使用交互式人工智能（CAI）实现语音转录成本降低高达90%

**交互式人工智能（CAI）简介**

什么是交互式人工智能（AI）？

交互式人工智能（CAI）使用机器学习（ML）的子集深度学习（DL），通过机器实现语音识别、自然语言处理和文本到语音的自动化。CAI流程通常用三个关键的功能模块来描述：

1. 语音转文本（STT），也称为自动语音识别（ASR）

2 自然语言处理（NLP）

3 文本转语音（TTS）或语音合成

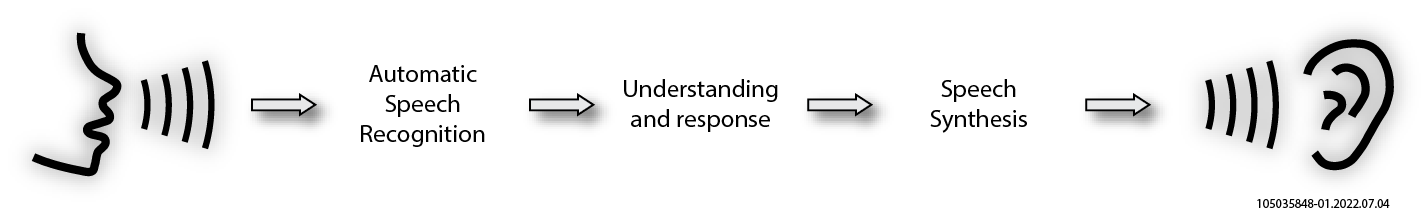


图1：***交互式AI构建模块***

本篇白皮书详细介绍了自动语音识别（ASR）的应用场景，以及Achronix如何在实现ASR解决方案的同时将相关成本降低高达90%。

细分市场和应用场景

仅在美国就有超过1.1亿个虚拟助手在发挥作用[1]，大多数人对使用CAI服务都很熟悉。主要示例包括移动设备上的语音助手，例如苹果的Siri或亚马逊的Alexa；笔记本电脑上的语音搜索助手，例如微软的Cortana；自动呼叫中心应答助理；以及支持语音功能的设备，例如智能音箱、电视和汽车等。

支持这些CAI服务的深度学习算法可以在本地电子设备上进行处理，或者聚集在云中进行远程大规模处理。支持数百万用户交互的大规模部署是一个巨大的计算处理挑战，超大规模的提供商已经通过开发专用的芯片和设备来处理这些服务。

现在，大多数小型企业都可以使用亚马逊、IBM、微软和谷歌等公司提供的云API，轻松地将语音接口添加到他们的产品中。然而，当这些工作负载的规模增加时（本白皮书后面将介绍一个具体的示例），使用这些云API的成本将会变得过高，迫使企业寻求其他解决方案。此外，许多企业运营对数据安全性有更高的要求，因此需要将解决方案必须保留在企业的数据安全范围内。

企业级CAI解决方案可用于以下应用场景：

* 自动呼叫中心
* 语音和视频通信平台
* 健康和医疗服务
* 金融和银行服务
* 零售和售货贩卖设备

详细介绍ASR处理过程

ASR是CAI流程的第一步，在这里语音被转录为文本。一旦文本可用，就可以使用自然语言处理（NLP）算法以多种方式对其进行处理。NLP包括关键内容识别、情感分析、索引、语境化内容和分析。在端到端的交互式AI算法中，语音合成用于生成自然的语音响应。

最先进的ASR算法是通过端到端的深度学习来实现。不同于卷积神经网络（CNN），递归神经网络（RNN）在语音识别中很常见。正如来自TechTarget [10]的David Petersson在《CNN与RNN：它们有何不同？》文章中提到：RNN更适合处理时间数据，与ASR应用非常适配。基于RNN的模型需要较高的计算能力和存储带宽来处理神经网络模型，并满足交互式系统所需的严格的延迟目标。当实时或自动响应太慢时，它们会显得迟缓和不自然。通常只有牺牲处理效率才能实现低延迟，这会增加成本，并且对于实际部署来说会变得过于庞大。

Achronix与采用现场可编程逻辑门阵列（FPGA）进行AI推理的专业技术公司[Myrtle.ai](http://myrtle.ai/)展开合作。[Myrtle.ai](http://myrtle.ai/)利用其MAU推理加速引擎在FPGA上实现基于RNN的高性能网络。该设计已集成到Achronix Speedster®7t AC7t1500 FPGA器件中，可以利用Speedster7t架构的关键架构优势（将在本白皮书后面进行探讨），大幅提高实时ASR神经网络的加速处理，从而与服务器级中央处理器（CPU）相比，可处理的实时数据流（RTS）的数量增加2500%。

数据加速器：如何实现资源的合理平衡分配

数据加速器可以卸载通常由主CPU执行的计算、网络和/或存储处理工作负载，从而可以显著减少服务器的占用空间。本白皮书介绍了用一台服务器和一个Achronix基于ASR的加速卡可取代多达25台服务器。这种架构大大降低了工作负载成本、功耗和延迟，同时提高了工作负载吞吐量。然而，只有在硬件得到有效使用并且部署具有成本效益的情况下，使用数据加速硬件来实现高性能和低延迟才有意义。

ASR模型对现代数据加速器来说是一种挑战，通常需要手动调整以实现比平台主要性能规格的个位数效率更高的性能。实时ASR工作负载需要高存储带宽以及高性能计算。这些大型神经网络所需的数据通常存储在加速卡上的DDR存储器中。将数据从外部存储器传输到计算平台是该工作负载中的性能瓶颈，特别是在进行实时部署的时候。

图形处理器（GPU）架构是基于数据并行模型，较小的批处理量（batch size）会导致GPU加速硬件的利用率较低，从而导致成本增加和效率降低。硬件加速解决方案数据表（以TOPS即每秒万亿次操作为单位进行衡量）中的性能数据并不能总是很好地表示实际性能，因为许多硬件加速器件由于与器件架构相关的瓶颈而未得到充分利用。这些数据以TOPS为单位，强调了加速器计算引擎的处理能力，但忽略了关键因素，例如外部存储器的批处理量、速度和规模，以及在外部存储器和加速器计算引擎之间传输数据的能力。对于ASR工作负载，关注存储带宽和在加速器内高效地传输数据为加速器性能和效率的实现提供了更强有力的指导。

加速器必须具有更大的外部存储规模和非常高的带宽。当今的高端加速器通常使用高性能的外部存储器，存储规模达8-16 GB，运行速度可高达4 Tbps。它还必须能够将这些数据传输到计算平台而不会影响性能。然而，无论如何去实现高速存储和计算引擎之间的数据通道，它几乎在所有情况下都是系统性能的瓶颈，特别是在实时ASR这样的低延迟应用中。

FPGA设计旨在存储和计算之间提供最佳的数据路由通道，从而为这些工作负载提供一个出色的加速平台。

Achronix解决方案与其他FPGA解决方案的对比

在机器学习（ML）加速领域中，已有FPGA架构宣称其推理速度可高达150 TOPS。然而在实际应用中，尤其是在那些对延迟敏感的应用（如ASR）中，由于无法在计算平台和外部存储器之间高效地传输数据，所以这些FPGA远不能达到其声称的最高推理速度。由于数据从外部存储器传输到FPGA器件中的计算引擎时出现了瓶颈，从而造成了这种性能上的损失。Achronix Speedster7t架构在计算引擎、高速存储接口和数据传输之间取得了良好的平衡，使Speedster7t FPGA器件能够为实时、低延迟的ASR工作负载提供高性能，可实现最高TOPS速率的64%等级。

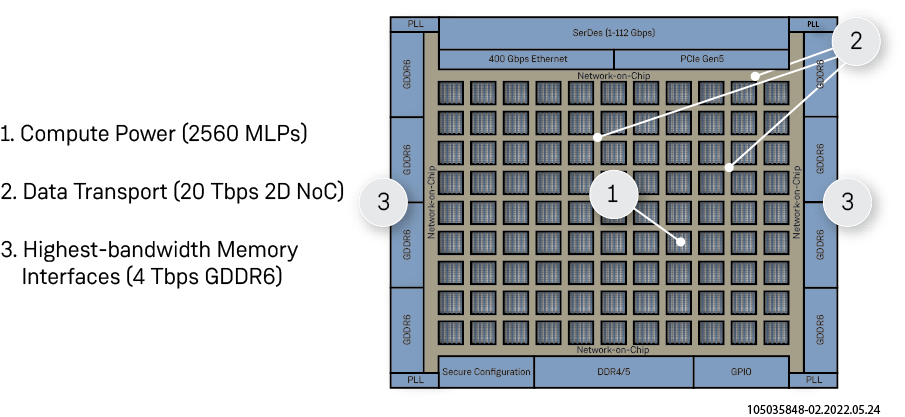


图2：*Speedster7t器件的计算、存储和数据传输能力*

Speedster7t架构如何实现更高的计算效率

在Speedster7t上搭载的机器学习处理器（MLP）是一种优化的矩阵/向量乘法模块，能够在单个时钟周期内进行32次乘法和1次累加，是计算引擎架构的基础。AC7t1500器件中的Block RAM（BRAM）与2560个MLP实例都处于同一位置，这意味着更低的延迟和更高的吞吐量。

借助于这些关键的架构单元，Myrtle.ai的MAU低延迟、高吞吐量的ML推理引擎已集成到Speedster7t FPGA器件中。

在构建最佳的ASR解决方案时，集成了之前提到的来自Myrtle.ai的MAU推理引擎，使用了2560个MLP中的2000个。由于MLP是一个硬模块，它可以运行在比FPGA逻辑阵列本身更高的时钟速率上。

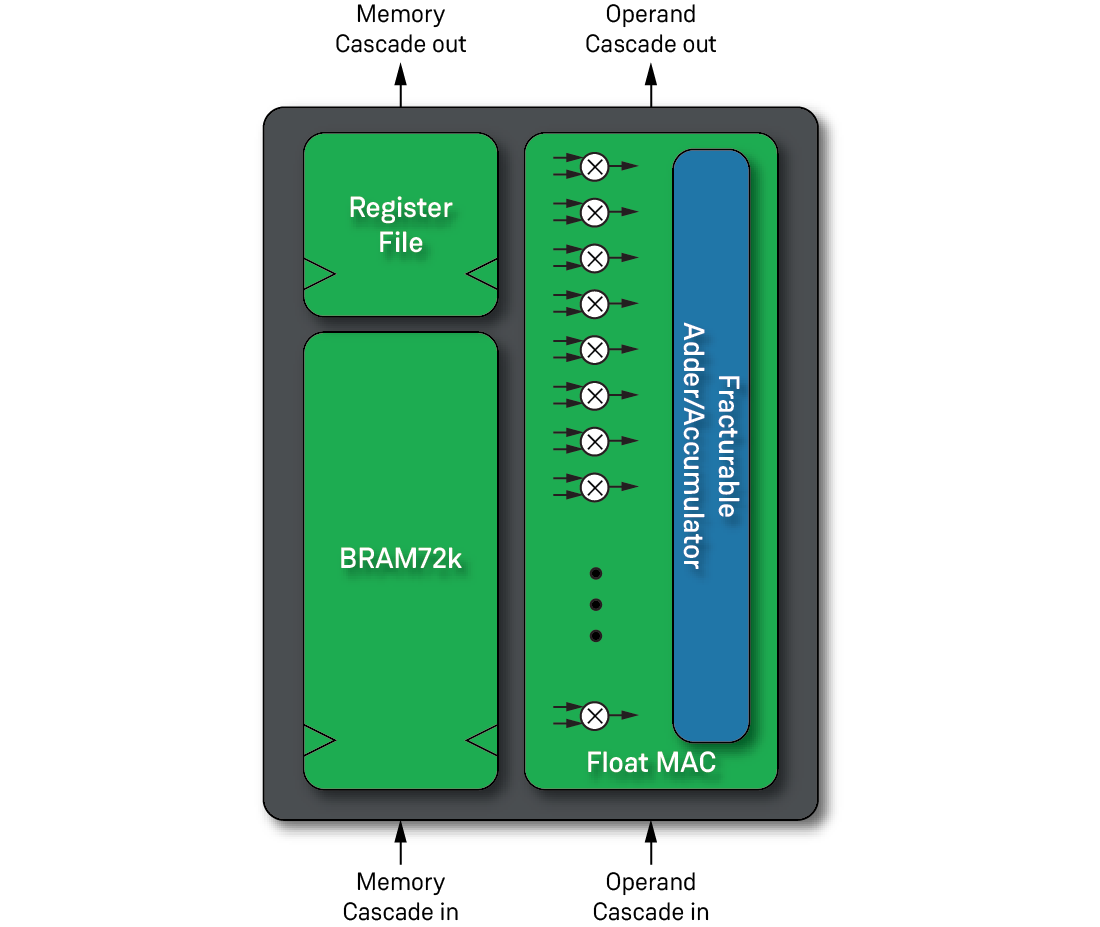


图3：*机器学习处理器*

在AC7t1500器件中使用了8个GDDR6存储控制器，它们总共可提供高达4 Tbps的双向带宽。如上所述，强大的计算引擎和大容量、高带宽的存储依赖于高速、低延迟和确定性的数据传输，以提供低延迟ASR应用所需的实时结果。

随后这些数据进入到Speedster7t的二维片上网络（2D NoC）。该二维片上网络是Speedster7t架构中的另一种硬结构，时钟频率高达2 GHz，可与所有I/O、内部硬模块和FPGA逻辑阵列本身互连。凭借20 Tbps的总带宽，2D NoC提供了最高的吞吐量，并通过适当的实现方式，可以在外部GDDR6存储器和支持MLP的计算引擎之间提供最具确定性的、低延迟的数据传输。

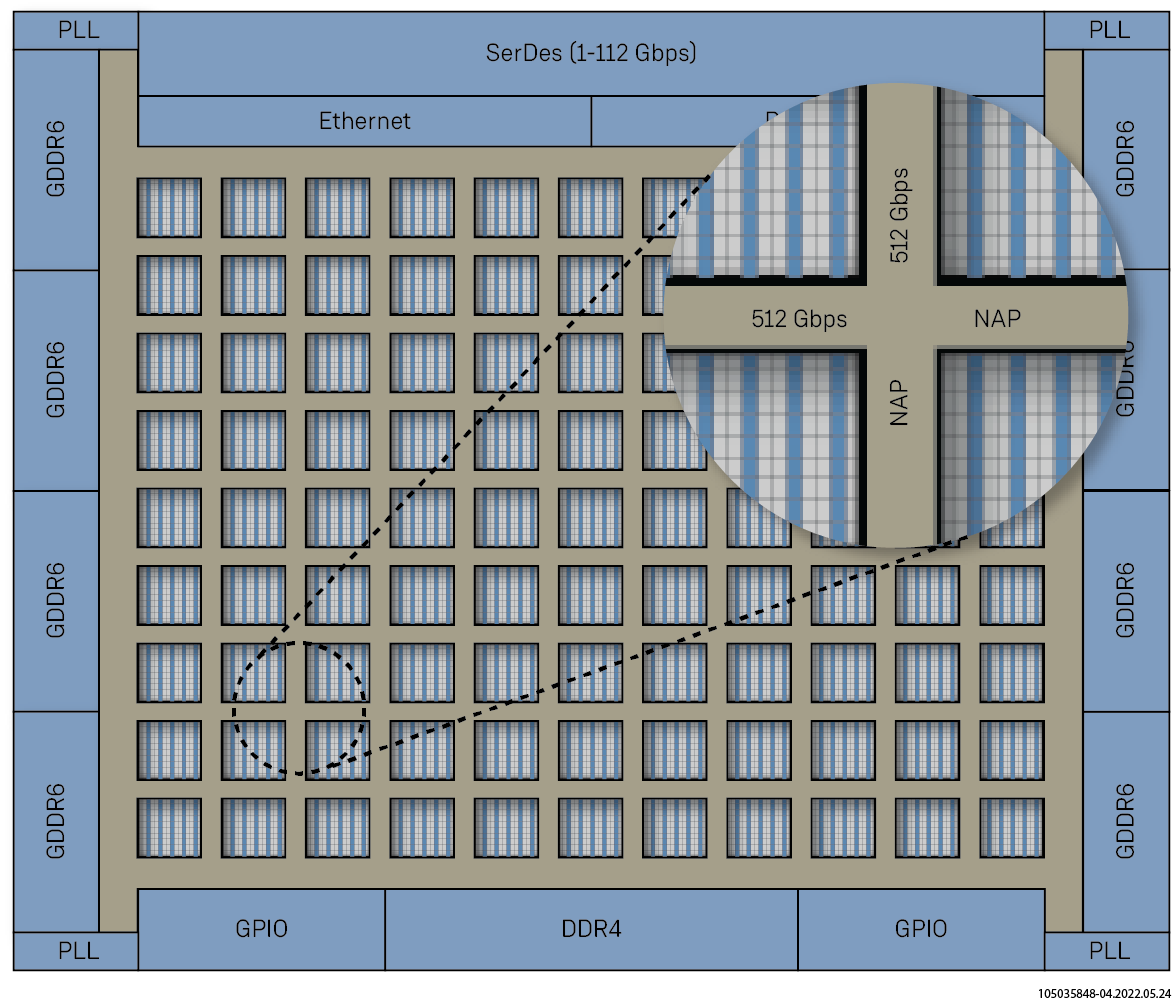


图4：*总带宽为20 Tbps的2D NoC*

与其他竞争性厂商的解决方案不同，2D NoC消除了Speedster7t ASR解决方案在存储器和计算引擎之间的任何瓶颈，在这些非常低的批处理速率下，硬件加速器的利用率达到最佳状态。

将所有这些功能放在一个roofline图中，就可以清楚地说明Achronix Speedster7t器件在低延迟ASR应用中相对于其他竞争性FPGA解决方案的优势。该roofline图使用了由每个制造商公布的经过验证的TOPS数据，展示了这些器件在实际应用中可以达到的效果。

下图显示了一个有效TOPS的roofline模型，它使用了Achronix为微基准（GEMV和MLP）和测试而构建的子集，以及公司A [4] [5]和公司B（基于架构）发布的数据。橙色的竖线表示批处理量为8毫秒和80毫秒音频模块的最佳操作点，用于低延迟、实时ASR数据流应用。在这个最佳操作点上，Achronix的有效TOPS比公司A提高了44%，比公司B的解决方案提高了260%。

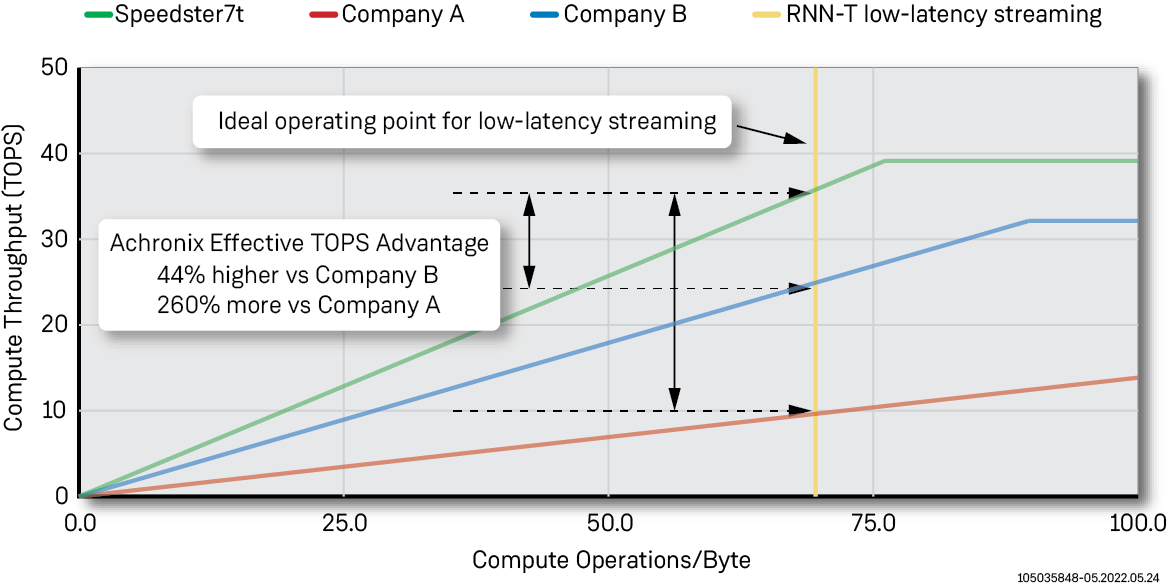


图5：*有效TOPS的Roofline模型*

在一年内实现ASR处理成本降低高达90%的目标

大多数ASR解决方案由Google、Amazon、Microsoft Azure和Oracle等大型云服务提供商提供。随着运营规模的扩大，以及这些产品在市场上取得的成功，在这些云API基础上构建产品的服务提供商面临着越来越高的成本压力。较大规模的ASR提供商公开宣传成本从每分钟0.01美元到0.025美元不等[6]、[7]、[8]、[9]。行业报告显示，呼叫中心的平均呼叫时间约为5分钟。考虑一个大型企业数据或呼叫中心服务公司每天要处理50,000通电话，每通电话5分钟。按照上述费率计算，ASR处理成本将是每天1,500至6,000美元或每年50万至200万美元。Achronix和Myrtle.ai的解决方案可以集成在一个加速卡上支持处理4000个RTS，每天可以处理超过一百万次的呼叫。

有许多因素会决定独立ASR设备的成本。在这个特定示例中，假设Achronix ASR加速解决方案是通过基于FPGA的PCIe卡提供，并集成到基于x86架构的2U服务器中。该设备从系统集成商那里出售，价格可能为50,000美元，而每年运行服务器的成本可能是这个成本的两倍。这样一来，本地ASR设备第一年的费用就达到了10万美元。将这种本地解决方案与云API服务进行比较，终端用户可以在第一年节省5到20倍的费用。

**表1：*Achronix ASR解决方案与云API服务的对比总结***

|  |  |  |  |
| --- | --- | --- | --- |
| **项目** | **Speedster7t解决方案** | **云API**  **（最低成本）** | **云API**  **（最高成本）** |
| 每天成本（50k RTS） | $275 | $1.5k | $6k |
| 年度成本 | $100k | $500k | $2,000k |
| 降低成本vs云API | 5X to 20X | – | – |

高度紧凑的系统使企业能够随着其业务的增加而扩展，而无需依赖日益昂贵的ASR云API，也无需构建庞大的数据中心基础设施来提供本地解决方案。

**总结**

CAI中的ASR功能要求RNN机器学习算法具有低延迟、高吞吐量的计算，这对现代AI加速器提出了挑战。声称推理速度高达150 TOPS的FPGA硬件加速器在大型计算引擎和高速存储器之间传输数据时会遇到瓶颈，这些瓶颈可能导致硬件利用率低至5%。Achronix和Myrtle.ai携手推出一个ASR平台，该平台由一个200W、x16 PCIe Gen4加速卡和相关软件组成，可以同时支持多达4000个RTS，每24小时可以处理多达100万个、时长5分钟的转录文件。将单台x86服务器上的PCIe加速卡与云ASR服务的成本相比，第一年的资本支出（CAPEX）和运营成本（OPEX）就可以降低高达90%。

欲了解更多信息，请通过网站[www.achronix.com/contact-us](http://www.achronix.com/contact-us)联系Achronix。

**参考资料：**

1. <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>
2. <https://www.microsoft.com/en-us/research/project/project-brainwave/>
3. Myrtle.ai未发表的白皮书，关于在Speedster AC7t1500器件中托管的ASR应用
4. M. Langhammer, G. Baeckler和S. Gribok，《SpiderWeb-高性能FPGA NoC》，*2020 IEEE国际并行和分布式处理研讨会（IPDPSW），*2020年，第115-118页，doi:10.1109/IPDPSW50202.2020.00025。
5. <https://arxiv.org/pdf/2010.06075.pdf>
6. <https://aws.amazon.com/transcribe>
7. <https://cloud.google.com/speech-to-text>
8. <https://www.ibm.com/cloud/watson-speech-to-text>
9. <https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service>
10. <https://www.techtarget.com/searchenterpriseai/feature/CNN-vs-RNN-How-they-differ-and-where-they-overlap>