白皮书

FPGA
内存带宽，HBM

利用英特尔® Agilex™ M 系列 FPGA 应对内存带宽和计算密集的挑战

作者

Supriya Velagapudi
内存 IP 产品营销经理
英特尔可编程解决方案事业部

Mark Honman
FPGA IP 软件设计工程师
英特尔可编程解决方案事业部

目录

引言 1
业界努力应对内存挑战 1
利用英特尔® Agilex™ M 系列 FPGA 应对内存挑战 2
- 内存层次结构 2
- 片上内存 2
- 封装内存（HBM2e） 2
- 片外内存（DDR5、LPDDR5 等） 3
- 英特尔® 傲腾™ 持久内存 3
- 大 I/O 容量 3
- 出色的数字信号处理（DSP）能力 3
- 片上网络（NoC）功能 3
- 横向与纵向网络 3

用例 4
5G 射频模拟硬件在环测试 4
浅水模型（SWM） 5
- 数据处理流程 5
- 内存要求 6
- 实施 6

结论 7

引言

从数据中心到网络，再到边缘，FPGA 在现代应用中正发挥越来越重要的作用。FPGA 的灵活性、出色能效、大规模并行架构和超高的输入/输出（I/O）带宽使其在加速高性能计算（HPC）、人工智能（AI）、存储和网络等广泛任务方面非常具有吸引力。这些应用中有很多对内存提出了严苛要求，包括内存的容量、带宽、时延和能效等。

为了满足这些应用的严苛要求，英特尔开发了英特尔® Agilex™ M 系列 FPGA，这也是颇为成功的英特尔® Stratix® 10 MX 设备系列的下一代产品。英特尔® Agilex™ M 系列设备基于英特尔® 7 制程工艺实施，实现了更高的可编程逻辑结构容量和性能，而且功耗更低。

英特尔® Agilex™ M 系列 FPGA 达到了 FPGA 行业内少有的内存带宽水平，也是英特尔® Agilex™ 设备系列中第一款提供封装 HBM2e 内存的产品。英特尔® Agilex™ M 系列设备还包括面向其他先进内存技术（如 DDR4、DDR5 和 LPDDR5）的硬核控制器。两个硬核内存片上网络（NoC）功能使 FPGA 逻辑结构能够提供对封装 HBM2e 和板载内存资源的高带宽、资源高效型访问。

此外，英特尔® Agilex™ M 系列 FPGA 还具备出色的收发器数据速率，这对于当今需要处理海量数据负载的系统而言至关重要。英特尔® Agilex™ M 系列设备支持 PCI Express（PCIe）Gen5、Compute Express Link、400 G 以太网以及运行速度可高达 116 Gbps 的串行收发器，可满足数据中心和边缘等严苛应用的吞吐量需求。

英特尔® Agilex™ M 系列 FPGA 将为多个市场带来益处，包括但不限于：测试与测量（任意波形发生器、5G/6G 蜂窝网络测试、GHz 射频测试）、数据中心（高性能计算、云计算、货币数字化）、无线和有线网络（888G+ 高速率数据传输、光传输网络、网络功能虚拟化、5G 基带）以及航空航天和国防（雷达、电子战）。

业界努力应对内存挑战

如今的计算工作负载比过去规模更大、更复杂、更多样化。高性能计算、人工智能、机器视觉、视频、游戏、数据分析和其他专业性任务的爆发式增长正在推动数据量呈指数级增长。根据 Statista 的预测[1]，仅 2021 年产生的数据量就多达 74 ZB（1 ZB 等于 1 万亿 GB）。相比之下，2019 年产生的数据量为 59 ZB；2020 年为 41 ZB；且数据增长仍在加速。

[1] https://www.statista.com/statistics/871513/worldwide-data-created/

长期以来，DDR 内存因能满足许多开发人员的内存需求而受到青睐。然而近年来，对更大带宽、更高容量和更高能效的需求增长超过了 DDR 性能的增长速度。因此，我们需要更强大的解决方案。此外，固定的能耗预算和小尺寸外形规格限制也意味着我们必须在相同的空间内实现更多的功能。

利用英特尔® Agilex™ M 系列 FPGA 应对内存挑战

在我们考虑英特尔® Agilex™ M 系列 FPGA 如何满足当今的内存带宽和容量需求之前，让我们先概览一下该系列 FPGA 的架构。这些设备采用系统级封装（SiP）技术构建（见图 1）。

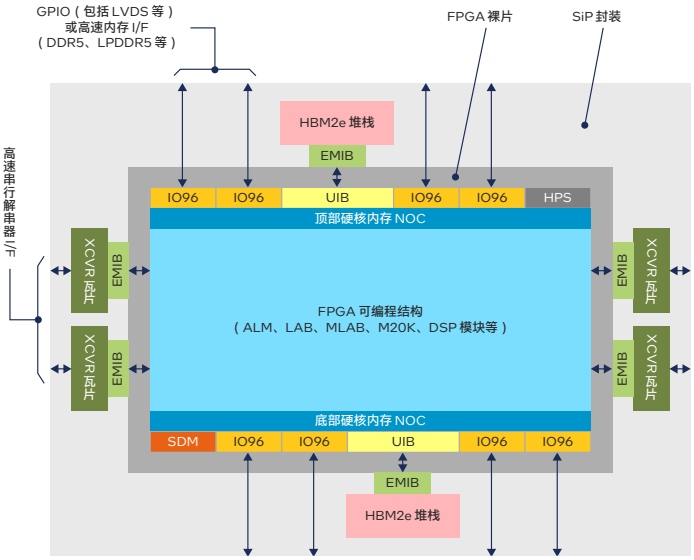

图 1. 英特尔® Agilex™ M 系列 FPGA 平面图

除了主 FPGA 裸片，还有四个收发器瓦片（XCVR tile）和两个 HBM2e 堆栈。XCVR 瓦片（tile）和 HBM2e 堆栈使用英特尔® 嵌入式多芯片互连桥接（EMIB）技术集成到 FPGA 裸片，这是实现异构芯片封装内高密度互连的一种简练且经济高效的方法。这样，所有这些芯片就都可以作为一个大型裸片来使用。

内存层次结构

当今有很多应用都需要对内存资源进行层次划分。这种层次结构使得设计团队能够在超低时延、超高带宽的片上内存（MLAB 和 M20K 模块）、较高容量、较高带宽的封装内存（HBM2e）和超高容量的板载内存（DDR4、DDR5、LPDDR5 等）之间做出时延与容量上的权衡（见图 2）。

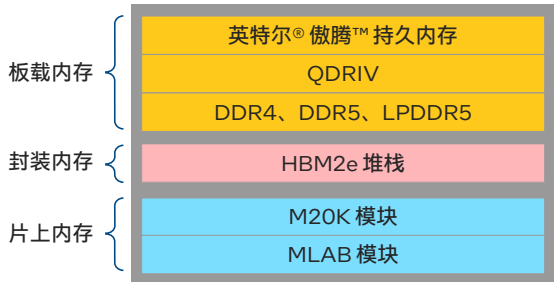

图 2. 英特尔® Agilex™ M 系列 FPGA 内存层次结构

英特尔® Agilex™ M 系列 FPGA 可提供大量资源来应对内存带宽挑战。我们将从内到外逐一介绍该系列产品内存的各个层次，包括 FPGA 逻辑结构中的超本地化片上内存、以 HBM2e 堆栈形式提供的本地封装内存，以及 DDR5 和 LPDDR5 等外部内存架构和接口。

片上内存

当设计人员需要最本地化的内存时，没什么能比得上以 MLAB 模块和 M20K 模块的 RAM 形式集成在可编程逻辑结构中的片上内存资源。

封装内存（HBM2e）

封装 HBM2e 弥合了内存层次结构中一个巨大而又关键的缺口，以满足当今数据密集型应用的需求。其容量远大于片上内存（两个数量级以上），同时带宽又远大于片外内存（两个数量级以上）。

通过将高性能的 HBM2e 与 FPGA 裸片集成在同一封装中，我们便可以在小尺寸外形规格中实现更高带宽、更低功耗、更低时延。此外，由于 HBM2e 集成在封装中，因此也不需要使用外部 I/O 引脚，从而节省了电路板空间，并消除了它们会带来的功耗和互连时延。

每个 HBM2e 堆栈可包含 4 层或 8 层，每层提供 2 GB 内存，因此单个英特尔® Agilex™ M 系列 FPGA 可包含 16 GB 或 32 GB 的高带宽内存。每个堆栈具有一个通用接口总线（UIB），其中包括八个硬核控制器和物理层。每个硬核控制器服务于一条 HBM2e 通道，每条通道又可分解为两条伪通道（PC）。这样就可以尽可能提高所有事务的性能，实现每堆栈高达 410 Gbps 的内存带宽，较 DDR5 组件的带宽提升高达 18 倍，较 GDDR6 组件提升 7 倍。两个 HBM2e 堆栈加起来可提供高达 820 Gbps 的峰值内存带宽[2]。每 HBM2e 堆栈支持 8 条通道或 16 条伪通道，可用于数据进出堆栈的路由。

[2] 当使用 HBM 设备中的纠错码（ECC）位来存储数据时，系统级吞吐量可提高 12.5%。

片外内存（DDR5、LPDDR5 等）

对于超出 HBM2e 容量的应用需求，或在需要独立内存的灵活性时，英特尔® Agilex™ M 系列 FPGA 支持高性能 DRAM 的全新规格：DDR5 和 LPDDR5，以及其他主流的内存架构。

英特尔® 傲腾™ 持久内存

英特尔® 傲腾™ 持久内存填补了内存/存储结构中 DRAM 等易失性内存和闪存等传统持久性存储之间一个关键空白。英特尔® 傲腾™ 持久内存的密度比 DRAM 高，存取速度又比闪存快。它与 DRAM 一样封装在 DIMM 中，与 DRAM 位于相同的总线/通道，能够以与 DRAM 相同的方式工作，区别是存储在英特尔® 傲腾™ 持久内存中的数据是非易失性的。

大 I/O 容量

对于当今的数据密集型应用来说，数据进出 FPGA 至关重要。英特尔® Agilex™ M 系列 FPGA 通过能够支持高达 116 Gbps PAM4、CXL、PCIe Gen5、400 G 以太网和多种其他协议的收发器，实现了非常大的 I/O 带宽。

英特尔® Agilex™ M 系列设备可支持高达 768 个主 I/O 连接到增强型 IO96 子系统。这些引脚可以用作通用型 I/O（GPIO），支持各种电气接口，如低压差分信号传输（LVDS），或作为板载内存设备的高速接口。同时，XCVR 瓦片提供高速 SERDES（串行/解串器）接口，可执行 PCIe Gen5 和 400 G 以太网等通信协议。

出色的数字信号处理（DSP）能力

英特尔® Agilex™ M 系列设备包含多达 12,300 个可变精度 DSP 模块，每个模块包含两个 18x19 DSP 增效器。这些 DSP 模块可支持高达 18.5 TFLOPS 的单精度浮点运算、37 TFLOPS 的半精度浮点运算和 88.6 TOPS 的 INT8 运算。这些 DSP 模块的浮点运算能力使英特尔® Agilex™ FPGA 的性能优于只支持定点运算的传统 FPGA。

片上网络（NoC）功能

英特尔® Agilex™ M 系列 FPGA 的一个关键差异化优势在于增加了两个硬核内存片上网络（NoC）功能，这有助于在 FPGA 的可编程逻辑结构和集成 NoC 的内存之间实现高带宽数据传输，而无需使用现有的 FPGA 路由资源。每个片上 HBM2e 堆栈通过 UIB 与其 NoC 通信。片外内存（DDR4、DDR5 等）则通过 IO96 子系统与 NoC 通信（见图 3）。

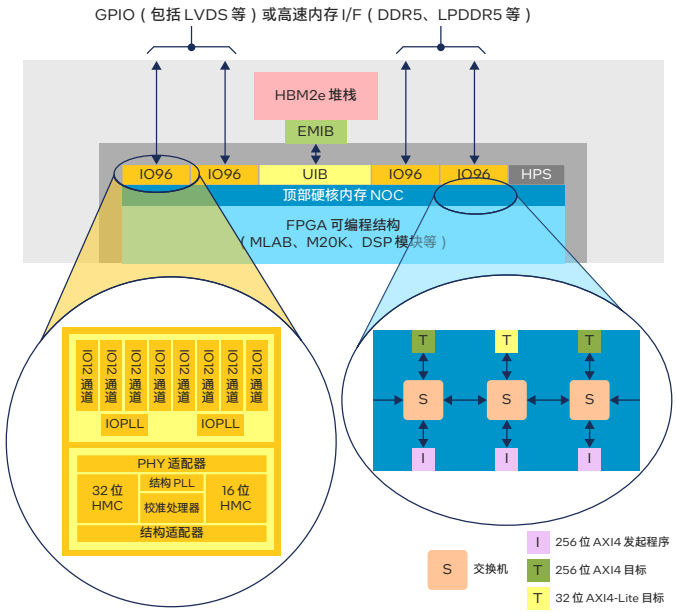

图 3. IO96 子系统和部分顶部硬核内存 NoC 详解

NoC 通过一个由交换机（路由器）、互连链路（导线）、发起程序（I）和目标（T）组成的网络，将数据从数据源传输到目的地。

横向与纵向网络

每个 NoC 都提供一个横向网络，通过 AXI4 发起程序将可编程逻辑结构中的逻辑连接到集成 NoC 的目标内存。此外，每个 NoC 也都提供一个纵向网络，通过优化的路由将横向网络路径读取的内存数据分发到 FPGA 的可编程逻辑结构深处（见图 4）。

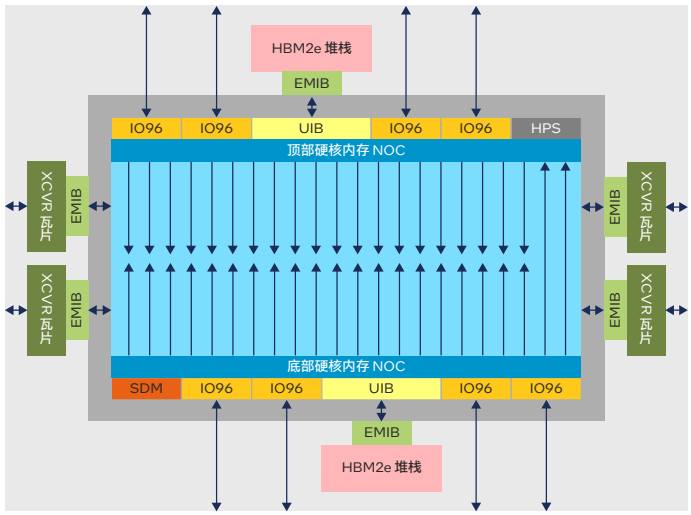

图 4. 纵向网络可将读取的数据从集成 NoC 的内存传输到可编程逻辑结构深处的 M20K 模块

顶部 NoC 有 20 个 256 位的发起程序，底部 NoC 有 22 个 256 位的发起程序。每个发起程序的带宽是 256 位 x 700 MHz = 22.4 Gbps。将读取数据从集成 NoC 的内存传输到可编程逻辑结构和/或 M20K 模块时，顶部 NoC 的带宽为 256 位 x 700 MHz x 20 个发起程序 = 3.58 Tbps，底部 NoC 的带宽为 256 位 x 700 MHz x 22 个发起程序 = 3.94 Tbps，所以总带宽为 3.42 + 3.76 = 7.52 Tbps。

通往 NoC 的接入点即上述的发起程序和目标。可编程逻辑结构中的用户逻辑连接到 256 位 AXI4 发起程序以发起请求并发送数据（可编程逻辑结构上的每个发起程序可通过用户时钟独立计时），而 256 位 AXI 目标则提供来自集成 NoC 的内存的响应。每个发起程序都可以跟目标通信，从而使用户可以灵活地实现一个充分强化的交叉开关矩阵 (crossbar)。NoC 中的交换机使用专有协议在发起程序和目标之间传输请求和响应。

用例

下面介绍的用例[3] 旨在反映计算工作负载与内存流量的比率，以及真实场景下工作负载的内存存取模式。这些用例包括 5G 射频模拟硬件在环测试和浅水模型 (SWM)，后者是天气预报中会用到的代表性工作负载。

5G 射频模拟硬件在环测试

对于制造射频和微波模拟设备（例如用于雷达和 5G 基站等应用的放大器或其他模拟电路）的公司来说，必须使用高输入频率对设备的响应进行测试。响应中的异常会在很宽的频率范围内表现出来，可能比工作频率更低，也可能比谐波噪声等引起的频率更高。

常见的测试方案是使用硬件在环 (HIL)。在这个特定的用例中，英特尔® Agilex™ M 系列 FPGA 就是测试系统中的一部分 (见图 5)。

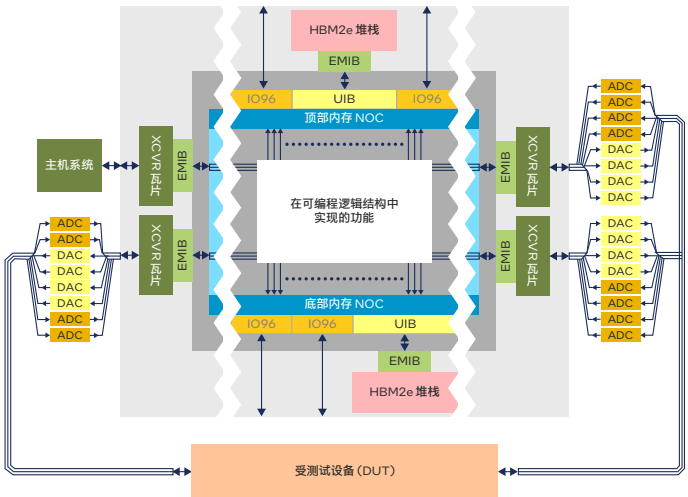

图 5. 英特尔® Agilex™ M 系列 FPGA 用于 5G 射频 HIL 测试台

[3] 由英特尔工程部门开发的用例

主机系统通常采用 CPU，它会生成自定义波形，一般持续几毫秒。这一步经过了精心设计，用来运行受测试模拟设备 (DUT)，从而暴露潜在的问题和缺陷。这些缺陷将表现为 DUT 对刺激波形的响应中的错误。

对于高度集成的模拟产品，主机系统将上传（输入）多个完全一致的超高频波形，并下载 (采集) 多个响应输出波形，随后分析结果。数据加载和检索同时进行。根据分析结果，新的刺激波形再投入应用。

在这个特定的用例中，FPGA 通过其中一个 XCVR 瓦片与主机系统进行双向通信，这个 XCVR 瓦片被配置为一个具备 64 Gbps 带宽的 PCIe Gen5 x16 接口。FPGA 使用剩余的三个 XCVR 瓦片在 12 条数模转换器 (DAC) 通道上输入这些相互同步的波形，同时在 12 条模数转换器 (ADC) 通道上采集 DUT 的输出波形。编排这一切的数据加载和检索器逻辑就是在 FPGA 的可编程逻辑结构中实现的（见图 6）。

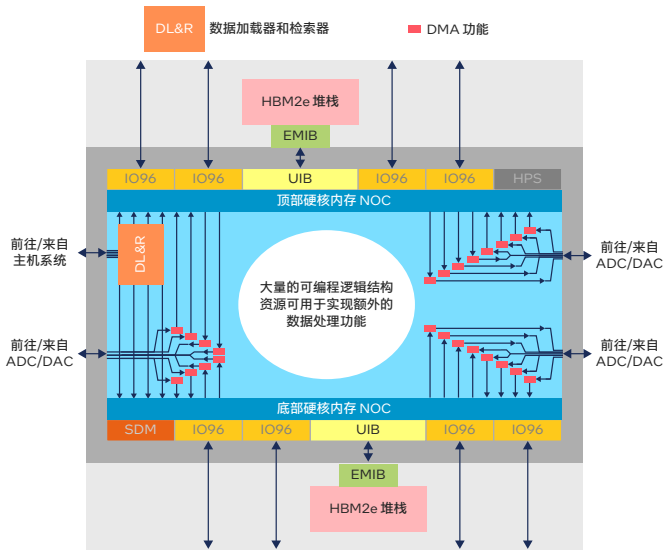

图 6. 大量的可编程逻辑结构资源仍可用于实现额外的功能，如数据缩减逻辑

12 条 DAC 和 12 条 ADC 通道分别以高达 4 Gsps（每秒千兆次采样）的速度运行。每个复杂样本的字长为 32 位 (16I + 16Q)，相当于每通道的带宽为 15.625 Gbps，因此连接 DAC 和 ADC 的瓦片的 I/O 带宽为 24 x 15.625 = 375 Gbps。如果考虑与主机系统通信的 PCIe Gen5 配置瓦片，则总 I/O 带宽为 375 + 64 = 439 Gbps。

在这样的采样率之下，会有太多数据需要存储在片上 M20K 内存模块中；否则，就要存储在封装 HBM2e 堆栈中。在 FPGA 内部，主机系统的波形通过 NoC 加载到 HBM2e 堆栈中。随后来自 HBM2e 堆栈的输出数据再通过 NoC 传送到可编程逻辑结构中的直接内存

访问（DMA）功能。采用优化的 NoC 纵向网络路由方式可以保证 Fmax 闭包，并能释放 FPGA 的常规逻辑资源用于其他任务。

除了将数据传递给驱动 DUT 的 DAC，DMA 功能还要对数据进行缓冲，这是因为自动刷新等事务会导致 DRAM 存取并不是 100% 正常。同样，从 DUT 驱动的 ADC 采集的数据也会被传送到可编程逻辑结构中的 DMA 功能。这些功能会将数据传输给 NoC，而 NoC 又将数据传输给 HBM2e 堆栈。

大量的可编程逻辑结构资源仍可用于实现数据缩减逻辑，比如大规模的快速傅里叶变换（FFT）引擎，这样就可以同时访问和处理来自顶部和底部硬核内存 NoC 的数据。

为了使该应用能高效利用 HBM2e，数据将基于伪通道（PC）作为并行数据流来处理。顶部和底部 HBM2e 堆栈上的四条伪通道保留给 PCIe Gen5 x16，从而同时读取和写入四条伪通道（将波形从主机系统传输到 HBM2e 时使用双缓冲）。在设备的每一半，将预留 6 条伪通道用于输出数据，6 条用于采集数据。在这种设计中，每条伪通道仅用作数据源或数据接收，从而避免争用。

英特尔® Agilex™ M 系列 FPGA 中的 NoC 有诸多优势，包括数据源和目的地之间的关联可以动态改变（每个 NoC 可配置为 16 x 16 交换机），因此任何 DAC 或 ADC 都可以与任一伪通道通信。NoC 能够实现从不同伪通道同时读取数据。如果在数据采集过程中，PCIe 也在将新数据加载到 FPGA，那么 NoC 将使输出引擎能够将新波形从不同的伪通道读取到前一组波形所用的伪通道，这意味着应用不会拘泥于 DAC 和伪通道之间的 1:1 关联。如果需要，还可以将可编程逻辑结构中用户定义的逻辑连接到 NoC 中的发起程序，来执行全双工读写。

浅水模型（SWM）

浅水模型（SWM）属于一类叫做“模板问题”（Stencil Problem）的高性能计算算法，该类算法通过将空间划分为独立单元来模拟物理现象。为了模拟随时间变化的行为，每个单元都会根据一组方程以迭代方式重新计算，这些方程取决于每个单元中一组变量的当前值，以及一组叫做 “halo” 的相邻单元中的变量值。

模板（Stencil）是一个包含当前单元及其相邻单元的结构。使用模板，我们就可以实现并行计算，这样就可以同时对大量的单元进行重新计算。

此 SWM 示例可以作为众多应用和问题的一个代表。例如，热力学问题就会使用基于模板的计算来模拟热量在半导体封装内的移动方式。天线模拟和计算流体动力学等其他应用也会使用类似的方法，将空间和时间离散化，并在每个时间步长上应用一组方程。

天气预报模型在创建和更新预报时，就会使用基于模板的复杂方法来模拟压力、温度、风速和其他变量。我们的例子将用于天气建模（这是一个三维问题，在每个点上有 10 多个变量需要计算）。我们将使用 SPECfp2000 基准测试套件中的 SWM 代码来说明 HBM2e 和 DDR5 在基于模板的计算中的性能优势。SWM 将天气模型简化成了一个具有三个变量的二维问题，这三个变量分别为：p（压力）、v（速度的垂直分量）和 u（速度的水平分量）。

在天气预报中，单元会以独立的时间步长进行多次更新，从而预测未来一两天的天气。SWM 会在内存中保留每个单元当前及以前的变量值，并根据一组更新方程在每个时间步长内更新变量值，再将新值写回内存。这一操作将会一直执行到完成对整个空间的计算。

当使用 FPGA 来加速这类计算密集型工作负载时，大多数设计人员都会从最大化数据路径并行性的角度出发。这是理所当然的，因为 FPGA 拥有大量的 DSP/算术资源，具备为操控大型数据组的算法提供大幅加速的潜力。

不过，由于英特尔® Agilex™ M 系列 FPGA 是专为应对计算瓶颈在于内存和/或 I/O 的情况而设计的，因此更好的做法可能是从用尽所有可用内存带宽的角度去探索解决方案，然后再确定支持这一特定带宽所需的数据并行性水平。

数据处理流程

我们的计算采用以下三个变量：

u – 速度的水平分量

v – 速度的垂直分量

p – 压力

在对 FPGA 加速进行优化之前，计算分为以下几个阶段：

1. 使用 p, (u, v) 来计算一组中间变量 (U, V), z, h，从而减少计算量。
2. 通过对 (U, V), z, h 施加循环边界条件，交换新计算的中间变量值。
3. 使用 (U, V), z, h 的中间值来计算新的 p, (u, v) 值。
4. 对 p, (u, v) 施加循环边界条件。
5. 通过保留旧版本的数据，对 p, (u, v) 进行时间平滑处理（实际上，我们会保留两份历史数据以方便顺序读/写）。

而我们的实施则采用了 oneAPI 来促进异构目标的无缝开发。它结合了阶段 1 和阶段 2，并合并了阶段 3、4、5，来提高内存效率，从而形成了一个由两个计算阶段（由边界数据交换分隔）组成的算法（见图 7）。

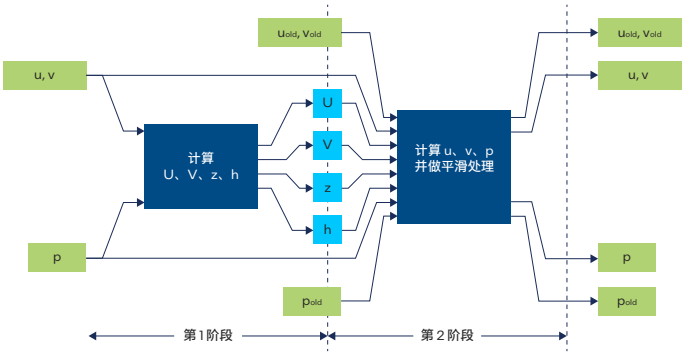

图 7. 迭代中的数据处理流程

第 1 阶段： 根据 u、v 和 p 计算出中间值 U、V、z 和 h。在第 1 阶段和第 2 阶段的边界，有一次边界条件更新，将新计算的中间值沿二维网格的边缘复制到 halo 单元中。

第 2 阶段： 读取中间值和当前值来计算新的 u、v 和 p 值。时间平滑计算使用当前、更新的和以前的变量值来确保数值稳定性。在进行时间平滑处理时，我们还需要读取旧值。

内存要求

由于这种计算受内存限制，我们将尽量减少存储在 HBM2e 和 DDR5 中的数组数量，将一些变量存储在 M20K 中从而提高性能。

在我们的方案中，M20K 用来存储 u、v、p、U、V、z 和 h 数组值。896 个 M20K 用来存储 SWM 的 3 个变量和 4 个中间变量，形成一个 64 x 64 大小的区域。这就减少了下一次迭代中第 2 阶段和第 1 阶段之间的时延，并降低了计算中的空闲时间。时间平滑处理中使用的“旧”p、u、v 数组值则存储在 DDR5/HBM2e 中。只有第 2 阶段才会使用这些“旧”的 p、u、v 值，因此 DDR/HBM 访问时延便不再那么重要。在下一次迭代的第 1 阶段进行期间，第 2 阶段的 DRAM 输出可以提交到内存中。

为了获得尽可能大的吞吐量，我们可以利用起所有可用的 32 条 HBM2e 伪通道（顶部 16 条加底部 16 条），同时利用 8 x 32 GB DDR5 内存（顶部 4 x 32 GB 加底部 4 x 32 GB）。

实现第 2 阶段的高效需要使用两份历史数据。首先，从一半的内存通道中读取 uold、vold 和 pold 进行计算，然后将新值写入另一半的内存通道中。这种纯粹的顺序存取模式使每个内存通道能够以 22.4 GB/s 的速度运行。因此，我们的设备可以达到 716 GB/s（仅 HBM2e）或 896 GB/s (HBM2e + DDR5) 的总速度（针对 -2 速度级别的 FPGA 设备）。

在双精度浮点运算中，内存带宽需匹配数据路径并行性。如果使用全部 40 条内存通道 (32 HBM2e + 8 DDR5)，我们的目标应该是达到 40 的数据并行性。即，可编程逻辑结构将同时读取每行的 40 个元素，为 40 条计算管道提供数据。我们需要合理安排数据，使所有的内存不断地进行顺序读写，从而确保 DRAM 带宽得到了充分利用。

实施

为了保持顺序访问模式，我们必须创建两个缓冲区，用来存储三个变量的“旧”值。我们将其称为 (pold, vold, uold) top，(pold, vold, uold) top'。这让我们能够从一条伪通道读取的同时写入另一条伪通道。我们将数组的每个副本分成两半，一半存储在顶部 HBM2e 堆栈中，另一半存储在底部。40 条计算管道以类似方式分布在设备结构的上半和下半部分之间，每条管道对来自最近内存的数据进行处理（见图 8）。

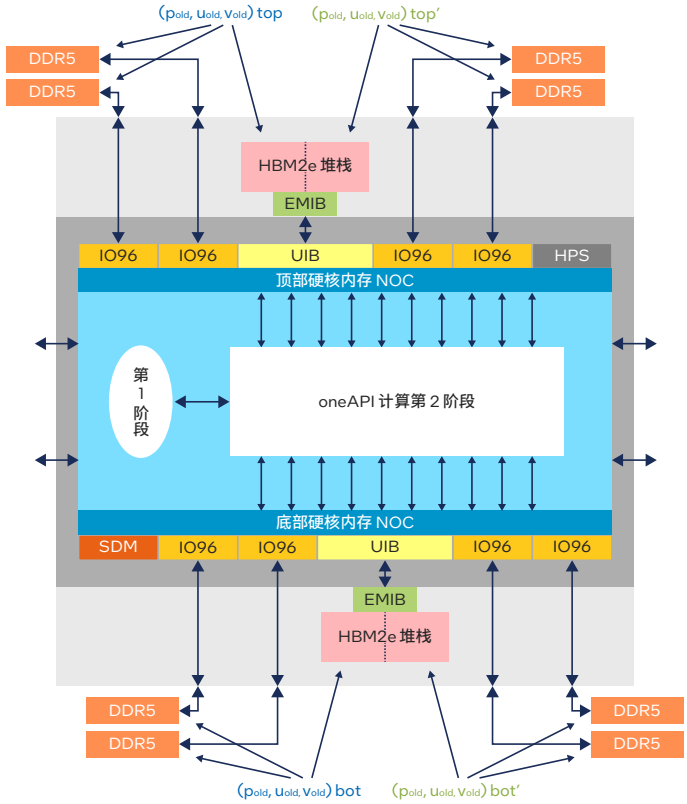

图 8. 在一次迭代中，我们从顶部通道 (top) [HBM2e+DDR5] 读取数据，然后写入 top' 通道；在下一次迭代中，我们从 top' 通道读取，然后写入 top 通道；底部也是如此

该分析基于 64 x 64 的计算区域。它的表面积与体积之比（也就是通信与计算之比）在许多并行算法中具有代表性。我们的数组是 64 x 64 的。数组中的每个元素都包含三个双精度浮点值。我们需要在多条伪通道上对数组进行剥离，将第一个 64 字节放入第一条伪通道，下一个 64 字节放入下一条伪通道，以此类推，以便实现从逻辑地址到实际地址的良好映射。这需要用到全部 16 条伪通道，以及每个顶部和底部 HBM2e 阵列上的四条 DDR5 通道。我们使用 NoC 的 10 个发起程序访问每边的 20 条通道。由于我们的发起程序是全双工的，它们可以同时对一对伪通道进行读取和写入。这使我们能够在不耗尽发起程序的情况下有效地利用全部内存带宽。

该算法的每一次迭代都会对每个单元进行 65 次浮点运算，其中包括：25 次乘法、39 次加法和一次除法。在双精度实施下，40 条管道需要消耗英特尔® Agilex™ M 系列 FPGA 的 3,360 个 DSP 模块（共 12,500 个）。由于这两个计算阶段并不同时进行，因此我们预测，基于 500 MHz 的结构频率，应用可实现高达 650 GFLOPS 的双精度浮点运算性能水平。

结论

如今的计算工作负载比过去规模更大、更复杂、更多样化。有些应用需要对大量数据进行流式传输，而有些则需要能应对短时随机的数据量激增。同样，有些算法在访问内存时需要尽可能低的时延，而有些对时延的包容性则更大。

为了满足这些应用的严苛要求，英特尔开发了英特尔® Agilex™ M 系列 FPGA。这是第一款基于英特尔® 7 制程工艺，并且支持封装 HBM2e 内存的英特尔® Agilex™ FPGA。英特尔® Agilex™ M 系列 FPGA 还包括面向其他先进内存技术（如 DDR5 和 LPDDR5）的硬核控制器。硬核内存 NoC 功能使 FPGA 可编程逻辑结构能够提供对封装内的 HBM2e 和封装外的（板载）内存资源的高带宽、资源高效型访问。

要应对网络、数据中心和边缘领域的艰巨挑战，就需要将高算力、高内存和高 I/O 带宽相结合。英特尔® Agilex™ M 系列 FPGA 提供出色的 INT8 TOPS 和 FP32 TFLOPS。其总内存带宽也相当优秀：在两个 HBM2e 堆栈和八个 DDR5 接口全部启用时，可提供超过 1 Tbps 的超高总内存带宽。此外，英特尔® Agilex™ M 系列 FPGA 为下一代 800 G/1.6 T 网络和网络功能虚拟化基础设施（NFVI）应用在每个方向上都提供了超过 2.65 Tbps 的总串行收发器带宽（超过 5.3 Tbps 全双工）。

所有开发人员都可以利用英特尔® Agilex™ M 系列 FPGA 的强大功能。硬件设计工程师可使用英特尔® Quartus® Prime 软件设计工具，而软件开发人员则可使用 oneAPI（一套核心工具和库）针对不同架构开发以数据为中心的高性能应用。

intel®