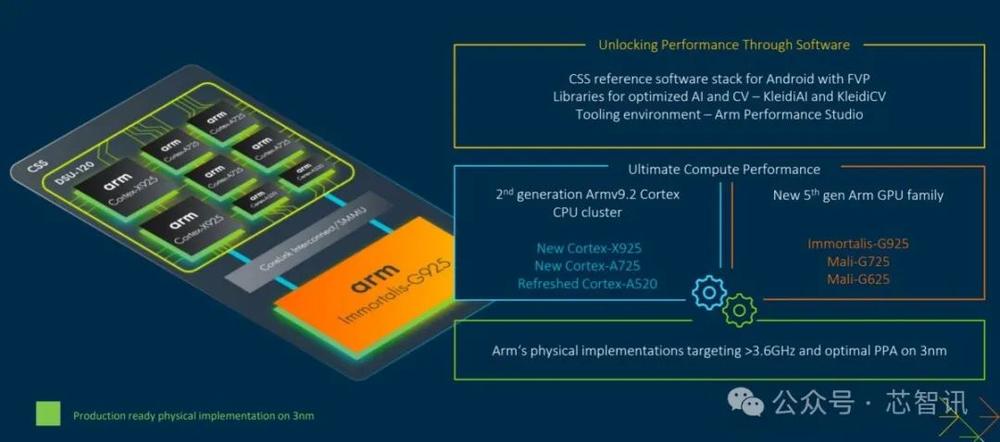


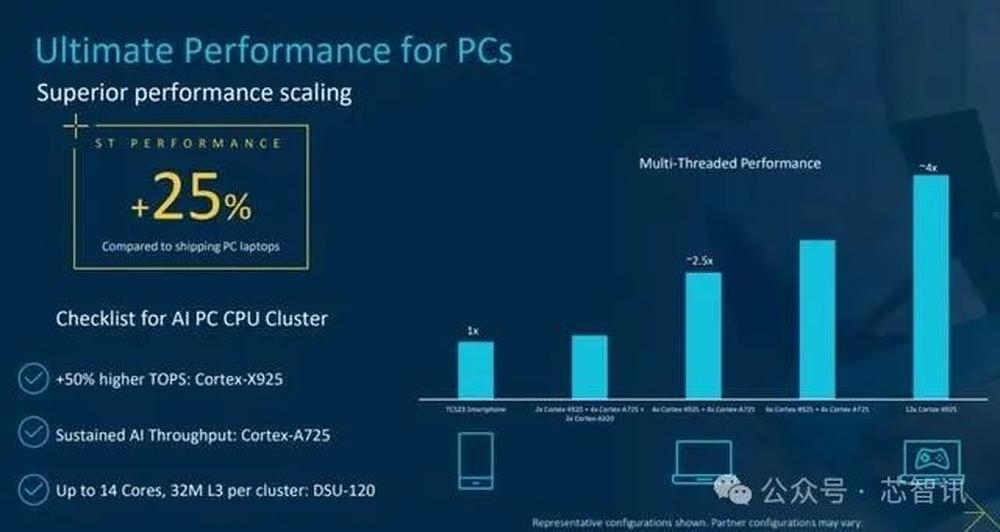
5月30日消息，当地时间周三，Arm在其全面计算解决方案（CSS）取得成功的基础上，正式发布了首款面向客户端产品的 Arm 计算子系统 ——CSS for Client，以及新的 Arm Kleidi 软件，大大简化了运行 Android、Linux 和 Windows 的台式机、笔记本电脑、平板电脑处理器的开发和人工智能（AI）的部署。  
  
  
CSS for Client包括最新的Armv9.2指令集的CPU集群，即最高性能的Cortex-X925 CPU、最高效的Cortex-A725 CPU和更新后的Arm Cortex-A520 CPU；Arm Immortalis GPU和 Mali GPU；CoreLink 互连系统 IP；以及用于领先代工厂 3nm 工艺的 CPU 和 GPU 的生产就绪物理实现。



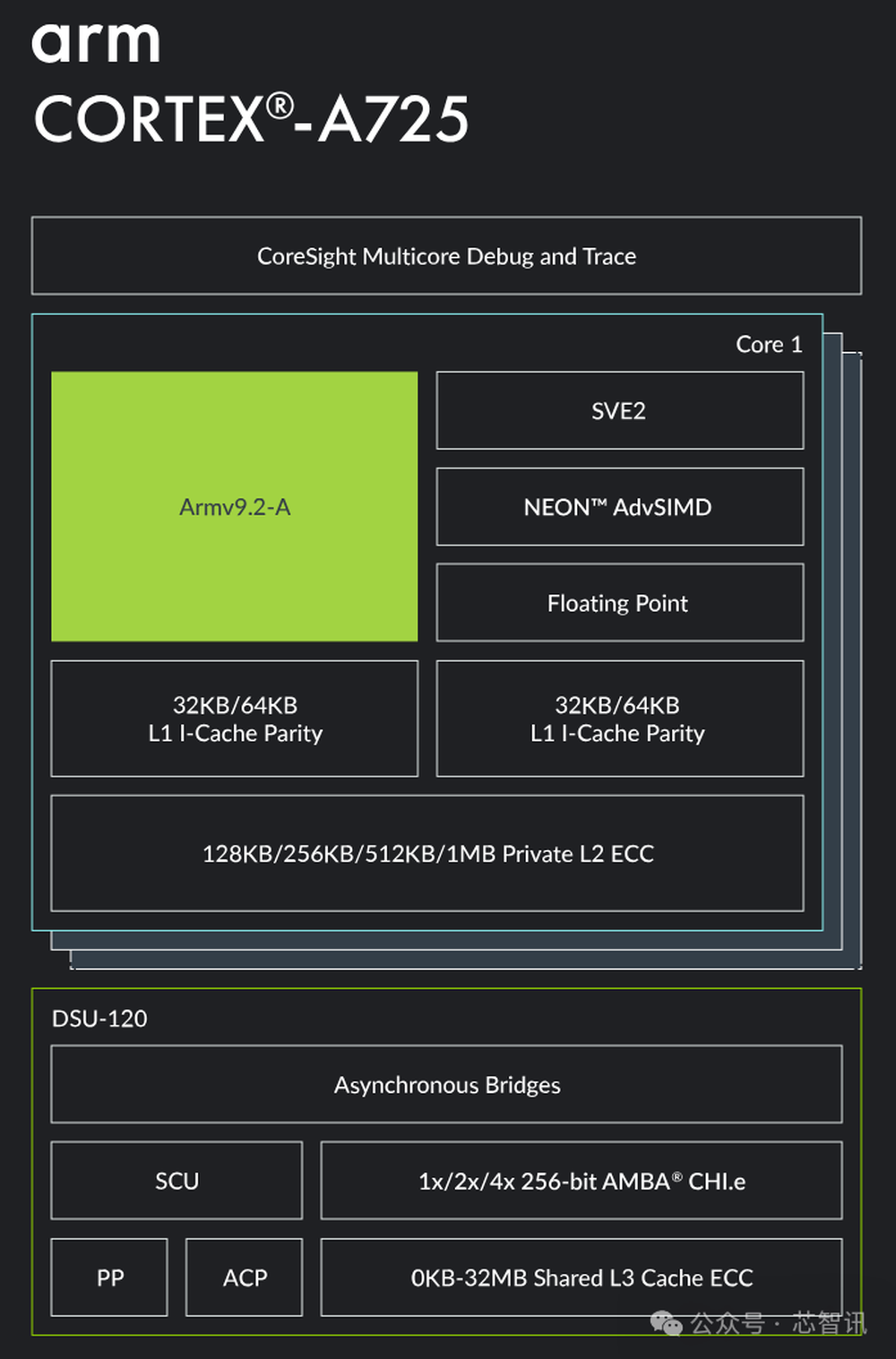
Arm表示，CSS for Client 是 Arm 迄今为止最快的 Android 平台，与 TCS23 平台相比，在关键基准测试和通用计算用例方面有了显著改进。同时，CSS for Client也是 AI 驱动体验的计算基础，可在最广泛的消费类设备中实现性能、效率和可扩展性的阶跃式变化。  
  
  
“该平台为我们的合作伙伴提供了最快的芯片生产途径。这些物理实现使 Arm 的合作伙伴能够释放领先 3nm 工艺的所有优势，同时实现高度灵活、可定制的硅设计。”Arm在其官方博客中写道。  
  
  
Cortex-X925 CPU：单线程性能提升36%，AI算力提升50%  
  
  
Arm台湾总裁曾志光指出，Cortex-X925的前一代是Cortex-X4，由于IPC（每周期指令）史上大幅提升，因此更改名称来凸显与前代产品的差异。但下一代名称是否从这个编号继续沿用下去，要等到明年新产品揭晓。  
  
  
Cortex-X925是Arm迄今为止最为强大的CPU内核，基于最新的Armv9.2指令集，可以支持支持 SVE、SVE2指令，主频最高可达3.8GHz（常规为3.6GHz），并且通过DSU-120（DynamIQ共享单元）可以实现最高14核集群，适用于从旗舰智能手机到笔记本电脑的设备。  
  
  
具体到性能方面，Arm终端产品事业部产品管理副总经理James McNiven表示，Cortex-X925这次带来了两方面的性能提升：一是在Geekbench中，单线程性能提升了36%，达到了历史上最大的同比性能提升；二是相对于上一代的Cortex-X4 CPU，Cortex-X925的AI性能大幅提升，基于Tiny-Llama模型的首个token生成时间，缩短了41%，可大幅改善大型语言模型的初始响应速度。



McNiven指出，Cortex-X925之所以能够有如此大的性能提升，主要是因为Arm对微构架进行重大改变，使AI算力TOPS增加了50%，并采用更大的3MB的私有L2缓存（也可以配置为2MB），同时通过RTL和实体设计团队间的合作，达成在3nm制程的最优化。



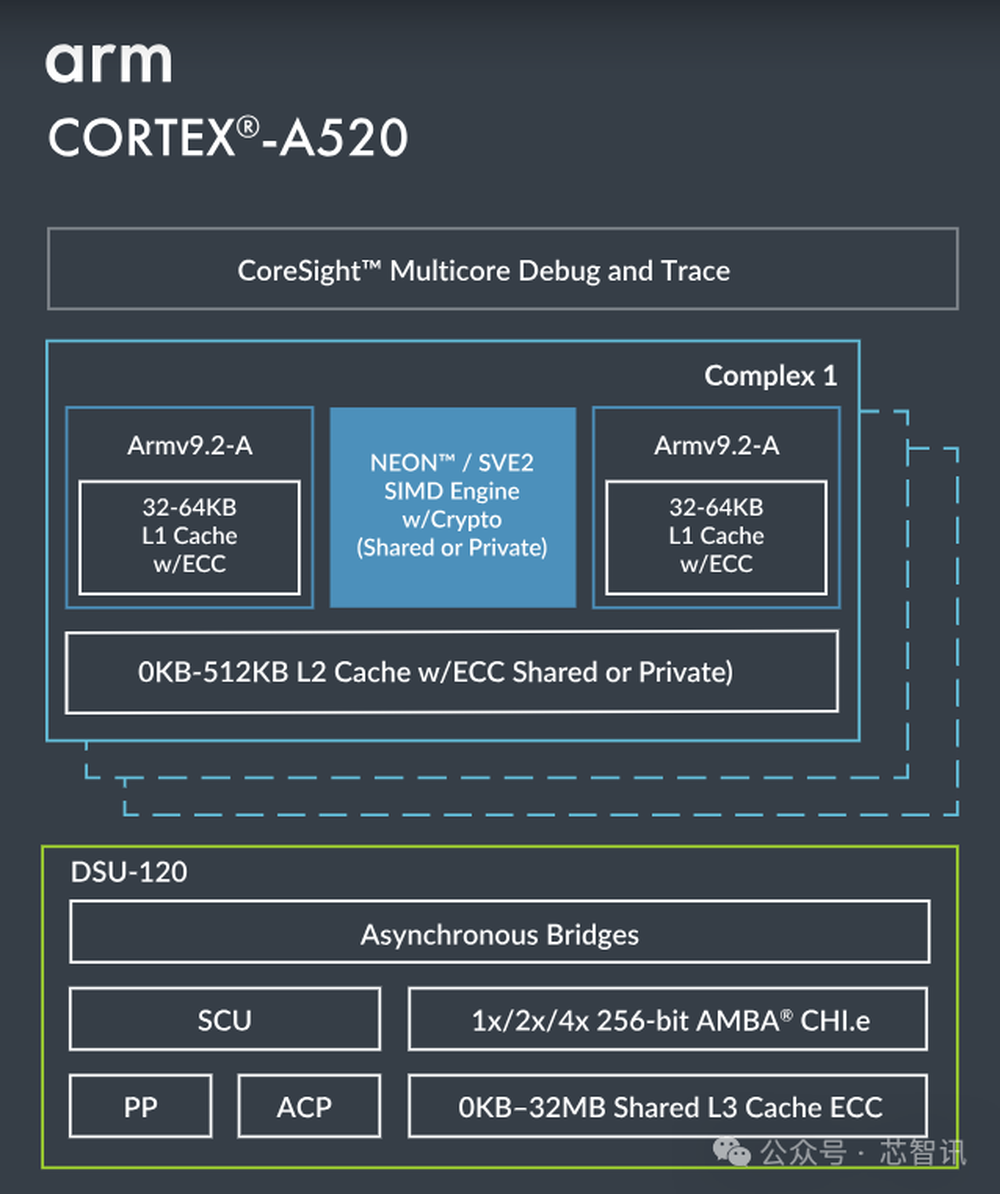
Arm称，通过DSU-120实现的最高14核由于Cortex-X925集群，可以提供“个人电脑的终极性能”，相比当前已上市的笔记本电脑性能高出25%，将助力Windows on Arm笔记本电脑。  
  
  
Cortex-A725：性能提升12%，能效提升25%  
  
  
Cortex-A725 也是一款超高效的基于 Armv9.2 指令集架构的CPU，可在受限的功率包络内提供业界领先的持续性能。这款 AArch64 位 CPU 配备了更新的 DSU-120（DynamIQ共享单元），可以与 Cortex-X925 和 Cortex-A520 结合使用。LITTLE 配置。



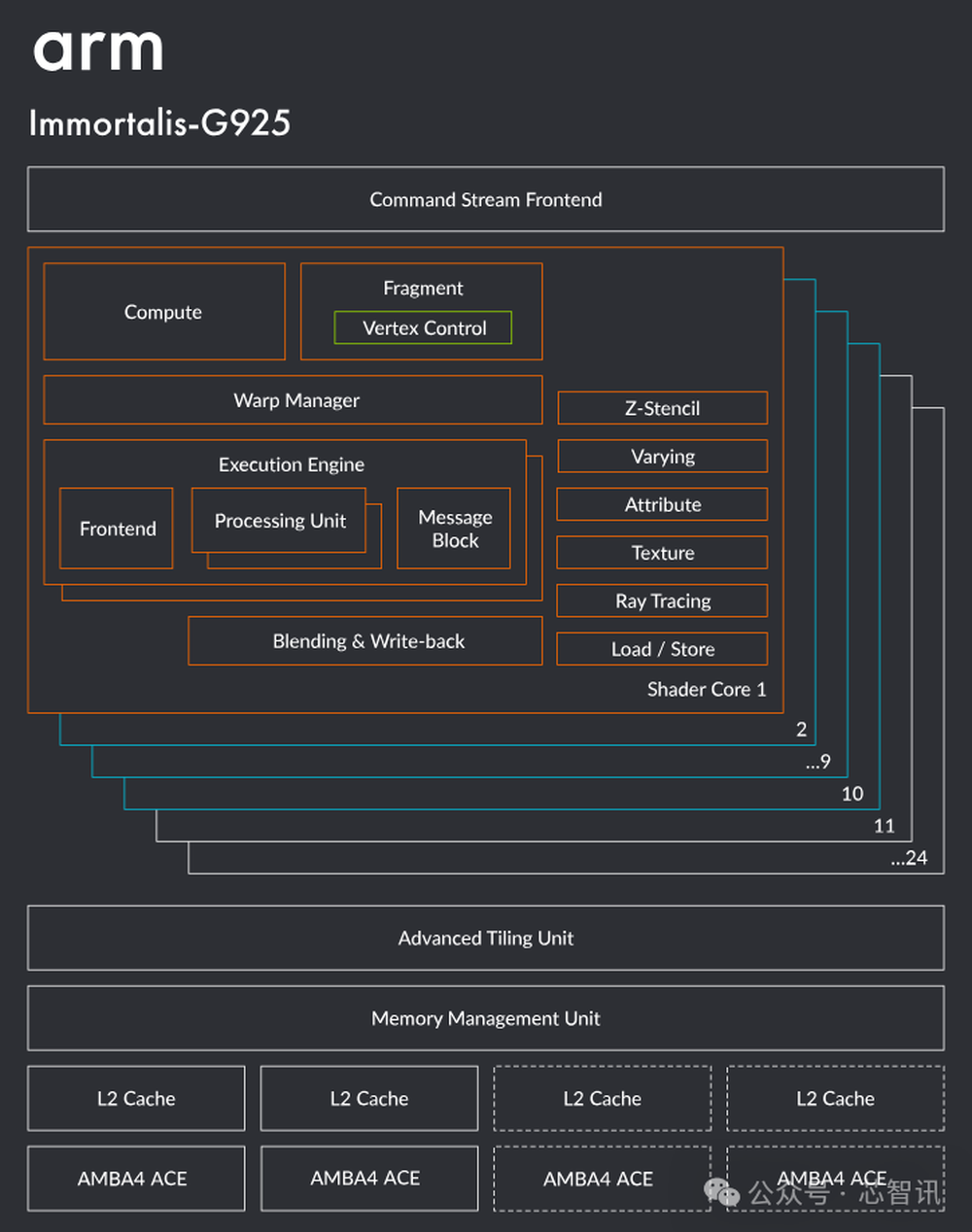
与 Arm Cortex-A720 相比，Cortex-A725的微架构改进可提供 25% 的效率和 12% 的性能提升。为成本受限的市场量身定制的配置选项，在相同的硅面积范围内，与 Armv8.2 的Cortex-A720相比，性能提高了两位数。  
  
  
另外，Cortex-A725的L2 缓存大小增加了一倍，再加上更新的 Arm DSU-120 的功耗改进，可大幅节省功耗，从而提高效率。



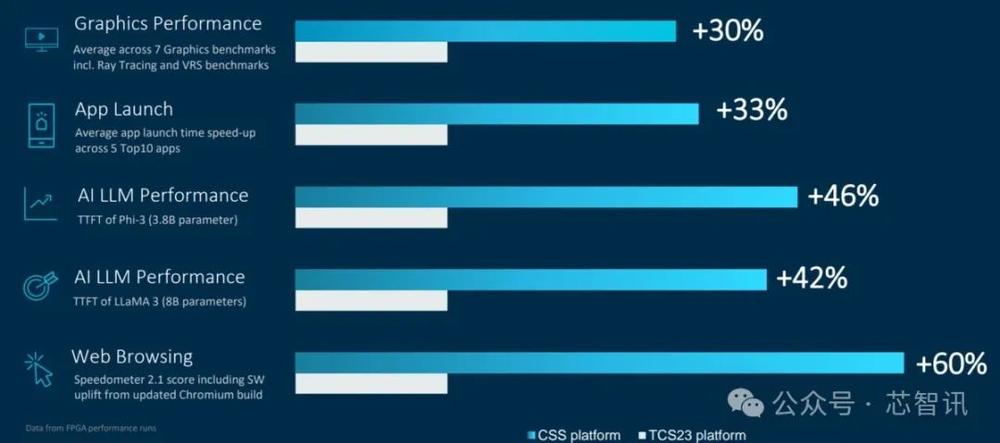
总体来看，Cortex-A725在性能效率上比上一代A720提高了35%。  
  
  
新版Cortex-A520 CPU：功耗可降低22%，性能提升8%  
  
  
Cortex-A520是Arm去年发布的2023全面计算解决方案（TCS23）当中的一款全新64位CPU，此次，Arm对其进行了进一步的更新，升级为最新的Armv9.2 架构，并带来了新的 QARMA3 PAC 算法，降低了性能成本，加强了 PAC 在消费技术市场的部署，可提供更高的电源效率，并针对后台和轻量级工作负载进行调整，以延长电池寿命。这与新的 DSU-120 一起提供，并与 Cortex-X925 和 Cortex-A725 兼容，用于构建大型的CPU集群。



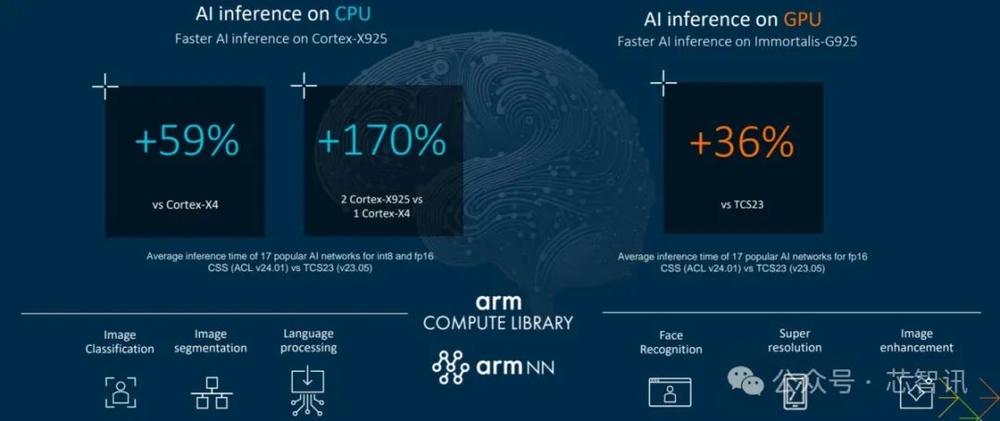
性能方面，与 Cortex-A510 相比，Cortex-A520的功耗可降低高达 22%，性能可提升8%，可跨多个细分市场进行扩展。如果Cortex-A520 采用先进的3nm制程，可将能效提高多达15%。  
  
  
Immortalis-G925  
  
  
Arm Immortalis-G925 是专为旗舰智能手机设计的最新旗舰 GPU，基于第 5 代 Arm GPU 架构构建。Immortalis-G925 提供片段预传递和双平铺和移位转换单元吞吐量，从而可以实现更好、更持久的帧速率，实现功能丰富和更长的游戏时间。这也是 Arm 迄今为止性能最高、效率最高的 GPU。



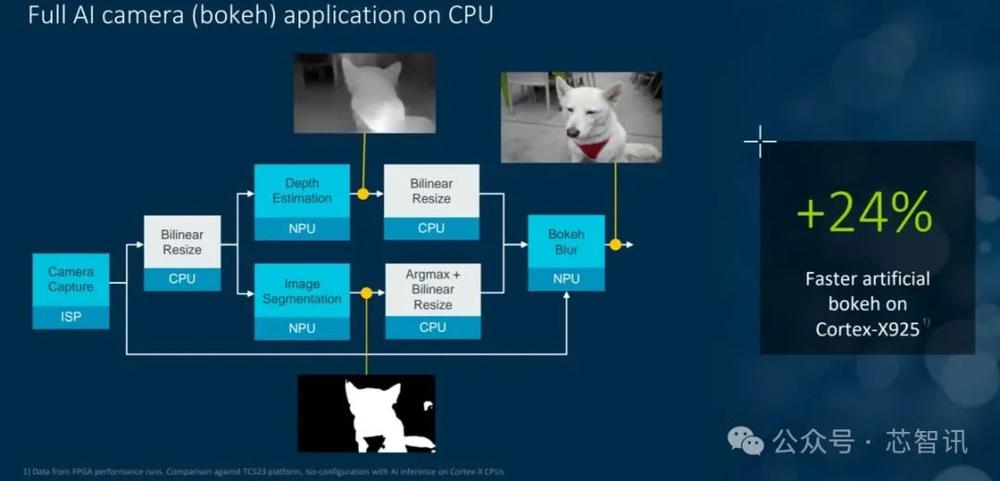
与Immortalis G720相比，Immortalis G925在图形性能方面提升了37%，而相同性能下功耗可降低30%，在一系列流行的手机游戏中支持以平均每秒 120 帧的帧率运行。面对复杂对象的光线追踪性能提升52%。  
  
  
在AI和机器学习工作负载方面，Immortalis G925性能平均提升了34%。比如，AI推理速度比当前GPU要快36%，自然语言处理速度最高可提升50%，图像分割速度最高可提升41%，语音转文本速度最高可提升32%。  
  
  
有分析称，Arm对GPU设计的升级“虽小但很关键”，看起来强大得多的一个重要原因是增加了GPU的最大着色器核心（shader core）数量，G720提供10到16个着色器核心，而全新的Immortal G925提供10到24个着色器核心，这将提供游戏内的性能与效率并在渲染线程上显著释放CPU。  
  
  
CSS for Client所带来的整体提升  
  
  
首先，CSS for Client 的系统集成和扩展是通过最新的 CoreLink Interconnect 实现的。集成的系统级缓存 （SLC） 通过减少 DRAM 带宽和访问来实现最佳系统能效。系统内存管理单元 （SMMU） 通过阶段 2 转换提供增强的安全性，以支持虚拟化安全框架，例如 Android 虚拟化框架 （AVF）。  
  
  
CSS for Client 也将成为 3nm 工艺节点上第一代 Android SoC 的一部分，从而在芯片中实现一流的 PPA（功耗、性能和面积）。作为 CSS for Client 的一部分，Arm 的物理实现释放了 3nm 技术的全部潜力，最大限度地提高了高端平台的 PPA 优势，并为我们的合作伙伴创造了最快的芯片之路。  
  
  
Arm表示，其与领先的代工合作伙伴合作，共同设计和交付 CPU 和 GPU 物理实现，其中包括用于 3nm 的流片就绪 Cortex-X925 CPU 和 Immortalis-G925 物理实现。这有助于合作伙伴在 3nm 工艺上获得全部 PPA 优势，同时通过生产就绪型芯片解决方案缩短芯片开发和部署时间。它还使我们的合作伙伴能够灵活地使用 CSS for Client 构建特定于市场的差异化 CPU 集群和 GPU。  
  
  
可以说，CSS for Client 是 Arm 迄今为止最快的 Android 平台，与 TCS23 平台相比，在关键基准测试和通用计算用例方面有了显著改进。这些包括：  
  
  
得益于新的 Cortex-X925，峰值性能提高了 36%，以 Geekbench 6 单核分数衡量;  
  
  
在前 10 个应用程序中，有 5 个应用程序的应用程序启动时间平均缩短了 33%，从而提高了工作效率，并在移动设备上提供了流畅的用户体验;  
  
  
使用 Speedometer 2.1 浏览器基准测试测得网页浏览速度提高 60%；  
  
  
得益于新的Immortalis-G925 GPU，在七个图形基准测试中，峰值图形性能平均提高了 30%，这些测试包括了光线追踪和可变速率着色 （VRS） 基准测试。



在AI性能方面，今年早些时候，Arm展示了大型语言模型 （LLM） 如何在移动设备上的 Arm CPU 上本地运行。借助 CSS for Client，LLM 将在 Arm CPU 上运行得更好，响应时间更快。目前，Cortex-X925 CPU 在运行 Llama 3 LLM 时将首次令牌的时间缩短了 42%，在运行 Phi-3 LLM 时将首次令牌的时间缩短了 46%。  
  
  
此外，由于新 Arm CPU 和 GPU 的进步，CSS for Client 还实现了在广泛的通用 AI 网络上实现 AI 推理的显着性能飞跃。这包括在 Cortex-X925 上将推理速度提高 59%，在 Immortalis-G925 上将 AI 推理速度提高 36%。通过在 CSS for Client CPU 集群配置中利用额外的 Cortex-X925 CPU，Arm观察到 int8 和 fp16 数据类型的 17 种流行网络的 AI 推理性能提升了 2.7 倍。AI 推理的这些改进可在一系列 AI 用例中实现无缝的用户体验。



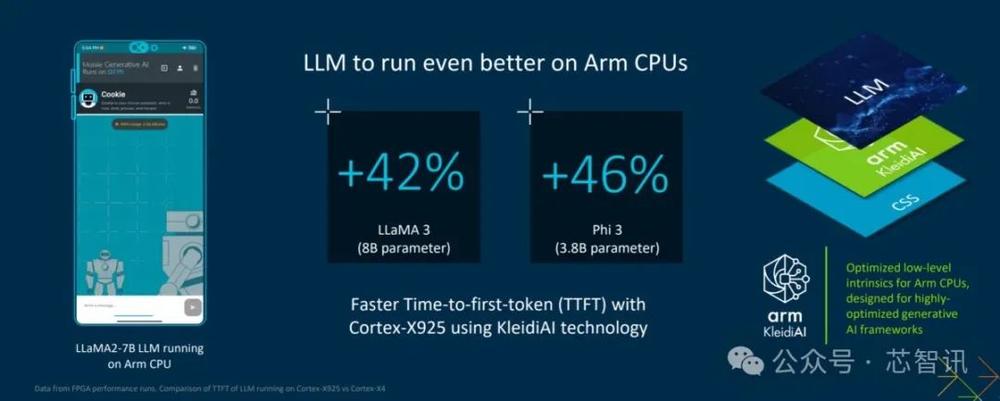
CSS for Client 特别突出的这些 AI 用例之一，是计算摄影和 AI 相机。能够拍摄具有逼真的散景效果的令人惊叹的照片和视频，这些效果会模糊背景并专注于选定的主题，这很复杂。AI相机散景管线由多个阶段组成，如深度估计、分割、遮罩和混合，以产生高质量的结果。与 TCS23 相比，CSS for Client 通过在 CPU 上进行散景工作负载的 AI 处理，将散景性能提高了 24%。这意味着用户可以在照片和视频上享受更快、更流畅的散景效果，而不会影响电池寿命。



然后，通过软件和工具的组合，客户可以在 Client for CSS 上进行进一步的性能和功耗优化。Arm 推出新的 Kleidi 库，该库具有 KleidiAI（高度优化的机器学习 （ML） 内核的集合），使开发人员能够在通过高度优化的生成式 AI 框架运行 AI 工作负载时释放 Arm CPU 的全部潜力。这意味着开发人员可以在最广泛的设备上以尽可能高的性能快速构建基于 AI 的应用程序。



与 Cortex-X4 相比，新的 Cortex-X925 平台在使用旨在加速现代 Arm CPU 上的 AI 应用的 KleidiAI 库时，在 LLaMA 3（80 亿个参数）中性能提升高达 42%，在 Phi 3（38 亿个参数）AI 模型中性能提升高达 46%。



对于更身临其境和更长的游戏会话，CSS for Client 提供了两位数的性能和能效改进。这包括在相同功耗下平均性能提升 37%，GPU 功耗降低 30%，在一系列流行的手机游戏中平均以每秒 120 帧 （fps） 的速度播放。  
  
  
跨平台、面向所有人的 AI  
  
  
Arm 表示，CSS for Client 是专为跨各种消费类设备提供下一代 AI 体验的平台，致力于为每个人提供 AI，通过 CSS for Client 可在广泛的消费类设备和外形尺寸上扩展，不懈地推动性能和效率。  
  
  
在移动设备上，用户将以前所未有的方式体验 Android，CSS for Client 将是 Arm 最快的 Android 计算平台。该平台的 PPA 优势是通过物理实施实现的，这些实施为我们的芯片合作伙伴提供了更快的上市时间和平滑的部署机会。  
  
  
CSS for Client 还可针对进入市场的性能最高的消费类设备进行扩展，其中包括下一代 AI PC，与 Arm Cortex-X4 CPU 相比，Cortex-X925 提供的 TOPS 提高了 50%。CSS for Client 为 PC 市场提供了一个专门构建的可扩展平台。它采用 Cortex-X925，通过新更新的 DSU-120 提供一流的单线程性能和最佳的性能可扩展性，可在单个 CPU 集群中提供多达 14 个 CPU 内核。除了 SVE2 之外，PC 市场还推出了更多的 Armv9 架构功能，包括指针身份验证 （PAC）、分支目标识别 （BTI） 和内存标记扩展 （MTE），这些技术已经在移动生态系统中得到了验证。  
  
  
通过 CSS for Client，Arm 在消费类设备市场的所有性能和成本点上提供可访问的 AI。Cortex-A725 是高能效 AI 吞吐量的主要处理器，是成本更敏感的大众市场消费技术细分市场的 AI 处理的主要主力和开发目标。例如，此虚拟助手演示展示了在使用 3 个 Cortex-A700 系列 CPU 内核的现有 Android 智能手机上运行 Llama2-7B 和 Phi-3 3.8B LLM 的性能。最后，区域优化的 Cortex-A725 允许在广泛的消费技术领域中高效部署生成式 AI 工作负载。  
  
  
显然，Arm CSS for Client 有望在帮助Arm推升面向旗舰级AI智能手机性能的同时，进一步向更高性能的AI PC市场进行开拓。预计接下来，将会有更多厂商（此前传闻显示联发科、英伟达都有此计划）借助Arm CSS for Client 进入AI PC市场，与苹果M系列和高通骁龙X系列进行竞争。  
  
  
Arm强调，CSS for Client 的可扩展性能功能可提供“面向所有人的 AI”，有助于在各种不同设备和外形规格类型中释放所有成本点的 AI 性能。  
  
  
Arm表示，CSS for Client 允许我们的生态系统做更多的事情，无论是释放更多的性能、更多的 AI、更多的应用体验，还是更先进的芯片，我们涵盖了所有的基础。通过该平台，Arm 正在为当今和未来的基于 AI 的体验构建消费者计算的未来。  
  
  
联发科天玑9400将首发  
  
  
值得一提的是，最新的爆料显示，联发科天玑9400或将成为首批配备Cortex-X925 CPU和Immortalis G925 GPU的芯片，而首发机型极有可能是vivo X200系列。对此联发科技资深副总经理、无线通信事业部总经理徐敬全博士和vivo 首席芯片规划专家夏晓菲，在Arm的新闻稿中对此也予以了证实。  
  
  
徐敬全博士表示，“我们将于今年下半年推出新一代旗舰移动芯片天玑 9400，该芯片将搭载最新的 Armv9 Cortex-X925 CPU 和 Arm Immortalis-G925 GPU。我们与 Arm 保持着长期而紧密的合作关系，致力于不断提升移动芯片的性能和功能，共同推动计算技术的快速发展。”  
  
  
夏晓菲表示：“vivo 非常注重用户体验，在 Arm CSS 的技术基础之上，我们与 Arm 的密切合作，共同推动开发者生态，使手机更流畅更好用，同时也为设备端带来了前沿的 AI 体验。我们很高兴看到 Arm 终端 CSS 所引入的新方向，以及通过 Armv9 CPU 与 Arm GPU，这套计算子系统将赋能新一代设备端生成式 AI，从而实现沉浸式的智能移动端解决方案。”  
  
  
在代工方面，Arm与台积电、英特尔、三星都进行了合作，为CSS for Client 基于他们的尖端制程的量产做好了准备。  
  
  
台积电生态系统与合作联盟管理部门负责人 Dan Kochpatcharin表示：“AI 优化的 Arm CSS 是 Arm 与台积公司的一个合作典范，共同助力芯片设计商实现前所未有的性能和能效，突破面向 AI 的半导体创新极限。通过与 Arm 以及我们开放创新平台 (OIP) 生态伙伴携手合作，我们能够提供最先进的工艺技术和设计解决方案，赋能客户加速 AI 创新。”  
  
  
英特尔代工副总裁兼生态系统技术办公室总经理 Suk Lee：“英特尔代工与 Arm 深度合作，此次发布彰显了 Arm 在终端业务上的进展。我们双方在 Intel 18A 工艺等前沿技术节点上密切合作，以提供一流的功率、性能和面积指标，并基于新的 Arm 终端 CSS 赋能新一代移动端 SoC 产品。”  
  
  
三星电子执行副总裁兼代工设计平台开发主管 Jongwook Kye表示：“在移动端设备中嵌入生成式 AI 功能是客户的主要需求，这就需要将顶尖的芯片技术与领先的计算解决方案相集成。Arm Cortex-X925 CPU 解决方案与三星代工厂最新的3nm GAA 工艺节点相结合，可以满足这一需求。我们与 Arm 的长期合作伙伴关系使我们能够在最大化 DTCO 和 PPA 方面从早期便开始展开紧密合作，从而按时交付芯片，并满足性能和效率需求。”  
  
  
编辑：芯智讯-浪客剑