嵌入式芯片与系统设计竞赛

芯片设计赛道

赛题指南

正式赛题

目　录

01 赛题简介 2

02 芯片硬件框架 2

03 参赛要求 3

04 赛题说明 4

05 赛道阶段及提交内容 13

06 技术支持 13

07 其他 14

01 赛题简介

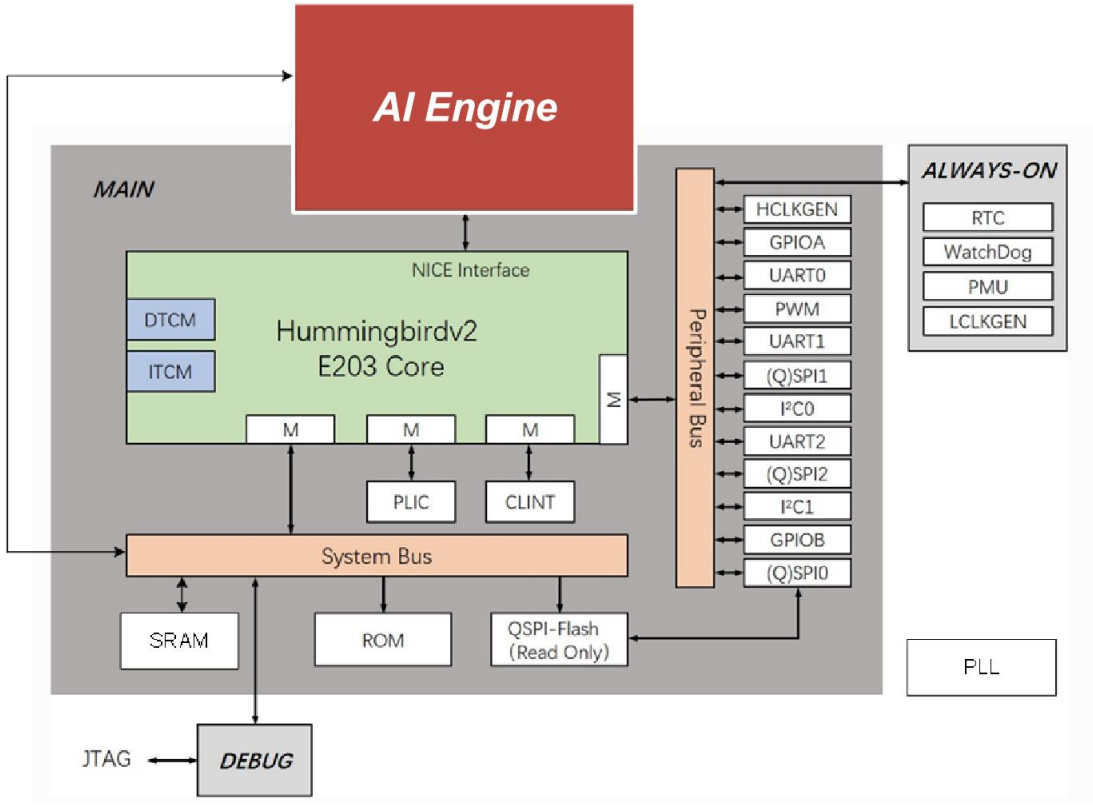
本赛题面向端侧离线语音指令识别，采用 KWT-1 开源预训练模型，在指定 E203 芯片硬件框架上设计计算加速器，完成以下任务：

1、完成本组规定的计算模块及控制逻辑设计。

2、完成加速器的 SoC 集成、驱动开发与功能验证。

3、完成逻辑综合及精度、吞吐量、功耗和面积评测。

02 芯片硬件框架



图：指定芯片硬件框架图

本赛题采用 组委会给定的硬件框架，包括：

- 芯来科技 E203 开源 RISC-V 处理器及配套软件开发环境，总线和外设采用组委会发布的平台工程。
- AI Engine 通过 System Bus 或 NICE Interface 接入，参赛队选择一种方式完成集成。
- 系统共享 SRAM 为 512 KB；加速器内部存储采用指定的 16 bits×128、16 bits×1024、16 bits×4096 单端口或双端口 SRAM 宏组合实现。
- SoC 提供 PLL，系统时钟为 150 MHz，该频率作为统一工作频率和逻辑综合时序目标。
- 常开域包括 RTC、WatchDog、PMU、LCLKGEN，低速时钟为 32.768 kHz。

系统采用分层参数装载。组委会提供参数加载仿真环境，通过平台已有访问接口向共享 SRAM 写入参数和输入数据；E203 执行驱动、调度及数据搬运，加速器完成本组规定的计算。全模型权重无需同时驻留共享 SRAM，不增设 DDR 或其他外部存储控制器。内部 SRAM 容量由参赛队根据面积与效率确定。

03 参赛要求

1）使用 Verilog 或 SystemVerilog 完成可综合 RTL。设计范围包括模块设计、仿真验证、SoC 集成和逻辑综合，不包括布局布线、物理验证及流片。

2）本科生组完成第 1—3 项任务，设计前馈线性加速器；其余任务自主选做。研究生组完成全部 7 项任务，设计十二层 Transformer 编码器加速器。必做计算不得由处理器软件或测试平台代算。

3）使用指定的公开预训练权重，不重新训练、微调或蒸馏，不改变模型层数、维度及基本运算结构。允许采用无梯度标定和数值格式转换。

4）数值精度和计算实现由参赛队自主设计，包括定点、浮点或混合精度、位宽、量化尺度、舍入与饱和规则，以及乘加结构、并行度、流水线、缓存组织和非线性近似方法。

5）提交与 RTL 一致的软件数值模型、参数转换脚本及数据格式说明。参赛队依据指定开源模型生成浮点参考结果，参赛队完成本设计格式的编码和解码，评测程序对解码结果进行比较。

6）加速器均须满足 1.5 mm×1.5 mm（2.25 mm^2）的面积上限。该数值作为逻辑综合阶段的面积预算，按本指南评分细则规定的统计范围核验。

04 赛题说明

面向端侧离线语音指令识别的 Transformer 计算引擎设计与实现

——赛题方向：开源预训练模型的数字前端实现

智能家居与交互终端通过本地语音识别接收“开启、关闭、停止”等短指令。本赛题采用 ARM 公开的 Keyword Transformer 最小型号 KWT-1，以英文语音指令为统一测试对象，设计适用于微控制

器系统的计算加速器。参赛队围绕语音时间片之间的动态相关性、矩阵数据复用和分层参数调度开展设计，不承担模型训练和音频预处理开发。

（一）模型与设计范围

1、采用 ARM-software/keyword-transformer 仓库提交 237e4097 的 models_data_v2_12_labels/kwt1 模型及 best_weights 权重。模型运行于推理模式，关闭训练、随机增强和 Dropout。

2、应用输入为 1 秒、16 kHz 单声道语音。参赛队利用指定代码离线完成音频规范化、98×40 MFCC 特征提取、输入投影、分类 token 拼接及位置编码，提供 99×64 的首层编码器输入。上述处理在计算机端离线完成，不列入参赛硬件任务。

3、模型包含 12 层编码器，特征维数为 64，注意力头数为 1，前馈网络维度为 64→256→64，采用 GELU 和 PostNorm 结构。硬件支持逐层参数更新，参赛队自行确定计算资源复用、并行度和调度方式。

4、输出类别为 yes、no、up、down、left、right、on、off、stop、go、静音及其他语音，类别编号采用权重目录中的 labels.txt。单次任务批量为 1，不设置变长输入和因果掩码要求。

5、本科生组的“前馈线性加速器”完成编码器前馈网络第一线性层的矩阵乘法与偏置加法，输入为 99×64，权重为 64×256，输出为 99×256；GELU、第二线性层和注意力均不属于本科必做内容。

6、研究生组的“十二层 Transformer 编码器加速器”完成全部编码器内的注意力、残差、LayerNorm 和前馈网络计算，输入与最终输出均为 99×64。分类头由组委会统一的软件评测程序执行，用于计算分类准确率，不计入硬件任务吞吐量。

7、原始参数和统一参考结果采用浮点表示。参赛队自行确定硬件数值精度及计算方法，GELU、Softmax 和 LayerNorm 允许使用查表、分段或迭代近似。数值设计须满足本章评分细则规定的精度门槛。

（二）数据集与测试数据

1、采用 Google Speech Commands V2，即 speech_commands_v0.02。参赛队自行下载公开数据集，按照本指南指定版本的模型源代码完成样本划分、标签映射及数据预处理。

2、识别类别包括十个目标词、静音及其他语音。其他语音类从十个目标词以外的词类中选取；静音类按照官方评估流程生成 1 秒全零波形。音频裁剪、补零及特征提取均采用指定版本模型源代码的处理流程。

3、参赛队从训练划分中自行选取样本进行无梯度数值标定，使用验证划分进行功能调试和精度验证。不得使用测试划分中的样本调整权重、量化参数或数值配置。

4、参赛队利用开源模型代码和预训练权重，自行生成硬件仿真所需的输入数据及浮点参考结果。本科生组提取各编码器层注意力、残差及第一处 LayerNorm 之后的前馈线性层输入、对应权重和偏置；

研究生组提取首层编码器输入及相关参考结果。上述处理在计算机端离线完成，不列入硬件必做内容。

5、组委会依据上述样本划分及类别规则选取 600 条样本作为统一评测集，评测结束后公布选取清单。组委会不另行制作和发布校准集、验证集或预处理数据包。最终评测采用参赛队冻结的提交工程及数值配置。

分项任务：本科生必做第 1—3 项，研究生必做全部 7 项

1、乘累加与数值处理模块

（1）实现长度为 64 的点积、偏置处理及本设计所需的数值转换。数据格式、乘法器结构、累加精度和运算顺序由参赛队确定。

（2）完成正常数据、边界值、复位和连续任务验证，说明舍入、溢出处理及误差来源。

2、前馈线性加速器

（1）实现 99×64 输入与 64×256 权重相乘并加偏置，输出 99×256 结果。完成参数装载、输入缓冲、计算控制及输出写回。

（2）使用指定预训练权重及自行提取的对应层输入验证，支持依次更换 12 层编码器的对应参数。GELU 和第二线性层列为本科选做内容。

3、前馈线性加速器 SoC 集成与综合

（1）将第 1—2 项设计接入 E203，完成驱动、数据交换、启动与完成状态管理，并通过 SoC 仿真验证。

（2）完成逻辑综合，提交 RTL、测试平台、综合网表、时序约束、面积和功耗估算报告，记录任务周期及吞吐量。本科生组完成本项及前两项即完成全部必做任务。

第 4—7 项为研究生必做，本科生自主选做

4、单头注意力模块

（1）完成 Q、K、V 投影、QK 转置乘法、缩放、逐行 Softmax、概率加权及输出投影。注意力分数矩阵维度为 99×99，单头特征维数为 64。

（2）验证各运算及注意力输出，设计矩阵分块、转置访问和缓存复用。运算近似和中间数据精度由参赛队确定。

5、单层编码器

（1）集成注意力模块、两处残差连接、两处 LayerNorm、64→256→64 前馈网络及 GELU，完成 PostNorm 编码器。

（2）按官方模型运算顺序验证单层结果，对照本设计软件数值模型检查中间输出，并说明近似计算带来的误差。

6、十二层编码器与参数调度

（1）顺序执行全部 12 层编码器，完成层间数据传递与逐层参数更新。E203 负责控制命令下发与数据搬运，编码器内数值计算由加速器执行。

（2）将最终编码器输出送入统一软件分类头，完成分类验证。针对权重装载、层间数据复用或注意力中间存储完成优化，并提交可复核的对照实验。

7、编码器加速器综合与性能分析

（1）完成加速器及 SoC 的逻辑综合、综合前后功能检查和 150 MHz 时序检查，报告关键路径、时序裕量及面积。

（2）提交运行周期、吞吐量、平均功耗、单任务能耗、SRAM 访问量和参数加载开销。全部指标使用同一设计配置、同一综合网表及规定测试条件。

（三）本科自主选做

本科生组完成第 1—3 项后，自主选择流水线、双缓冲、权重复用、低功耗控制、数值表示、完整前馈网络或后续任务开展创新。选做不要求按任务编号顺序扩展，注意力不作为本科达标或创新项满分的必要条件。将根据选做情况酌情提高分数。

（四）评分细则

两组分别评审和排名，满分均为 100 分。本节百分制评分适用于决赛。初赛按照本组规定的阶段任务进行验收。分值如下：

评分项目	本科生组	研究生组
功能与数值精度	40 分	35 分
SoC 集成、综合与复现	20 分	20 分
吞吐量	8 分	12 分
功耗与单任务能耗	8 分	12 分
加速器面积	4 分	6 分
自主创新	20 分	15 分
总分	**100 分**	**100 分**

1、功能与数值精度。

本科生组：点积及数值处理 10 分；前馈线性计算 20 分，其中矩阵与偏置计算 10 分、多层参数切换 5 分、缓冲与写回 5 分；数值精度 10 分。

研究生组：注意力模块 10 分、单层编码器 10 分、十二层执行与参数调度 5 分、分类精度 10 分。功能项依据测试用例通过率及对应任务完成情况评分。

本科数值门槛为归一化均方根误差不超过 3%。该误差按 12 层分别汇总全部评测样本计算，取最大值；各层误差等于输出误差均方根除以浮点参考结果均方根，分母下限为 10 的负 8 次方。误差为零得 10 分，每增加 0.3 个百分点扣 1 分，按比例计分，最低为零。

研究生分类门槛为 Top-1 准确率相对统一浮点参考下降不超过 2 个百分点。达到或超过参考准确率得 10 分，每下降 0.2 个百分点扣 1 分，按比例计分，最低为零。参赛结果与浮点参考使用同一评测集和同一分类头。

全部解码结果须为有限值。组委会在 600 条评测样本上运行参赛队的软件数值模型计算精度；本科覆盖各样本的 12 层前馈线性输入，研究生完成各样本的十二层编码器推理。RTL 对照使用本节规定的 12 个性能任务及独立组件向量。RTL 输出须与提交数值模型逐位一致；评审发现差异时扩大复核，差异未消除的设计视为功能验收不通过。

2、SoC 集成、综合与复现。

两组均为 20 分：驱动及 SoC 联调 5 分，综合与约束完整性 5 分，综合前后功能一致性 5 分，工程和报告可复现性 5 分。工程须满足 150 MHz 时序要求，不得使用隐藏时序违例的无效约束。

3、吞吐量、功耗与面积的测量。

本科一个任务为完成一次规定前馈线性层计算并写回结果；研究生一个任务为完成一次十二层编码器计算并写回结果。硬件计时不包含计算机端音频预处理、一次性文件格式转换和统一软件分类头执行。

性能测试在提交冻结后，由组委会从评测数据中统一抽取 12 个任务，同组采用相同输入。本科为覆盖 12 层参数的 12 次线性计算，每次均须单独满足 3% 的误差门槛；研究生为覆盖 12 个类别的 12 次十二层编码器计算。每个任务从输入和参数开始装载计时，至结果全部写回结束；任务开始前清空任务数据与参数缓存。参数装载、总线等待及 E203 调度时间全部计入，任务之间的测试平台复位时间不计入。

吞吐量等于完成任务数除以总运行时间，单位为任务/秒。平均功耗采用统一库、工艺角、负载、时钟与互连估计条件，根据同一计时窗口内的 SAIF 或 VCD 活动信息进行综合阶段估算；统计 E203、加速器、互连及 SRAM 的动态与静态功耗，模拟模块和片外设备不计入。活动信息须通过映射与覆盖检查，未标注电路不得按零活动处理，未建模 SRAM 不得按零功耗处理。

单任务能耗等于平均功耗乘以总运行时间，再除以完成任务数，单位为焦耳/任务。功耗项以单任务能耗计分，同时公布平均功耗，保证低功耗与处理效率共同接受评价。

面积指标为新增加速器、接口逻辑及内部 SRAM 宏面积之和，上限为 2.25 mm²。标准单元按指定库的综合映射结果统计，SRAM 宏按组委会提供的面积数据统计，不得遗漏或重复计入。复用的 E203、512 KB 共享 SRAM 及固定总线和外设不计入该预算，其面积单列报告。面积以逻辑综合结果核验，不进行物理边长或布局布线验收。

4、效率项计分。

吞吐量得分为该项满分乘以本队吞吐量与本组合格参赛设计最高吞吐量的比值。能耗得分为该项满分乘以本组合格参赛设计最低单任务能耗与本队单任务能耗的比值。面积得分为该项满分乘以本组合格参赛设计最小面积与本队面积的比值。各项得分保留两位小数。

只有完成本组全部必做任务、通过全部必测功能用例、数值门槛及 RTL 对照并满足时序和面积上限要求的设计进入效率与创新项评分。吞吐、功耗和面积须来自同一冻结配置，不得拼接不同配置的最优结果。无合格参赛设计时，本组效率与创新项均计零分。

5、自主创新。

本科生组 20 分：设计方案及创新说明 8 分、同条件对照实验 8 分、答辩与贡献说明 4 分。研究生组 15 分：设计方案及创新说

明 6 分、同条件对照实验 6 分、答辩与贡献说明 3 分。根据方案独立性、验证完整性及精度、吞吐、能耗或面积的可量化收益分。

05 赛道阶段及提交内容

（一）初赛

本科生组提交第 1—3 项任务的 RTL、驱动、验证工程和综合报告。研究生组在此基础上提交注意力模块和单层编码器。两组均提交数值格式说明、与 RTL 一致的软件数值模型及参数转换脚本。

（二）决赛

本科生组提交全部必做任务和自主选做成果；研究生组提交全部 7 项任务。交付可复现工程、SDC、综合网表、功能与时序报告、面积报告、SAIF/VCD 及功耗报告、吞吐与能耗记录、创新对照实验及答辩材料。模型、数据、脚本和工具配置均记录版本及校验值。初赛和决赛时间执行竞赛通知。

06 技术支持

（一）公开资源

1、模型代码、预训练权重、配置、标签及预处理实现：ARM-software/keyword-transformer，https://github.com/ARM-software/keyword-transformer。版本和模型目录按本指南执行。

2、原始语音数据：Google Speech Commands V2，获取入口见上述仓库 Setup 部分。公开数据为音频与标签，由参赛队利用开源代码与预训练权重自行生成。

3、模型论文：Keyword Transformer: A Self-Attention Model for Keyword Spotting，https://arxiv.org/abs/2104.00769。

07 其他

参赛队提交全部设计来源和引用说明，标明自主设计部分。不得针对评测样本保存预计算答案，不得将必做硬件计算转移到输入输出适配脚本。评测使用冻结提交版本；技术更正通过竞赛平台统一公告。