速执行交易策略和高效训练 AI 模型，以适应金融市场对实时性和准确性的高要求。这种架构的设计和部署旨在确保做市业务的高效运行、模型训练的稳定性和可靠性，以及在各种情况下都能快速响应和恢复，从而为业务交易系统提供有力的技术支持和保障。

第二节 测试数据对比

以下给出本课题研制系统的相关测试数据对比。

表 5-1 测试数据对比

测试项	软件解决方案	FPGA 一体化平台方案
Level-2 行情解码功能	5~10us	300~500ns
行情网络传输延时	100~300ns	0
因子计算/策略准备	80us	2~2.5us
LSTM 模型推理延时	500us	40us
报单功能	3us	3us
整体穿透延时	590~600us	45~46us
系统延时抖动百分比	1%~3%	0.5%~1%

注：us 指微秒，ns 指纳秒，测试数据由广发证券总部大厦测试环境测试得到。

由以上测试数据可以看出，FPGA 一体化平台方案相较于软件解决方案，在 Level-2 行情解码功能，行情网络传输延时，因子计算与策略准备，LSTM 模型推理延时等方面，均有明显的延时降低。系统整体穿透延时较软件解决方案的 590 微秒至 600 微秒，缩短至 45 微秒至 46 微秒，整体性能提升 10 倍以上。并且考虑当前 FPGA 加速卡的资源占用率还未超过 50%，系统后续还有较大的提升空间。

第三节 生产环境整体效能分析

本课题系统的生产环境部署主要包括模型训练的 GPU 集群，交易机房的做市策略服务器和 FPGA 加速卡等。为了保障相关业务场景的顺利实施，计划在主数据中心调度 3~5 块 GPU 加速卡作为模型训练主集群，在同城备数据中心调度 1 块 GPU 加速卡作为备训练集群，另外在总部大厦研发环境配置 2 块 GPU 加速卡作为应急手段。在交易机房的配置中，配置 2 套做市策略主系统，并在每台服务器配置 1 块 FPGA 加速卡，此外，单独配置 1 套备用做市策略系统，不带 FPGA 加速卡，作为灾备和应急系统。通过以上

部署计划，可以看出系统在保障业务稳定运行时，没有增加过多的硬件设备部署，没有独立大规模训练集群，生产环境的整体效能表现较好。

第六章 总结与展望

通过本课题的研究，广发证券构建了一个将行情数据处理，AI 策略以及交易等功能集成于 FPGA 平台的一体化系统，并在实际生产中部署和应用，论证了本课题的实际应用价值。本课题主要的成果有以下几个方面：第一，高性能行情数据处理，通过优化 FPGA 的硬件架构，实现了对交易所行数据流的实时接收、解析和处理，能够在极短的时间内处理大量的市场行情信息，包括股票价格、成交量和买卖盘数据等，为后续的 AI 策略分析提供了数据基础；第二，精准的 AI 策略，利用深度学习算法在 FPGA 上的加速优势，对行情数据进行深度分析，挖掘潜在的市场趋势和交易机会，实现了对市场变化的快速响应和精准预测；第三，高效的交易执行，在一体化平台上集成了交易执行模块，采用了先进的交易订单管理技术，对交易指令进行优化和拆分，提高了交易的成功率和成交价格的合理性，同时，通过实时监控交易执行情况，及时反馈交易结果，为策略调整提供了可靠依据。

课题项目展望方面，相关业务系统可以针对系统软硬件，AI 技术和系统安全性方面进行持续优化。首先，芯片技术在不断发展进步，下一代 FPGA 芯片将具备更高的逻辑密度、更快的时钟频率和更低的功耗，广发证券将持续关注相关技术的发展动态，及时将新的硬件特性应用到业务系统中，进一步提升系统的运行效率。此外，广发证券正不断加大研发力度，深入探索新型 AI 算法（例如强化学习、大模型等）在金融交易领域的运用，持续优化 AI 策略模型，以增强其适应复杂市场环境和提高预测精度的能力，从而为投资者提供更加智能化、精准化的交易决策辅助。同时，在金融交易领域，业务系统的安全性和稳定性是核心要素，广发证券将不断强化平台的安全防护体系，引入更先进的技术手段，确保交易系统的持续稳定运行。

本课题研究的行情+AI 策略+交易的 FPGA 一体化平台在证券行业信息系统的数字化、智能化转型方面具有广阔的应用场景。例如，证券行业的专业投资机构需要高效、精准的交易工具来管理大规模的投资组合，用于应对复杂的市场环境，而采用一体化的硬件方案，使得市场智能感知、精准定价和快速响应成为可能。此外，对于零售投资者市场，跨境金融市场等，该方案也可以在快速感知市场变化、定制投资策略等方面发挥作用，提高投资决策的科学性和成功率，推动金融市场的普惠化发展。

在金融行业持续追求高质量发展的背景下，广发证券将继续坚守创新驱动、用户至

上的核心价值观，不断推进金融技术的革新与产品的迭代升级，拓宽市场应用范畴，强化团队凝聚力与合作机制，致力于将平台塑造为金融科技领域的标杆产品，为金融市场的繁荣发展和投资者的财富保值增值贡献更多力量。

参考文献

- [1] 皮六一. 沪深证券市场交易制度效率的综合评估[J]. 上海金融, 2013(01): 68-73.
- [2] 薛臻. 股票做市商制度发展及我国实践[J]. 金融纵横, 2023(07): 10-14.
- [3] 孙伟杰. 引入做市商对我国期货市场流动性的影响研究[D]. 湖南大学, 2023.
- [4] 陈雳, 梁燕子, 冯钦远. 证券公司服务实体经济的方向、机遇与挑战, 2018.
- [5] 王晓东. 证券公司盈利模式优化研究[D]. 江苏科技大学, 2018.
- [6] 林倩. 积极布局做市业务 券商提升综合服务能力[N]. 中国证券报, 2024-07-19.
- [7] 谭钦韬. 基于波动率的期权交易策略研究[D]. 西安理工大学, 2024.
- [8] A. F, N. A, O. F, et al. Bi-LSTM based model for efficient diagnosis of schizophrenia using time series EEG: 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON)[C], 2023.
- [9] K. S G, V. A, R. C, et al. CNN, LSTM, and Bi-LSTM based Self-Attention Model Classification for User Review Sentiment Analysis: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)[C], 2023.
- [10] A. U, M. W, N. W, et al. Convolutional neural network implementations using Vitis AI: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)[C], 2022.
- [11] A. P B, G. J P, S. T S. FPGA Based Acceleration of a Custom Deep Neural Network Model Inference: 2023 International Conference on Integration of Computational Intelligent System (ICICIS)[C], 2023.
- [12] 刘鑫. 基于 FPGA 的 AI 图像降噪方法研究[D]. 杭州电子科技大学, 2024.
- [13] 丁楠, 柴志雷, 高昊晖, 等. 低延迟抖动的证券行情数据解析系统设计与实现[J]. 计算机与数字工程, 2023,51(11): 2524-2530.
- [14] 冯一飞. 基于 FPGA 的超低延迟量化金融计算平台研究与实现[D]. 江南大学, 2022.
- [15] 王艳秋. 基于 FIX 协议的多版本间数据转换的设计与实现[J]. 信息与电脑(理论版), 2014(04): 111-112.