全新x86内核架构、XeSS神经网络超采样、千亿晶体管SoC，这次英特尔诚意满满

https://mp.weixin.qq.com/s/60n7AEI-mzJufG5YreIxOw

None Fri Aug 20 09:13

机器之心报道

机器之心编辑部

2021 年英特尔架构日上，英特尔发布一系列重大技术架构的改变和创新：两款全新的 x86 内核架构（能效核与性能核）、代号为 Alder Lake 的首个性能混合架构、英特尔硬件线程调度器、专为数据中心设计的下一代英特尔至强可扩展处理器 Sapphire Rapids 以及基础设施处理器（IPU）等内容。

8 月 20 日，2021 年英特尔架构日如约而至！
会上，英特尔公司高级副总裁兼加速计算系统和图形事业部总经理 Raja Koduri 携手多位英特尔架构师，全面介绍了英特尔在 CPU、GPU 及 IPU 架构方面的重大进展。

英特尔公布了一系列重大的改变与创新，总结如下：

- 首个性能混合架构 Alder Lake，以及其搭载的两款全新的 x86 内核框架——能效核和性能核，以及英特尔硬件线程调度器；
- 全新的独立游戏图形处理器（GPU）架构 X^e HPG 微架构，这将是英特尔首款基于台积电 N6 工艺的 GPU；
- 专为数据中心打造的下一代英特尔至强可扩展处理器 Sapphire Rapids；
- 基于 X^e HPC 微架构的数据中心 GPU 架构 Ponte Vecchio；
- 全新的基础设施处理器（IPU）；
- oneAPI 工具包：提供一个开放、规范、跨架构和跨厂商的统一软件栈，让开发者能够摆脱专有语言和编程模型的束缚。

两款全新 X86 内核

能效核（Efficient Core）

能效核曾用代号「Gracemont」，是一个高能效的 x86 微架构，旨在面向当今多任务场景，提高吞吐量效率并提供可扩展多线程性能。
该架构致力于通过低电压能效核降低整体功率消耗，为更高频率运行提供功率热空间。同时能效核的性能也得到了提升，将能满足更多动态任务负载。

能效核可以在不耗费处理器功率的情况下对工作负载进行优先级排序，并通过每周期指令数（IPC）改进一些功能以提高性能，包括：

- 拥有 5000 个条目的分支目标缓存区，实现更准确的分支预测；
- 64KB 指令缓存，在不耗费内存子系统功率的情况下保存可用指令；
- 英特尔的首款按需指令长度解码器，可生成预解码信息；
- 英特尔的簇乱序执行解码器，可在保持能效的同时，每周期解码多达 6 条指令；
- 后端宽度（Wide Back End）具备 5 组宽度分配（Five-wide allocation）和 8 组宽度引退、256 个乱序窗口入口和 17 个执行端口；
- 支持英特尔控制流强制技术和英特尔虚拟化技术重定向保护等功能；
- 实现了 AVX 指令集以及支持整数人工智能操作的新扩展。

相比于英特尔最多产的 CPU 内核 Skylake，在单线程性能下，能效核能够在相同的功耗下实现 40% 的性能提升，或在功耗不到 40% 的情况下提供同等性能。与运行四个线程的两个 Skylake 内核相比，四个能效核所提供的吞吐量性能，能够在功耗更低的情况下带来 80% 的性能提升，而在提供相同吞吐量性能的情况下，功耗将减少 80%。

性能核（Performance Core）

工作负载的代码体积正在不断增长，需要更强的执行能力。数据集也随着数据带宽的需求提升而大幅增加。英特尔全新性能核微架构旨在提高速度，突破低时延和单线程应用程序性能的限制，曾用代号「Golden Cove」。该架构在带来显著增速的同时，能够更好地支持代码体积较大的应用程序。

性能核是英特尔有史以来构建的性能最高的 CPU 内核，并通过以下功能突破了低时延和单线程应用程序性能的极限：

- 相比于目前的第 11 代英特尔酷睿处理器架构（Cypress Cove），在通用性能的 ISO 频率下，针对大范围的工作负载实现了平均约 19% 的改进；
- 呈现出更高的并行性；
- 搭载英特尔高级矩形扩展（AMX），内置下一代 AI 加速提升技术，用于学习推理和训练。AMX 包括专用硬件和新指令集架构，以显著加速矩阵乘法运算；
- 减少时延，对大型数据和代码体积较大的应用程序提供更好的支持。

英特尔硬件线程调度器（Intel Thread Director）

为了让性能核和能效核与操作系统无缝协作，**英特尔开发了一种改进版的调度技术，称为「英特尔硬件线程调度器（Intel Thread Director）」**。

该调度器直接内置于硬件中，可提供对内核状态和线程指令混合比的低级遥测，让操作系统能够在恰当的时间将合适的线程放置在合适的内核上。硬件线程调度器具有动态性和自适应性——它会根据实时的计算需求调整调度决策——而非一种简单的、基于规则的静态方法。
目前，英特尔正在优化 Thread Director，以在微软即将推出的 Windows 11 操作系统上获得最佳性能。

基于 Intel 7 制程的首个混合架构 Alder Lake

英特尔公布了**首款集成能效核和性能核、并搭载全新英特尔硬件线程调度器的混合架构「Alder Lake」，它基于 Intel 7 制程工艺打造而成，并采用了单一、高度可扩展的 SoC 架构，支持最新内存和最快 I/O**。该架构适用于从移动端到台式机的产品，并且基于 Alder Lake 的产品将在今年开始出货。

Alder Lake 将提供惊人的性能，支持**从超便携式笔记本到发烧级再到商用台式机的所有客户端设备**，它提供了以下三类产品设计形态：

- 高性能、双芯片、插座式的台式机处理器，具有领先性能和能效。支持高规格的内存和 I/O；
- 高性能笔记本处理器，采用 BGA 封装，并加入图像单元，更大的 X^e 显卡和 Thunderbolt 4 连接；
- 轻薄、低功耗的笔记本处理器，采用高密度的封装，配置优化的 I/O 和电能传输。

构建如此高度可扩展架构的挑战，需要在不影响功率的情况下满足计算和 I/O 代理对带宽超乎寻常的需求。为了解决这一挑战，英特尔设计了**三种独立的内部总线，每一种都采用基于需求的实时启发式后处理方式**：

- 计算内部总线可支持高达 1000GBps——即每个内核或每集群 100GBps，通过最后一级缓存将内核和显卡连接到内存；
- I/O 内部总线支持可高达 64GBps，连接不同类型的 I/O 和内部设备，能在不干扰设备正常运行的情况下无缝改变速度，选择内部总线速度来匹配所需的数据传输量；
- 内存结构可提供高达 204GBps 的数据，并动态扩展其总线宽度和速度，以支持高带宽、低时延或低功耗的多个操作点。

英特尔还公布了台式机处理器的 I/O 信息，从下图可以看到，**拥有最高 8 个性能核、8 个能效核、24 线程以及 30M 的 Non-inclusive LL 缓存**。

独立游戏显卡微架构 X^eHPG 及 Alchemist 系列 SoC

X^e HPG 是一款全新的独立显卡微架构，专为游戏和创作工作负载提供发烧级的高性能。**X^e HPG 微架构为 Alchemist 系列 SoC 提供动力**，首批相关产品将于 2022 年第一季度上市，并采用新的品牌名——英特尔锐炫 Arc。X^e HPG 微架构采用全新的 X^e 内核，是一款聚焦计算、可编程且可扩展的元件。
英特尔公布了基于 X^e HPG 架构的客户端显卡路线图——Alchemist（此前称之为 DG2）、Battlemage、Celestial 和 Druid SoC。

其中，基于 X^e HPG 微架构的 Alchemist SoC 产品能够提供出色的可扩展性和计算效率，并拥有以下关键架构特征：

- 使用台积电的 N6 制程节点上进行制造；
- 多达 8 个具有固定功能的渲染切片，专为 DirectX 12 Ultimate 设计；
- 全新 X^e 内核，拥有 16 个矢量引擎和 16 个矩阵引擎（被称为 XMX，即 X^e Matrix eXtension）、高速缓存和共享内部显存；
- 支持 DirectX Raytracing（DXR）和 Vulkan Ray Tracing 的新光线追踪单元；
- 通过架构、逻辑设计、电路设计、制程工艺技术和软件优化，相比 X^e LP 微架构实现 1.5 倍的频率提升和 1.5 倍的每瓦性能提升。

英特尔分享了试产阶段 Alchemist SoC 的真实游戏展示（虚幻引擎 5 测试良好），以及全新的基于神经网络的超取样技术 X^e SS 等。
X^ e SS是一种利用Alchemist的内置XMX AI加速、实现高性能和高保真视觉的全新升频技术。该技术使用深度学习来合成非常接近原生高分辨率渲染质量的图像。

目前，多家早期的游戏开发商已开始使用X^SS， 本月将向独立软件供应商（ISV）提供XMX初始版本的SDK，DP4a版本将于今年晚些时候推出。

专为数据中心打造的下一代英特尔至强可扩展处理器 Sapphire Rapids
Sapphire Rapids 处理器基于 Intel 7 制程工艺技术，采用全新的性能核微架构，旨在提高速度 / 突破低时延和单线程应用性能的极限。

Sapphire Rapids 的核心是一个分区块、模块化的 SoC 架构，采用英特尔的嵌入式多芯片互连桥接（EMIB）封装技术，在保持单晶片 CPU 接口优势的同时，具有显著的可扩展性。
Sapphire Rapids 提供了一个单一、平衡的统一内存访问架构，每个线程均可完全访问缓存、内存和 I/O 等所有单元上的全部资源，由此实现整个 SoC 具有一致的低时延和高横向带宽。该处理器的主要构建块如下图所示：

此外，Sapphire Rapids 提供业界广泛的数据中心相关加速器，包括新的指令集架构和集成 IP，以在各种客户工作负载和使用中提升性能。通过无缝集成的加速器引擎赋能常见模式任务的卸载，提升内核效率。

新的内置加速器包括：

- 英特尔加速器接口架构指令集（AIA）——支持对加速器和设备的有效调度、同步和信号传递；

- 英特尔高级矩阵扩展（AMX）——Sapphire Rapids 中引入的新加速引擎，可为深度学习算法核心的 Tensor 处理提供大幅加速。其可以在每个周期内进行 2000 次 INT8 运算和 1000 次 BFP16 运算，实现计算能力的大幅提升；

英特尔数据流加速器（DSA）——旨在卸载导致数据中心规模部署开销的最常见数据移动任务。DSA 改进了对这些开销任务的处理，提供了更高的整体工作负载性能，并可以在 CPU、内存和缓存以及所有附加的内存、存储和网络设备之间移动数据。

总之，这些架构上的改进使得 Sapphire Rapids 能够为云、数据中心、网络和智能边缘中广泛的工作负载和部署模式提供开箱即用的性能。

英特尔迄今最复杂、千亿晶体管的 SoC

英特尔发布了**迄今为止最复杂的 SoC Ponte Vecchio，它拥有 1000 多亿个晶体管**，提供业界领先的每秒浮点运算次数和计算密度，以加速 AI、HPC 和高级分析工作负载。

据悉，Ponte Vecchio 已走下生产线进行上电验证，并已开始向客户提供限量样品。Ponte Vecchio 预计将于 2022 年面向 HPC 和 AI 市场发布。

英特尔称，早期的 Ponte Vecchio 芯片展示了领先的性能，在流行的 AI 基准测试中创造了推理和训练吞吐量的行业记录。其中，英特尔 A0 芯片性能提供高于 45 TFLOPS 的 FP32 吞吐量、高于 5 TBps 的内存结构带宽，以及高于 2 TBps 的连接带宽。

Ponte Vecchio 基于 X^e-HPC 微架构，由多个复杂的设计组成，这些设计以单元形式呈现，然后通过嵌入式多芯片互连桥接（EMIB）单元进行组装，实现单元之间的低功耗、高速连接。这些设计均被集成于 Foveros 封装中，为提高功率和互连密度形成有源芯片的 3D 堆叠。高速 MDFI 互连允许 1 到 2 个堆栈的扩展。

Ponte Vecchio 的**核心是计算单元（Compute Tile）——一个密集的多 X^e 内核**。计算单元基于台积电先进的 N5 制程工艺，包含 8 个 X^e 内核和 4MB 一级缓存。此外，计算单元具有极其紧凑的 36 微米凸点间距，可与 Foveros 进行 3D 堆叠。英特尔也已经通过设计基础设施设置和工具流程以及方法，为测试和验证该节点的单元铺平了道路。

基础单元是 Ponte Vecchio 的连接组织。它是基于 Intel 7 制程工艺的大型芯片，针对 Foveros 技术进行了优化。

最后是 X^e 链路单元（X^e Link Tile），它提供了 GPU 之间的连接，支持每单元 8 个链路。该单元对 HPC 和 AI 计算的扩展至关重要，旨在旨在实现支持高达 90G 的更高速 SerDes，并且已被添加到「极光」（Aurora）百亿亿次级超级计算机的扩展解决方案中。

全新基础设施处理器（IPU）

IPU 设计旨在使云和通信服务提供商减少在中央处理器（CPU）方面的开销，并充分释放性能价值。Mount Evans 是英特尔的第一个 ASIC IPU，旨在解决多样化和分散的数据中心的复杂性。Oak Springs Canyon 是一个 IPU 参考平台，采用 Intel Xeon D 处理器和 Intel Agilex FPGA。Intel N6000 加速开发平台专为基于 Xeon 的服务器设计。

Mount Evans。

Oak Springs Canyon。

参考链接：

https://mp.weixin.qq.com/s/2i2SDY9jD-TVFtemX9GlzQ

https://www.intel.com/content/www/us/en/newsroom/resources/press-kit-architecture-day-2021.html

NVIDIA对话式AI开发工具NeMo实战分享

开源工具包 NeMo 是一个集成自动语音识别（ASR）、自然语言处理（NLP）和语音合成（TTS）的对话式 AI 工具包，便于开发者开箱即用，仅用几行代码便可以方便快速的完成对话式 AI 场景中的相关任务。

8月26日20:00-21:00，系列分享**第2期：使用NeMo快速构建智能问答系统**。

- 智能问答系统简介
- 智能问答系统的工作流程和原理
- 构建适合于NeMo的中文问答数据集
- 在NeMo中训练中文问答系统模型
- 使用模型进行推理完成中文智能问答的任务

直播链接：https://jmq.h5.xeknow.com/s/how4w（点击阅读原文直达）

报名方式：进入直播间——移动端点击底部「观看直播」、PC端点击「立即学习」——填写报名表单后即可进入直播间观看。

交流答疑群：直播间详情页扫码即可加入。

© THE END