



南開大學
Nankai University

计算机学院
并行程序设计第一次作业

Apple M1 体系架构调研

姓名：丁屹
学号：2013280
专业：计算机科学与技术

2022 年 2 月 26 日

目录

1	Apple M1 简述	2
2	设计	2
2.1	CPU	2
2.1.1	对比和分析	3
2.2	GPU	4
2.2.1	对比和分析	4
2.3	其他特性	4
3	性能和效率	5
3.1	分析	5
3.2	Benchmark 和对比	6
3.2.1	SPEC2006	6
3.2.2	SPEC2017	7
3.2.3	Cinebench R23	8
3.2.4	GeekBench 5	9
4	参考文献	9

1 Apple M1 简述

Apple M1 是苹果公司第一款基于 ARM 架构的自研处理器单片系统 (SoC)，为 Mac 产品线与 iPad 产品线提供中央处理器。M1 是首款用于个人电脑的 5 纳米芯片。苹果宣称该芯片在所有低功耗中央处理器产品中性能最佳，同时具有最佳的性能功耗比。

2 设计

Apple M1 成为主处理器之前，苹果从 T 系列就开始为迁移 arm 做准备。

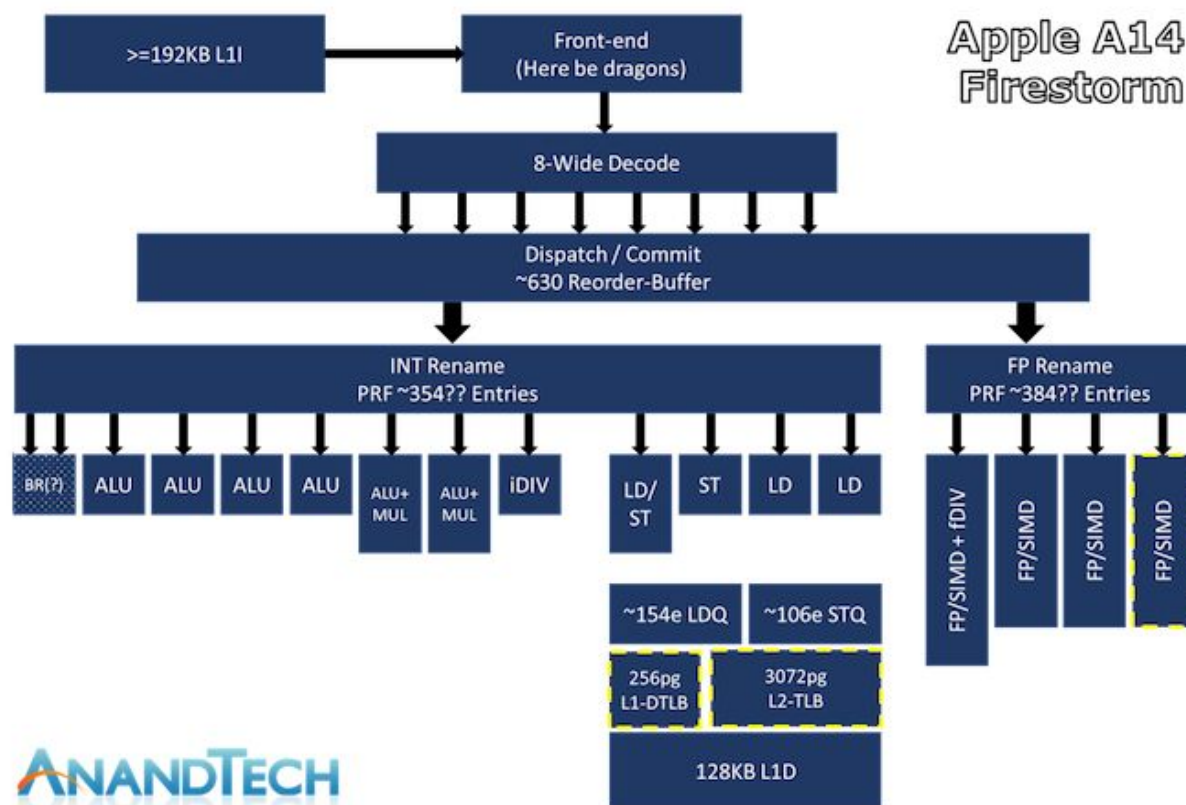
2.1 CPU

M1 拥有 4 个高性能 “Firestorm” 和 4 个节能 “Icestorm” 核心，提供类似于 ARM DynamIQ 和 Intel 的 Lakefield 和 Alder Lake 处理器的混合配置。这种组合可以实现以前的 Apple-Intel 架构设备无法实现的功耗优化。Apple 声称节能 (high-efficiency) 核心使用的功率是高性能 (high-performance) 核心的十分之一。高性能核心拥有特别大的 192 KB 一级指令缓存和 128 KB 一级数据缓存，并共享一个 12 MB 二级缓存；节能核心具有 128 KB L1 指令高速缓存、64 KB L1 数据高速缓存和共享的 4 MB L2 高速缓存。SoC 还具有由 GPU 共享的 16MB 系统级缓存。

由于苹果公司是一家无厂半导体公司，自身团队专攻芯片设计，最终的芯片制造需要晶圆代工厂来完成。苹果预订了台积电的 5 纳米生产线来制造该款芯片。苹果也是 TSMC 5 纳米生产线的前期客户之一。M1 Pro 和 M1 Max 具有 8 个高性能 “Firestorm” (M1 Pro 的低分档变体中有 6 个) 和 2 个节能的 “Icestorm” 核心，提供总共 10 个核心 (某些核心中有 8 个) 的混合配置 M1 Pro 的基本型号。

- 架构 AArch64
- 指令集架构 ARMv8-A
- 制作工艺/工艺 N5、N5P
- 最大 CPU 加速频率 3.2 GHz
- 核心数量
 - M1 8 ($4 \times \text{high-performance} + 4 \times \text{high-efficiency}$)
 - M1 Pro & MAX 10 ($8 \times \text{high-performance} + 2 \times \text{high-efficiency}$)
- L1 缓存
 - 192 + 128 KB per core (performance cores)
 - 128 + 64 KB per core (efficient cores)
- L2 缓存
 - 12 MB (performance cores)
 - 4 MB (efficient cores)

Anandtech 使用 Veedrac 的架构测试软件推测的架构图如下



2.1.1.1 对比和分析

从表1中可以看到，最重要的执行单元部分，M1 / A14 的 6 个 ALU，4 个 FPU 带来的最高理论性能直接就比其它 CPU 高 50% ~ 100%（暂不考虑 x86 256bit 的 AVX）。而 M1 的 L1 / L2 容量、解码宽度、ROB 规模都非常大，往往是其它 CPU 的两三倍（其中 12 MiB 的 L2 容量是四个大核共享，平均 3 MiB / 核心，但运行单线程应用的时候，理论上可以全部由单个核心使用），前端、调度单元、缓存的庞大规模，保证了执行单元能最高效率发挥性能。所以，不考虑更多细节的话，假设大家性能效率发挥基本一致，理论上 M1 的 IPC，整数方面应该比 Zen 2 / 3、SKL、SNC 都高 50%；不考虑 AVX，浮点方面比 Zen 2 / 3、SKL / SNC 高 100%。而 x86 的高频率所必须的长流水线，在分支预测失败需要重新填充流水线的时候，惩罚会比 M1 高，所以实际 IPC 要更低些；Intel 家的 FPU 和 ALU 共享发射端口，IPC 可能会更低一点。

解码器部分，x86 的变长指令解码器需要解析完一条指令得知本条指令长度后，才能计算出下一条指令的起始地址；在此之前，CPU 无法知道下一条指令从哪里开始，对连续的指令序列无法分拆后并行解码。虽然程序中会有类似跳转 (JMP)、调用 (CALL) 这样的分支指令，其操作数是某条指令的起始地址，另一个解码器可以从该地址开始解码，实现对指令序列进行并行解码。但毕竟程序中这样的指令密度有限，即便 CPU 中有更多的解码器对更多分支的指令并行解码，往往某个解码器解码后的 OPs 在多个时钟周期内也不会进入执行序列。今天的 x86 CPU，AMD 已经是四个复杂解码器，Intel 是一个复杂解码器 + 四个简单解码器，已经有点过于富裕，这也是两家的 CPU 都支持 SMT (Simultaneous MultiThreading, 同时多线程，也就是 Intel 的 Hyper Threading, 超线程)，并且开启 SMT 后多线程性能往往有相当幅度提升的一个重要原因，因为对两个线程并行解码后可以进入执行序列的 OPs 数量更多。

CPU	A13-Ligthning	A14-FireStorm	Cortex-A77	Cortex-A78	Cortex-X1	Zen2	Zen3	Skylake	Sunny Cove
L1l(KiB)	128	192	64	32/64	64	32	32	32	32
L1D(KiB)	128	128	64	32/64	64	32	32	32	48
L2(MiB)	8	12	0.25/0.5	0.25/0.5	1	0.5	0.5	0.25	0.5
解码宽度	7	8	4	4	5	4	4	5	5
ROB		630	160	160	224	224	256	224	352
Int Reg File		354				180	192	180	
FP Reg File		384				160	160	168	
ALU		6	3	4	4	4	4	4	4
FPU		128bx4	128bx2	128bx2	128bx4	256bx2	256bx2	256bx2	256bx3

表 1: 架构对比

Apple Silicon processor	M1	M1 Pro	M1 Pro	M1 Max	M1 Max
GPU core count	8	14	16	24	32
Teraflops	2.6	4.5	5.2	7.8	10.4
AMD equivalent GPU	Radeon RX 560 (2.6TF)	Radeon RX 5500M (4.6TF)	Radeon RX 5500 (5.2TF)	Radeon RX 5700M (7.9TF)	Radeon RX Vega 56 (10.5TF)
Nvidia equivalent GPU	GeForce GTX 1650 (2.9TF)	GeForce GTX 1650 Super (4.4TF)	GeForce GTX 1660 Ti (5.4TF)	GeForce RTX 2070 (7.4TF)	GeForce RTX 2080 (10TF)

表 2: Apple M1 的 GPU 性能参考表

而 RISC 的定长指令，指令抓取单元直接对指令序列按长度分割后，交由不同的解码器并行解码即可。因此可以看到两个超宽的 RISC 架构：IBM 最新的 POWER 可以配置为支持 SMT8，以及虽然不支持 SMT 但 IPC 惊人的苹果 A 系列。在 CISC 的 x86 上则没有出现过类似的超宽架构。

2.2 GPU

M1 集成了 Apple 设计的 8 核（在某些基本型号中为七核）图形处理单元 (GPU)。每个 GPU 核心被分成 16 个执行单元，每个执行单元包含 8 个算术逻辑单元 (ALU)。M1 GPU 总共包含多达 128 个执行单元或 1024 个 ALU，Apple 表示它们可以同时执行多达 24,576 个线程，其最大浮点 (FP32) 性能为 2.6 TFLOP。

M1 Pro 集成了 Apple 设计的 16 核（在某些基本型号中为 14 个）图形处理单元 (GPU)，而 M1 Max 则集成了一个 32 核（在某些基本型号中为 24 个）GPU。每个 GPU 核心分为 16 个执行单元，每个执行单元包含 8 个算术逻辑单元 (ALU)。M1 Max GPU 总共包含多达 512 个执行单元或 4096 个 ALU，其最大浮点 (FP32) 性能为 10.4 TFLOP。

2.2.1 对比和分析

如表2，Apple 的 M1 Pro 和 M1 Max 的 GPU 速度可与 AMD 和 Nvidia 的新版本显卡竞争，更高端的配置有望与游戏台式机和现代游戏机竞争。在 Apple 的主题演讲中，公司声称 M1 Pro 和 M1 Max 的核心比 M1 芯片多得多。据说，M1 Pro 的 16 核 GPU 比现代“8 核 PC 笔记本电脑芯片”上的集成显卡快 7 倍，并且比独立笔记本 GPU 提供更高的性能，同时功耗降低 70 据说 M1 Max 具有更高的性能，它显然可以在具有类似的能效的同时与紧凑型专业 PC 笔记本电脑中的高端 GPU 相媲美。

2.3 其他特性

M1 在处理器所有组件共享的统一内存配置中使用 4,266 MT/s LPDDR4X SDRAM。SoC 和 RAM 芯片以系统级封装设计安装在一起。提供 8 GB 和 16 GB 配置。

在统一内存中，GPU 核心和 CPU 核心可以同时访问内存，并且在这种情况下，共享内存没有开销。例如 CPU 和 GPU 可以互相告知一些内存的位置：以前 CPU 必须将数据从其主内存的区域复制到 GPU 使用的区域。而现在 GPU 可以直接开始使用该内存，而不需要做任何复制。这意味着，通过使用相同的内存池，M1 上的所有特殊的协同处理器都可以彼此快速交换信息，从而可以显著提高性能。

M1 在 16 核神经引擎中包含专用的神经网络硬件，每秒能够执行 11 万亿次操作。其他组件包括图像信号处理器 (ISP)、NVMe 存储控制器、Thunderbolt 4 控制器和 Secure Enclave。支持的编解码器包括 H264 和 H265 (8/10 位，最高 4:4:4)、VP9 和 JPEG。

3 性能和效率

3.1 分析

M1 在大量的基准测试中体现了其竞争力强大的性能和效率。配备 2020 M1 的 Mac mini 在空闲时消耗 7 瓦，在最大负载时消耗 39 瓦，而 2018 年 6 核 Intel i7 Mac mini 的空闲时为 20 瓦，最大负载为 122 瓦。M1 的能效使基于 M1 的 MacBook 的电池寿命比之前基于英特尔的 MacBook 增加了一倍。在发布时，MacBook Air (M1, 2020) 和 MacBook Pro (M1, 2020) 被认为是 Apple 生产的最快的 MacBook。

在提升性能的道路上，苹果并没有选择增加更多通用 CPU 核心，而是采取了另一种策略：添加更多专用芯片来完成一些专门的任务。这样做的好处是：与通用 CPU 核心相比，专用芯片可以使用更少的电流执行任务，而且速度还更快。

这并不是什么全新的技术。多年来，图形处理单元 (GPU) 等专用芯片已经存在于英伟达和 AMD 的显卡中，执行与图形相关的操作，速度比通用 CPU 快得多。苹果只是在这个方向上走得更加彻底。除了通用核心和内存之外，M1 包含了一系列专用芯片：GPU (图形处理器)、IPU (图像处理单元)、DSP (数字信号处理器)、NPU (神经网络处理器)、视频编码器 / 解码器、Secure Enclave。

而在 CPU 加速的方面，原则上可以通过以下两种策略来完成 CPU 加速的任务：

- 以更快的速度顺序执行更多指令
- 并行执行大量指令

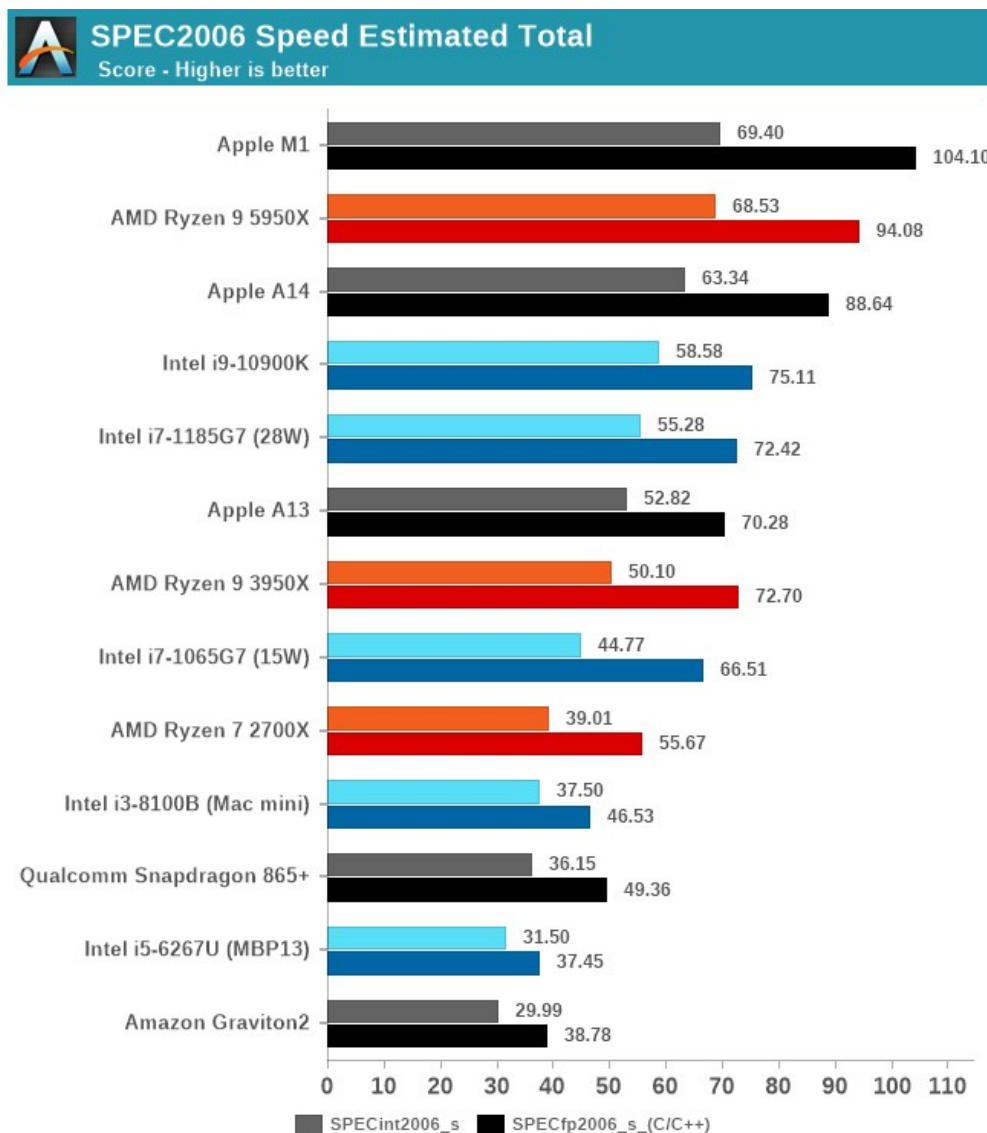
在上世纪 80 年代，这很容易做到。只要增加时钟频率，就能更快执行指令。每个时钟周期表示计算机执行某项任务的时间，但是这项任务可能非常微小。一条指令由多个较小的任务构成，因此可能需要多个时钟周期。但是现在已经几乎不可能增加时钟频率了，所以第二个策略“并行执行大量指令”是目前研发的重心。

苹果生产单用户设备，大量线程并不是优势。苹果的设备多用于游戏、视频编辑、开发等。苹果希望台式机具有精美的响应图形和动画。桌面软件通常不需要利用很多核心，例如，电脑游戏通常需要 8 个核心，在这种情况下 128 个核心就完全是浪费了。因此，用户需要的是更少但更强大的核心。

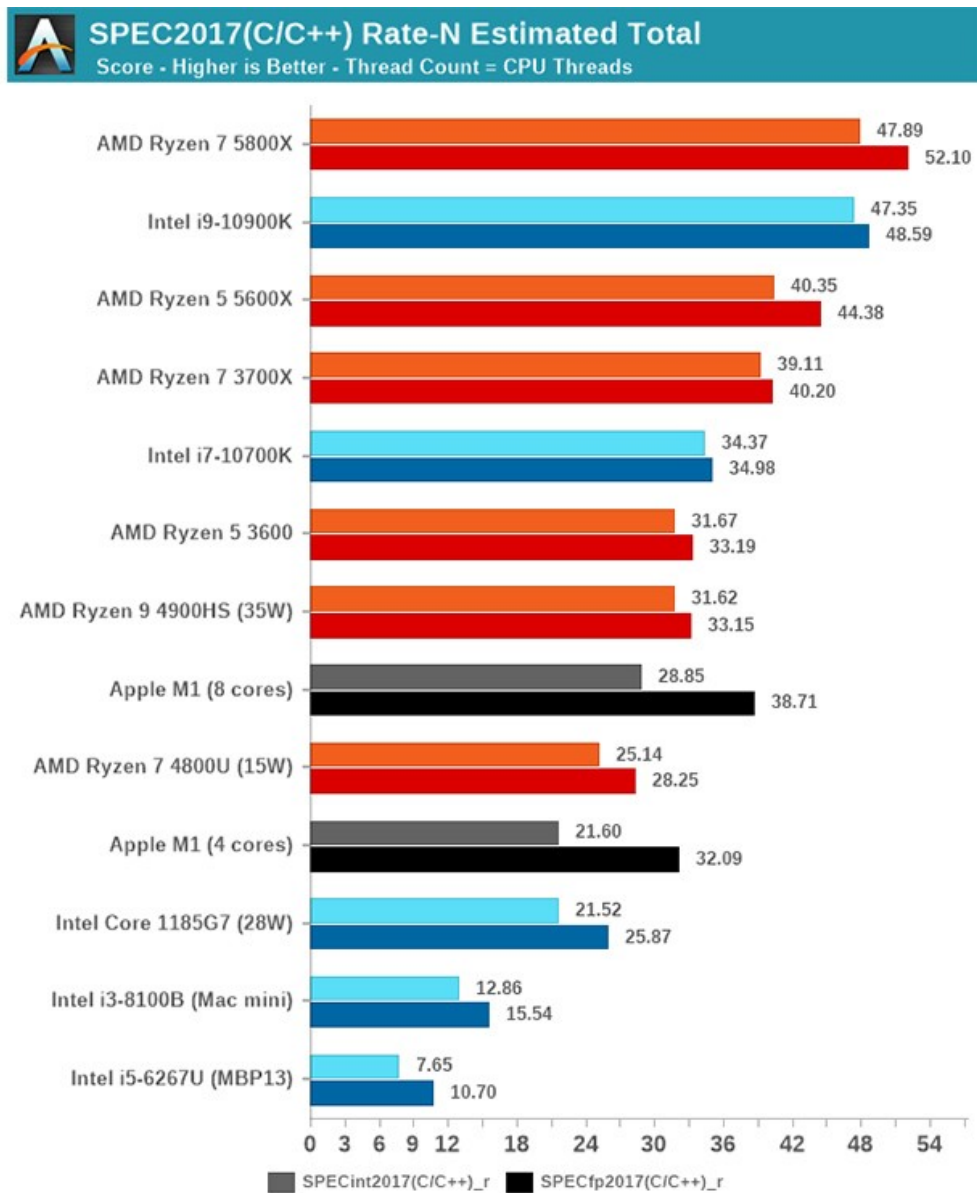
英特尔和 AMD 微处理器核心具有 4 个解码器，这意味着它可以并行解码 4 条指令，并输出微操作。但是苹果有 8 个解码器。不仅如此，ROB 还大了约 2 倍，基本上可以容纳 3 倍的指令。没有其他主流芯片制造商的 CPU 拥有如此多的解码器。

3.2 Benchmark 和对比

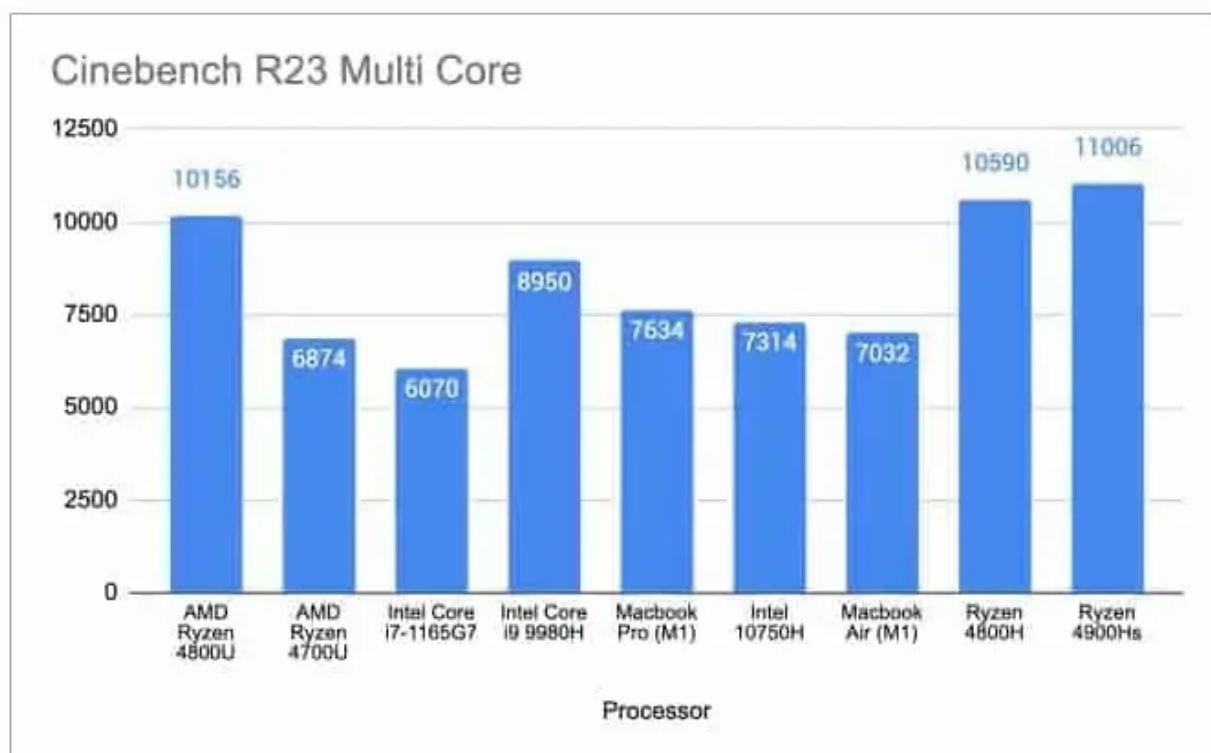
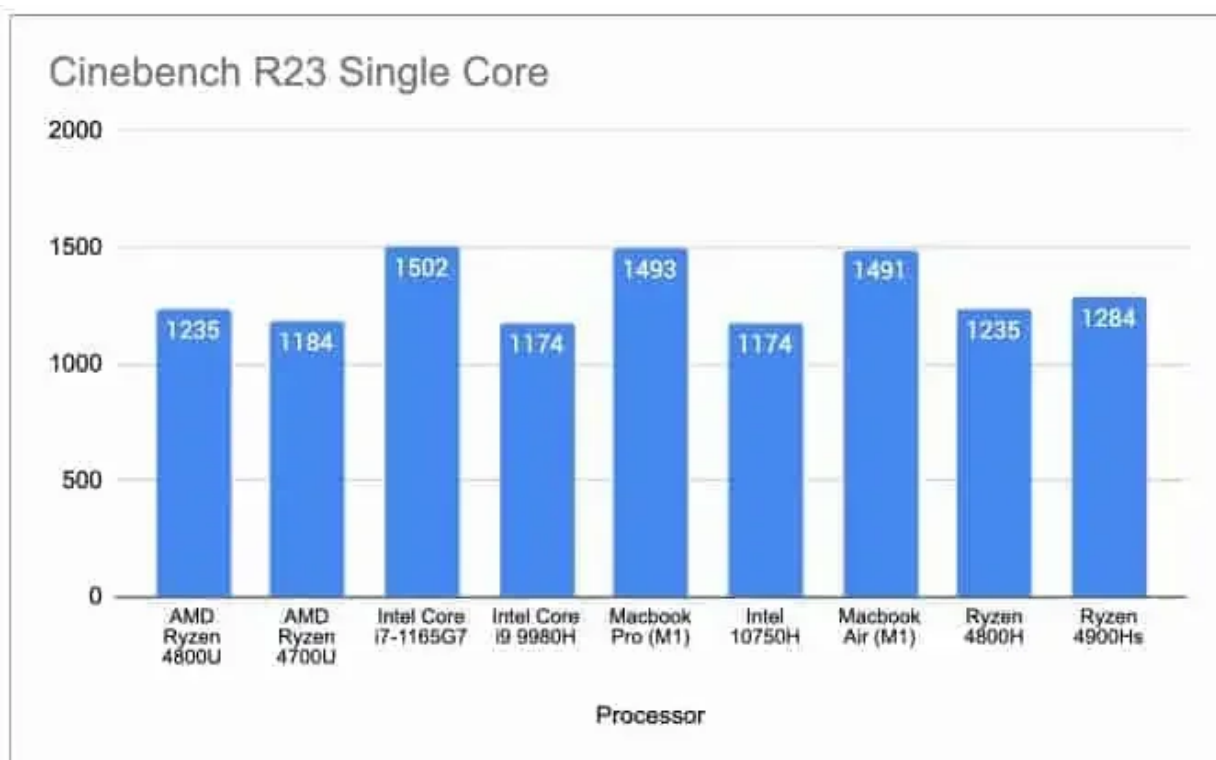
3.2.1 SPEC2006



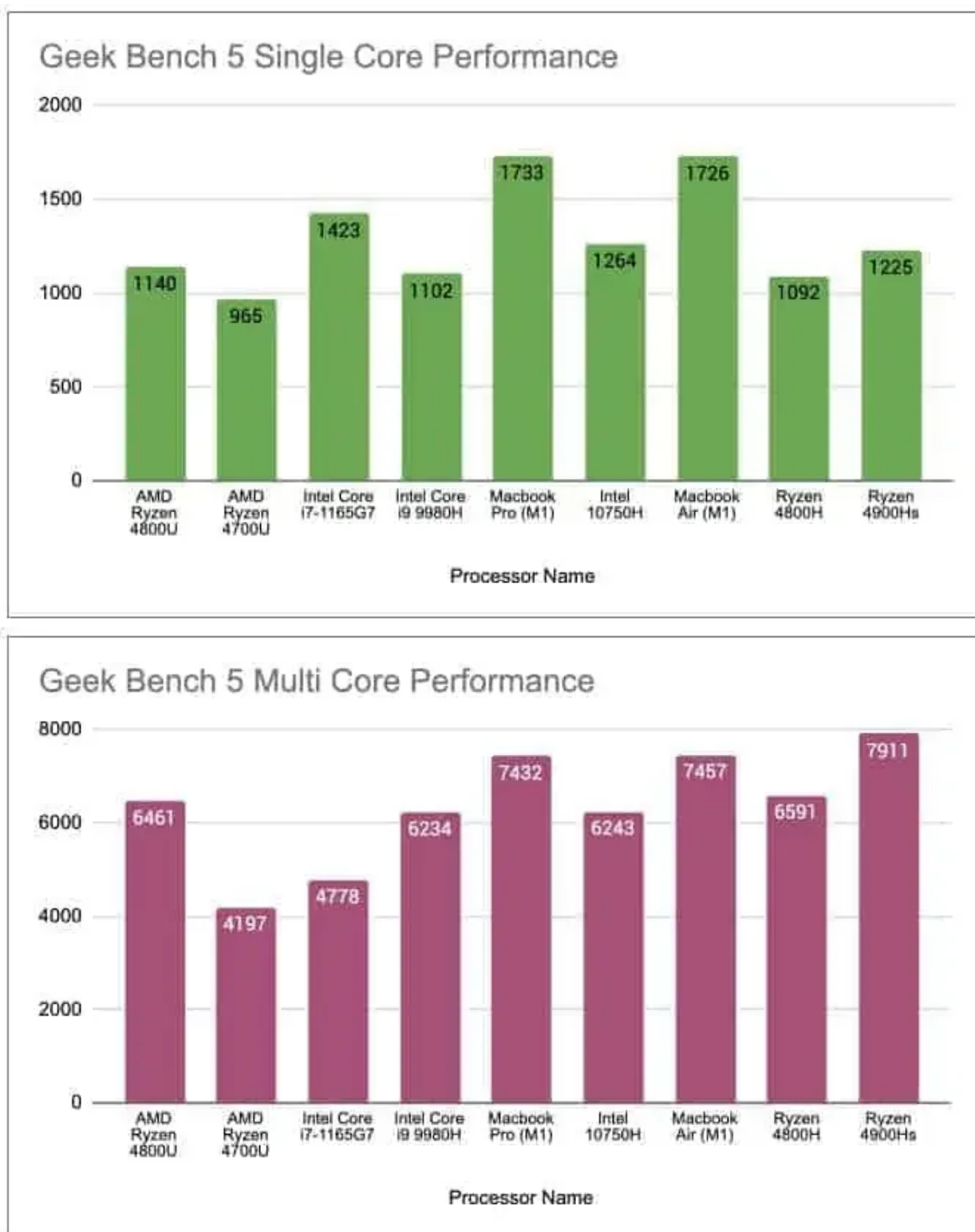
3.2.2 SPEC2017



3.2.3 Cinebench R23



3.2.4 GeekBench 5



4 参考文献

参考文献 [3][1][2]

参考文献

- [1] Gene H Golub and James M Ortega. *Scientific computing: an introduction with parallel computing*. Elsevier, 2014.
- [2] Stephen Bassi Joseph, Emmanuel Gbenga Dada, Sanjay Misra, and Samuel Ajoka. Parallel faces recognition attendance system with anti-spoofing using convolutional neural network. In *Illumination of Artificial Intelligence in Cybersecurity and Forensics*, pages 123–137. Springer, 2022.
- [3] Michael J Quinn. *Parallel computing theory and practice*. McGraw-Hill, Inc., 1994.