SYNOPSYS®

WHITE PAPER

应用于边缘人工智能，基于USB4的灵活接口IP解决方案

作者

Morten Christiansen
USB/DisplayPort技术营销经理
新思科技

对于智能设备配备人工智能（AI）最新发展成果一事，消费者已经习以为常。为了拓展设备的整体目标市场，有创新意识的设备设计人员构建了边缘人工智能加速器和边缘人工智能SoC，以支持多种应用场景和集成方式。本白皮书介绍了一种应用于边缘人工智能加速器和SoC的基于USB4的灵活IP解决方案。该IP解决方案支持传统的PCIe 4.0、USB4、USB 3.x和USB 2.0连接，可用于多种类型的主机。本文所介绍的解决方案描述了一种小面积的人工智能加速器，可在多个应用中重复使用，从而创建一种具有成本优势的解决方案，可加快将新产品推向市场。

介绍

许多老牌和初创企业正在交付人工智能芯片，并且忙于规划新设计，旨在使芯片更具有成本优势、速度更快、更加先进。对于芯片设计，最重要的一个选择是人工智能芯片和处理其它功能主机之间的接口。接口的选择，决定了将边缘人工智能加速器芯片系统（SoC），与用于配置、控制和/或协同处理的支持应用处理器（也称为主机）进行连接的复杂性。接口决定了人工智能加速器支持人工智能应用和系统的容易程度及效率。

有些边缘人工智能系统是独立存在，而有些则采用了混合云/边缘配置，并使用网络连接进行进一步处理。视觉处理SoC就是混合云/边缘配置的例子，其任务是检测、分类并计算鱼类、野生动物、汽车、自行车或行人，然后将带有地理标记和时间戳的结果发送到云端，以供后续处理和分析。该用例不需要在视觉处理SoC和主机之间配置高吞吐量接口。

又如，我每天都使用智能网络摄像头，它配备了一个视觉处理SoC，与高清广角摄像头传感器和先进的麦克风阵列连接。视觉处理SoC经过训练，可以检测人的面部和躯干。它能够自动平移、倾斜和变焦，始终对焦会议室内的参与者，而且所有这些动作都实时完成。根据检测到的演讲者所在的位置，麦克风阵列可以对准该方向，以拾取演讲者的声音。摄像头以适当的分辨率和质量提取视频会议的视频流，对于视频和音频流，使用USB 3.x连接进行传输。在本例中，选择USB 3.x可以使摄像头连接大量视频会议主机。

野生动物计数器和智能摄像头是边缘人工智能SoC的例子，它们无需连接到主机即可执行人工智能加速功能。这些设备与主机连接仅用于执行有用的人工智能系统功能。主机和边缘人工智能加速器之间的带宽要求视具体用例而定，并且还确定了对于主机接口解决方案的需求。相同的用于野生动物计数器和智能摄像头的视觉处理芯片也可用于实时汽车驾驶辅助应用。在本例中，视觉处理SoC使用PCIe接口与一个强大的主机紧密连接在一起。

设想一下这样的情况：野生动物视觉处理器配有摄像头和其他传感器，使用隧道 PCIe与现成的低成本USB4主机连接。PCIe使边缘人工智能芯片和主机中的CPU、内存、存储和网络之间，实现快速且架构层面密切的连接。由于典型的 USB4 主机体积很小、成本和功耗都很低，因此，支持USB4的边缘人工智能系统可以轻松部署。这样，视觉处理器可以使用实时数据进行实时训练，并与云端通信。此外，该应用可用于验证边缘人工智能应用，然后将应用融合到专用服务器解决方案中进行全面部署。

目前，许多具有人工智能加速能力的边缘视觉处理芯片仅使用PCIe与主机对接。增加USB4能力可简化边缘人工智能系统集成，支持多种应用场景，并且通过正确的实施，不会使面积增加。对于当前采用传统USB（USB 3.x或更早版本）主机连接的人工智能加速器设计，也应考虑加入USB4的能力。基于USB4的接口解决方案，允许边缘人工智能加速器芯片在不同主机上重复使用，以满足多个应用需求。且集成多个接口既加快了上市速度（无需开发多个芯片版本），又拓展了目标市场。与现有的传统接口解决方案相比，额外的芯片成本可以降到最低，原因是无需对PCIe和USB PHY分别进行实例化。

为了设计并部署边缘人工智能加速SoC，尽可能占领市场份额，设计人员可以考虑灵活的基于USB4的接口解决方案，因为这种解决方案可以支持PCIe 4.0、USB4、USB 3.x和USB 2.0工作模式。本白皮书介绍了这种解决方案的优点、挑战、实施选项和集成技巧。

边缘人工智能计算器举例

图1显示了典型边缘人工智能加速器系统的总体框图，这种系统可能会在野生动物计数器中用到。基于实际应用，边缘人工智能加速器使用USB4或隧道PCIe与通用还是专用主机连接。与当前正在开发的大多数人工智能加速SoC一样，图1显示了通用的情况，而且由于USB4和PCIe的集成，这种通用情况可支持多种用例。

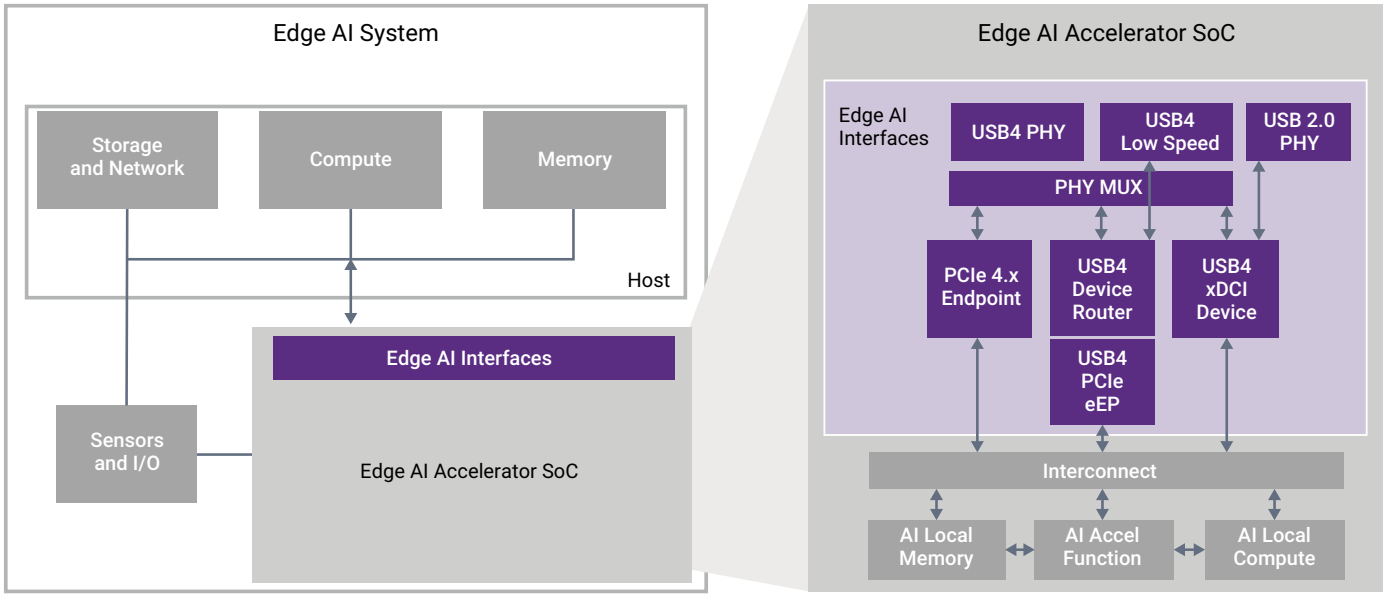

图1：边缘人工智能加速器系统集成了USB4和PCIe接口，以提高应用的灵活性。

图1所示的边缘人工智能加速器接口设计可以在传统PCIe模式、带有隧道PCIe的USB4模式、传统USB3.x流媒体或传输模式以及USB2.0传输模式下运行。设计多接口芯片，允许边缘人工智能加速器或SoC在设计中重复使用，并降低或者消除了对同一个边缘人工智能加速器，或SoC设计的多个变化进行重新设计、流片和验证的需求。边缘人工智能加速器可以根据需要连接多种类型的主机，用于边缘人工智能系统的多种用例和应用。

传统PCIe模式

PCIe 得到了用于边缘人工智能系统的典型主机的广泛支持。许多现有和规划中的边缘人工智能加速器都使用传统PCIe（PCIe 4.0或更早版本），因为这是一种高效且易于理解的接口选项，过去一直用于满足边缘人工智能市场的需求。如图1所示，传统PCIe模式使用与标准PCIe端点连接的USB4 PHY。USB4 PHY在传统PCIe模式下运行。USB4设备路由器、USB4 PCIe嵌入式端点和USB4 xDCI设备控制器都被禁用。边缘人工智能加速器可以使用PCIe规范的所有特性，并通过DMA对主机内存进行高效访问。边缘人工智能接口支持PCIe 4.0 x2，可提供高达32Gbps的原始吞吐量。

边缘人工智能加速器的物理实施方案（见图2）包含一个PCIe插卡以及一个或多个边缘人工智能加速器。然而，这通常需要一个服务器型主机，可容纳一个或多个PCIe插卡。许多物理体积更小的主机支持常用作固态磁盘的M.2插卡，但也兼容传统PCIe模式下的边缘人工智能加速器。就主机的物理尺寸而言，最不灵活但最优化的选择是将边缘人工智能加速器集成到主板上。这种解决方案需要实施定制的嵌入式的主机设计。

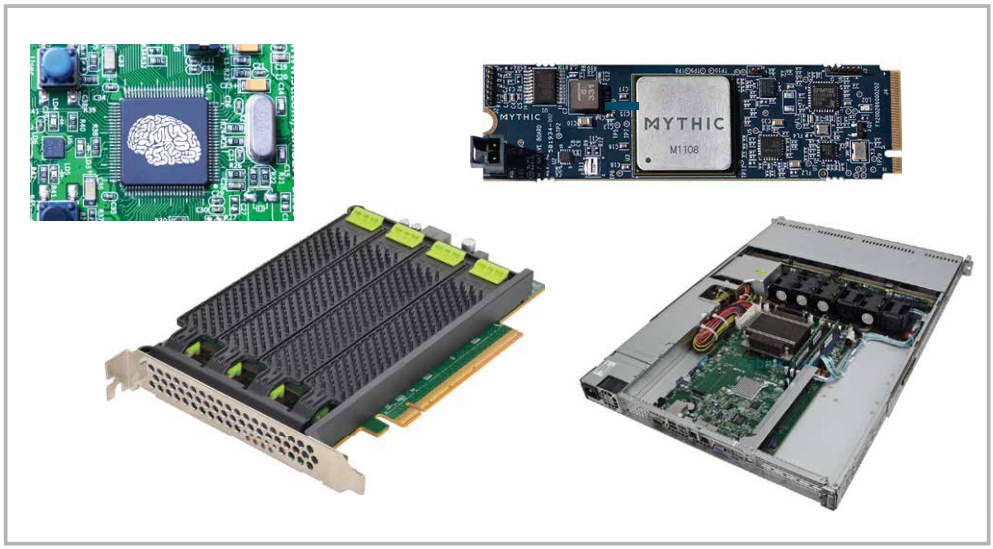

图2：当前传统PCIe实施选项举例

图片来源：https://techcrunch.com/2020/06/08/mits-tiny-artificial-brain-chip-could-bring-supercomputer-smarts-to-mobile-devices/; https://www.mythic-ai.com/product/m-2-cards/; https://docplayer.net/140492619-Qct-rackgo-x-ocp-ava-4-m-2-carrier-card-product-marketing-specification.html, page 6; https://www.abmx.com/1u-rackmount-server

USB4模式

对于配置定制服务器或无法使用PCIe插卡的边缘人工智能系统，可以使用USB4。在USB4模式下，USB4 PHY以USB4模式运行。PCIe端点和USB4 xDCI设备控制器被禁用。USB4设备路由器插入并提取通过USB4传输的PCIe流量。USB4设备路由器与PCIe嵌入式端点连接。由于USB4设备路由器和PCIe端点之间的物理连接无PHY，因此需要一个嵌入式端点。

图2展示了使用USB4的边缘人工智能系统的一些典型的物理实现方法。物理实现包括一个（小型）自供电或总线供电并带有Type-C接口的机箱，通过标准USB Type-C线缆与USB4主机连接。固定线缆不太常用。另一个实施选项是“计算棒”，即一个类似于USB指状储存器的设备，带有Type-C接口，可插入USB4主机、USB4集线器或USB4坞。

需要注意的是，目前，大多数服务器级的主机并不支持USB4。然而，分立式USB4主机控制器已得到广泛使用，使得多种主机都能支持USB4。另外，许多利用笔记本处理器的小型台式主机也支持USB4。这类主机可能比典型的服务器级主机更适合许多边缘人工智能系统，因为它们一般更便宜、体积更小，而更容易获得。

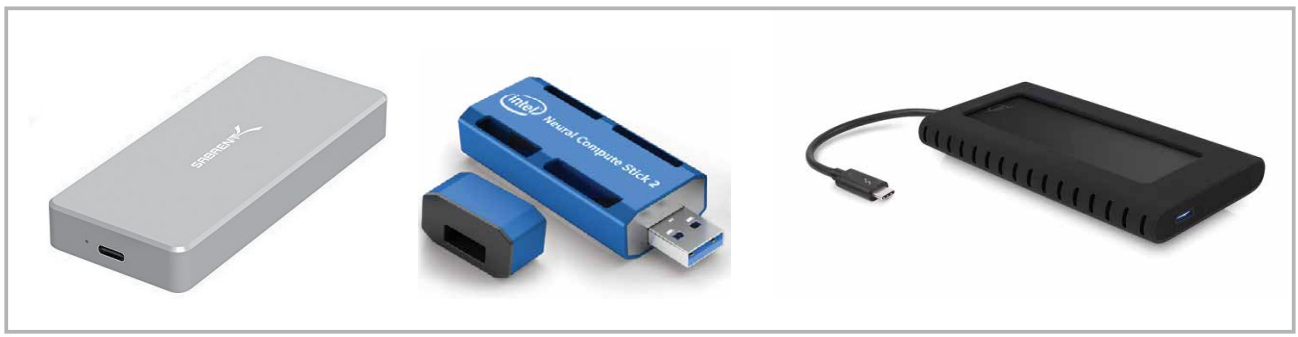

图3：支持隧道PCIe和传统USB的USB4的典型实施方法

图片来源：https://www.bhphotovideo.com/c/product/1448870-REG/sabrent_ec_nvme_usb_3_1_aluminum_enclosure.html; https://software.intel.com/content/www/us/en/develop/hardware/neural-compute-stick/where-to-buy.html; https://eshop.macsales.com/shop/owc-envoy-pro-ex-thunderbolt-3

USB4带宽

多个边缘人工智能加速器可以连接一个USB4主机，并行处理一个复杂的任务，或者执行由主机控制的多个独立任务。例如，一辆装有八个互连摄像头的汽车配有一个主机，来协调摄像头的输入数据。通常，对于带有集成式USB4控制器的USB4主机，每个USB4端口支持32Gbps的PCIe带宽，而对于带有外部独立USB4控制器的USB4主机，同一个控制器上的USB4端口之间通常共享32Gbps的PCIe带宽。高性能USB4主机可以添加多个外置USB4控制器，为需要高PCIe带宽的应用提供最大的USB4带宽。

与传统USB一样，USB4总线带宽由连接到同一USB4端口的所有设备共享。在多个边缘人工智能加速器通过USB4集线器或USB4坞连接到同一个USB4主机端口时，USB4主机端口的可用带宽将由多个边缘人工智能加速器共享。然而，两个具有隧道PCIe的USB4设备共享一个40Gbps的USB4端口时，获得的带宽与每个设备通过传统PCIe模式直接连接的情况相同或更高。

与传统USB一样，实际设备吞吐量取决于USB4端口上同时连接并保持活动的其他设备。例如，如果一个USB4主机在驱动4K显示器的同时，在同一个USB4端口上为一个边缘人工智能加速器提供服务，则DisplayPort流量将优先考虑。因此，对于使用隧道PCIe模式的USB4的应用，在设计时必须能够容忍不同的带宽。使用USB4模式的边缘人工智能系统的集成商必须清楚这一点，并确保有足够的带宽可用。

传统USB模式

传统USB模式的关键优势在于：传统USB应用广泛，允许边缘人工智能加速器与各种主机配合使用。在传统USB模式下，流式传输或批量传输都可以使用。如图1和图4所示，USB4/PCIe PHY在传统USB 3.x模式下运行，或在USB 2.0模式下被禁用。PCIe端点、USB4设备路由器和USB4 PCIe eEP都被禁用，而USB4 xDCI设备控制器被启用。

在传统USB 3.x模式下，连接速度为SuperSpeed 5Gbps (USB 3.0)、SuperSpeed 10Gbps (USB 3.1) 或SuperSpeed 20Gbps (USB 3.2)。连接速度取决于USB 3.x主机能力和线缆长度。只有长度小于1米的线缆才支持SuperSpeed 10Gbps和SuperSpeed 20Gbps。较古老的服务器级主机仅支持USB 3.0，而现代服务器级主机则支持USB 3.1和USB 3.2。大多数嵌入式主机和单板计算机基于广泛应用且价格低廉的移动、消费或工业应用处理器，可支持USB 3.0，而有些支持USB 3.1。基于常见笔记本电脑处理器的小型台式机或嵌入式服务器通常支持USB 3.1，而对USB 3.2的支持也开始出现。

对于USB4模式，边缘人工智能加速器的典型物理实施方案，通常是将其作为一个带有Type-C连接器或固定线缆的（小）盒子，或者作为“计算棒”。计算棒可以有Legacy-A或Type-C插头。Type-C插头是首选，因为这种插头允许一种设计可以在USB4模式或传统USB3.x模式下运行，而这取决于主机的能力。在使用legacy-A插头时，边缘人工智能加速器仅能达到传统USB3.0或USB3.1的速度。

盒子最好使用Type-C插座，这样才能以最灵活的方式实施，并实现与普通主机最全面的兼容性。需要注意的是，SuperSpeed micro-B已被废弃，不应继续采用。

传统USB 3.x模式——流模式

USB 3.x流模式使用USB同步传输，优点是每个服务间隔的带宽有保障。实际带宽取决于使用USB 3.0、USB 3.1还是USB 3.2，以及xHCI控制器和是否支持高带宽同步传输。每个服务间隔可以支持48kBytes或96kBytes，吞吐量最高可达768MB/秒。需注意的是，同步传输不能保证主机和设备之间所有数据包都能交付。这意味着设备和/或主机必须能够容忍偶尔的数据包丢失。规范规定并假设传统USB的误码率优于10E-12，这意味着数据包丢失数量相当少，但仍必须适当考虑并处理。

对于同步传输模式，在从数据可用到数据在总线上传输的过程中，这种模式增加了一些延迟。最好的情况是延迟局限于一个服务间隔，可低至125us。然而，实际延迟取决于使用的USB设备类别、编程模型以及主机和设备的同步传输调度算法。如果低延迟对于边缘人工智能加速器的用例至关重要，设计人员必须分析和设计实际实施的延迟。

边缘人工智能加速器可以使用批量传输或中断传输的混合模式，并且保证配置、状态、处理结果等可以正确执行，并对进出边缘人工智能加速器的‘原始’进行同步传输。

传统USB 3.x模式——传输模式

USB 3.x传输模式通常使用USB批量传输，在检测到错误和重新传输时保障数据的交付。然而，带宽和延迟无法保证，因为批量传输是“尽力而为”的做法，其优先级排在控制、同步和中断传输之后。需要注意的是，USB 3.0、USB 3.1或USB 3.2批量传输模式连接的实际吞吐量高度依赖传输的数据量。对于USB 3.2 SuperSpeed 20Gbps批量传输，最佳情况是1Mb的传输吞吐量约为2Gb/秒。然而，对于4Kb数据的传输，主机和设备之间直连的吞吐量为3-400Mb/秒，在使用USB集线器时，吞吐量甚至更低。

对于USB4模式，端口的总线带宽由连接到同一主机端口的所有设备共享。此外，对于一些传统USB3.x主机，全部或部分USB端口之间也共享带宽。例如，如果一个USB连接的网络摄像头用来向主机提供数据，边缘AI加速器的可用带宽可能会大大减少。系统集成商应该了解这些问题，并选择适当的主机和/或系统设计，以提供所需要的带宽。

传统USB 2.0模式

有些边缘人工智能系统可以成功使用USB 2.0连接。在传统USB3.x模式下，xDCI控制器处于活动状态。而USB4 PHY只有使用图2中的配套USB 2.0 PHY时才处于活动状态。根据使用场景，可以使用同步模式或传输模式。USB2.0连接的优点在于较低的功率和与主机的全面兼容性。价格低廉且广泛使用的嵌入式USB 2.0主机很适合大型边缘人工智能的部署，如低成本单板计算机，例如Raspberry Pi等。

USB 2.0模式的缺点是吞吐量较低，通常，高速USB的吞吐量为30Mb/秒，而全速USB的吞吐量则低于1Mb/秒。然而，如果边缘人工智能加速器集成摄像头或其他高带宽传感器接口，并使用USB 2.0连接传输配置、控制、状态和结果时，USB 2.0模式很有用。例如，前文所述的智能网络摄像头中的视觉处理器可以经过重新训练，以检测进出商店的人、检测没有戴口罩的人、检测商店某些区域的拥堵情况，并且只输出计数和/或警报，而非高带宽视频和音频流。

用于边缘人工智能加速器的DesignWare USB4 PHY IP

想要设计出灵活的边缘人工智能加速器接口，关键在于使用新思科技的DesignWare USB4 PHY IP。DesignWare USB4 PHY IP支持USB4（包括隧道PCIe）和传统USB。对于边缘人工智能加速器，DesignWare USB4 PHY IP可以通过升级，以额外支持传统PCIe。边缘人工智能加速器支持的操作模式如图4所示。DesignWare USB4 PHY IP正在为7/6/5/4nm等高级工艺节点和12nm等低成本工艺节点进行开发。

USB4/USB 3.x/PCIe 4.x © = USB 3.2 Config Lane							
Type-C name	TX1+/-	RX1+/-	TX2+/-	RX2+/-	D+/D-	SBU	**PHY Operating Mode**
Connector Orientation	USB TX or PCIe TX	USB RX or PCIe RX	USB TX or PCIe TX	USB RX or PCIe RX	USB 2.0	USB4 LS	
Type-C Normal	SS	SS	Not used	Not used	D+/D-	n/u	**USB 3.0 5G x1**
Type-C Flipped	Not used	Not used	SS	SS	D+/D-	n/u	
Type-C Normal	SS+	SS+	Not used	Not used	D+/D-	n/u	**USB 3.1 10G x1**
Type-C Flipped	Not used	Not used	SS+	SS+	D+/D-	n/u	
Type-C Normal	SS/ SS+ © Lane	SS/ SS+ © Lane	SS/ SS+ © Lane 1	SS/ SS+ Lane 1	D+/D-	n/u	**USB 3.2 5G/10G x2**
Type-C Flipped	SS/ SS+ Lane 1	SS/ SS+ © Lane 1	SS/ SS+ © Lane	SS/ SS+ © Lane	D+/D-	n/u	
Type-C Normal	Lane 0	Lane 0	Lane 1	Lane 1	D+/D-	TX/RX	**USB4 10G x2**
Type-C Flipped	Lane 1	Lane 1	Lane 0	Lane 0	D+/D-	TX/RX	
Type-C Normal	Lane 0	Lane 0	Lane 1	Lane 1	D+/D-	TX/RX	**USB4 20G x2**
Type-C Flipped	Lane 1	Lane 1	Lane 0	Lane 0	D+/D-	TX/RX	
Legacy PCIe	Lane 0	Lane 0	Lane 1	Lane 1	n/u	n/u	PCIe x2

图4：边缘人工智能加速器所需的操作模式

总结

下一代边缘人工智能加速器需要支持传统PCIe 4.0、USB4、传统USB 3.x和USB 2.0操作模式的灵活USB4接口。要想成功实施，设计人员需要了解集成的优点和挑战，并研究可以加速设计过程的IP。

借助DesignWare IP，设计人员可以创建支持USB4和PCIe协议的边缘人工智能加速系统，而这种系统的面积约为传统离散协议实施的一半。与单独的PCIe或单独的传统USB设计相比，在边缘人工智能加速器设备中，实施DesignWare USB4 PHY接口IP可提供相当高的灵活性，并增加边缘人工智能加速器的功能和接口选项。DesignWare USB4 IP允许在不同应用中重复使用边缘人工智能加速器芯片，根据每个应用的需要连接多种主机。无需开发多个版本的边缘人工智能加速器芯片，以便拓宽目标市场，并加快新型边缘人工智能应用的上市速度。

新思科技和USB-IF

新思科技是USB技术和标准的主要贡献者；新思科技员工对几乎所有已发布的USB标准都做出了贡献。对于从USB 1.1到USB4的所有USB标准，新思科技提供了世界上最受青睐且使用最广泛的 DesignWare USB IP。欲了解关于我们如何帮助您开启下一个设计，请联系新思科技。

USB4资源：

USB4白皮书: https://www.synopsys.com/dw/doc.php/wp/usb4_user_expectations_drive_design_complexity.pdf

USB4规范: https://www.usb.org/document-library/usb4tm-specification

USB4开发者日演讲: https://www.usb.org/documents?search=usb4&items_per_page=50

07/09/21.cs715103362-SG-USB AI wp-chinese.

Pub: July 2021