

21-面向流水线的指令设计（下）：奔腾4是怎么失败的？

上一讲，我给你初步介绍了CPU的流水线技术。乍看起来，流水线技术是一个提升性能的灵丹妙药。它通过把一条指令的操作切分成更细的多个步骤，可以避免CPU“浪费”。每一个细分的流水线步骤都很简单，所以我们的单个时钟周期的时间就可以设得更短。这也变相地让CPU的主频提升得很快。

这一系列的优点，也引出了现代桌面CPU的最后一场大战，也就是Intel的Pentium 4和AMD的Athlon之间的竞争。在技术上，这场大战Intel可以说输得非常彻底，Pentium 4系列以及后续Pentium D系列所使用的NetBurst架构被完全抛弃，退出了历史舞台。但是在商业层面，Intel却通过远超过AMD的财力、原本就更大的市场份额、无所不用的竞争手段，以及最终壮士断腕般放弃整个NetBurst架构，最终依靠新的酷睿品牌战胜了AMD。

在此之后，整个CPU领域竞争的焦点，不再是Intel和AMD之间的桌面CPU之战。在ARM架构通过智能手机的快速普及，后来居上，超越Intel之后，移动时代的CPU之战，变成了高通、华为麒麟和三星之间的“三国演义”。

“主频战争”带来的超长流水线

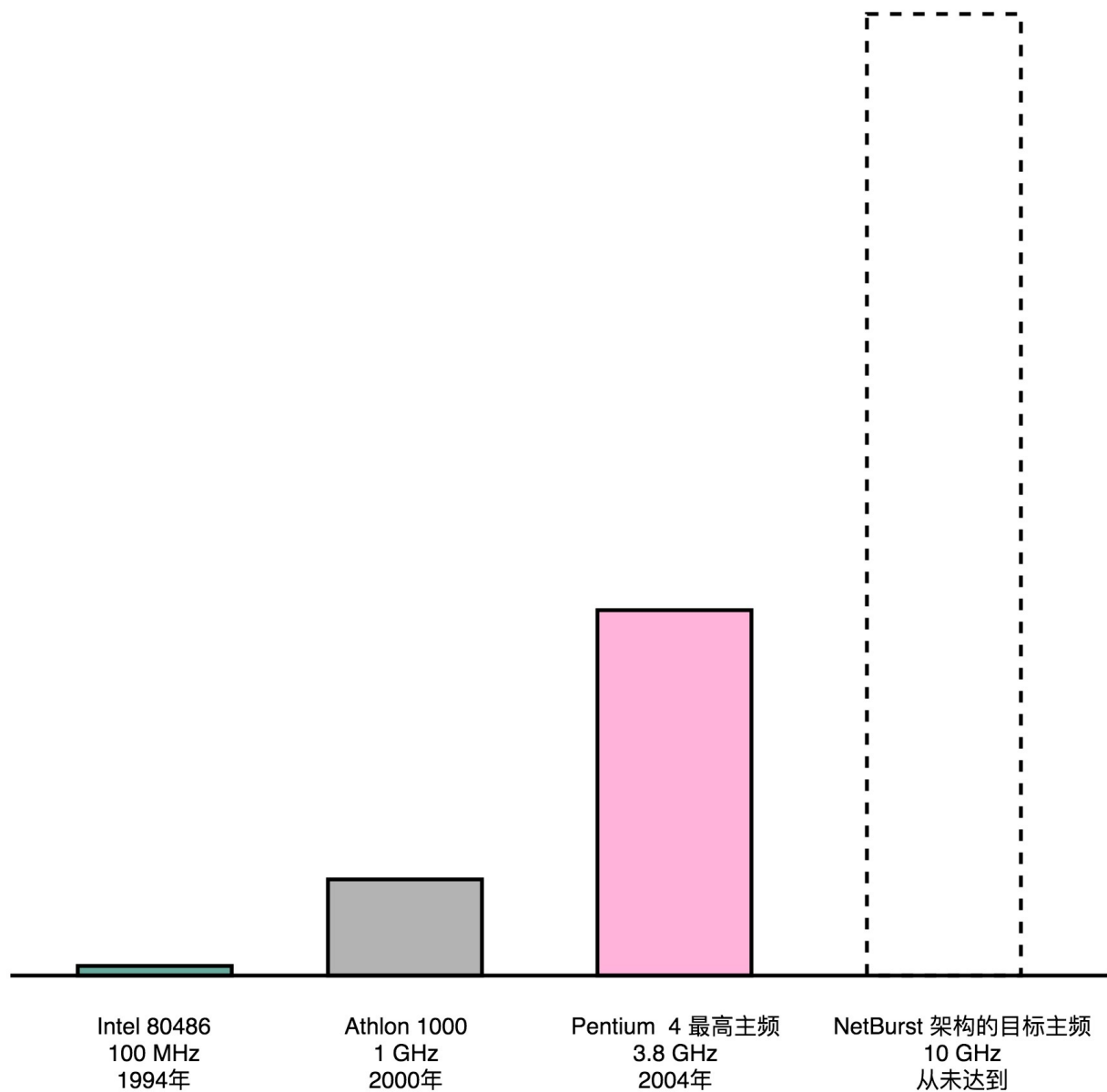
我们在[第3讲](#)里讲过，我们其实并不能简单地通过CPU的主频，就来衡量CPU乃至计算机整机的性能。因为不同的CPU实际的体系架构和实现都不一样。同样的CPU主频，实际的性能可能差别很大。所以，在工业界，更好的衡量方式通常是，用SPEC这样的跑分程序，从多个不同的实际应用场景，来衡量计算机的性能。

但是，跑分对于消费者来说还是太复杂了。在Pentium 4的CPU面世之前，绝大部分消费者并不是根据跑分结果来判断CPU的性能的。大家判断一个CPU的性能，通常只看CPU的主频。而CPU的厂商们也通过不停地提升主频，把主频当成技术竞赛的核心指标。

Intel一向在“主频战争”中保持领先，但是到了世纪之交的1999年到2000年，情况发生了变化。

1999年，AMD发布了基于K7架构的Athlon处理器，其综合性能超越了当年的Pentium III。2000年，在大部分CPU还在500~850MHz的频率下运行的时候，AMD推出了第一代Athlon 1000处理器，成为第一款1GHz主频的消费级CPU。在2000年前后，AMD的CPU不但性能和主频比Intel的要强，价格还往往只有Intel的2/3。

在巨大的外部压力之下，Intel在2001年推出了新一代的NetBurst架构CPU，也就是Pentium 4和Pentium D。Pentium 4的CPU有个最大的特点，就是高主频。2000年的Athlon 1000的主频在当时是最高的，1GHz，然而Pentium 4设计的目标最高主频是10GHz。



为了达到这个10GHz，Intel的工程师做出了一个重大的错误决策，就是在NetBurst架构上，使用超长的流水线。这个超长流水线有多长呢？我们拿在Pentium 4之前和之后的CPU的数字做个比较，你就知道了。

Pentium 4之前的Pentium III CPU，流水线的深度是11级，也就是一条指令最多会拆分成11个更小的步骤来操作，而CPU同时也最多会执行11条指令的不同Stage。随着技术发展到今天，你日常用的手机ARM的CPU或者Intel i7服务器的CPU，流水线的深度是14级。

可以看到，差不多20年过去了，通过技术进步，现代CPU还是增加了一些流水线深度的。那2000年发布的Pentium 4的流水线深度是多少呢？答案是20级，比Pentium III差不多多了一倍，而到了代号为Prescott的90纳米工艺处理器Pentium 4，Intel更是把流水线深度增加到了31级。

要知道，增加流水线深度，在同主频下，其实是降低了CPU的性能。因为一个Pipeline Stage，就需要一个时钟周期。那么我们把任务拆分成31个阶段，就需要31个时钟周期才能完成一个任务；而把任务拆分成11个阶段，就只需要11个时钟周期就能完成任务。在这种情况下，31个Stage的3GHz主频的CPU，其实和11个Stage的1GHz主频的CPU，性能是差不多的。事实上，因为每个Stage都需要有对应的Pipeline寄存器的开销，这个时候，更深的流水线性能可能还会更差一些。

我在上一讲也说过，流水线技术并不能缩短单条指令的**响应时间**这个性能指标，但是可以增加在运行很多条指令时候的**吞吐率**。因为不同的指令，实际执行需要的时间是不同的。我们可以看这样一个例子。我们顺序

执行这样三条指令。

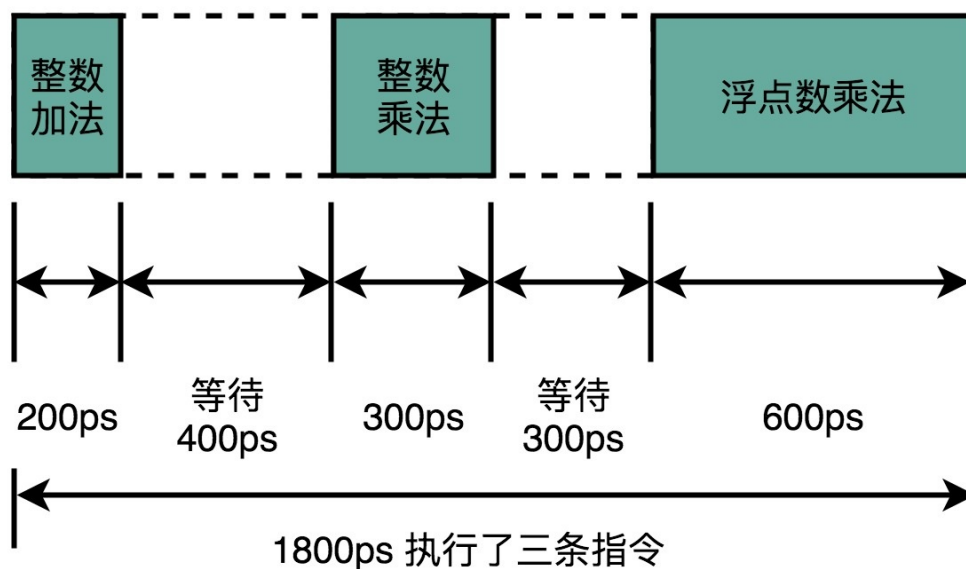
1. 一条整数的加法，需要200ps。
2. 一条整数的乘法，需要300ps。
3. 一条浮点数的乘法，需要600ps。

如果我們是在单指令周期的CPU上运行，最复杂的指令是一条浮点数乘法，那就需要600ps。那这三条指令，都需要600ps。三条指令的执行时间，就需要1800ps。

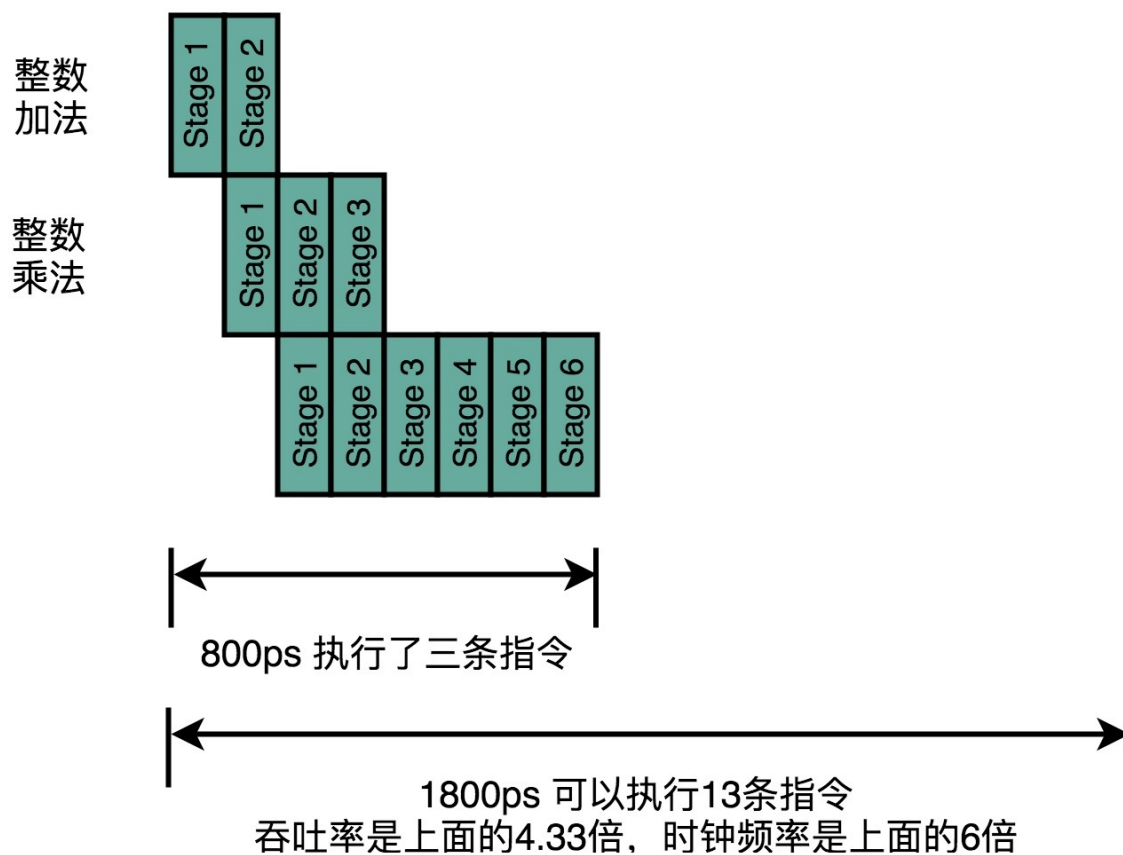
如果我们采用的是6级流水线CPU，每一个Pipeline的Stage都只需要100ps。那么，在这三个指令的执行过程中，在指令1的第一个100ps的Stage结束之后，第二条指令就开始执行了。在第二条指令的第一个100ps的Stage结束之后，第三条指令就开始执行了。这种情况下，这三条指令顺序执行所需要的总时间，就是800ps。那么在1800ps内，使用流水线的CPU比单指令周期的CPU就可以多执行一倍以上的指令数。

虽然每一条指令从开始到结束拿到结果的时间并没有变化，也就是响应时间没有变化。但是同样时间内，完成的指令数增多了，也就是吞吐率上升了。

没有流水线



6级流水线



新的挑战：冒险和分支预测

那到这里可能你就要问了，这样看起来不是很好么？Intel的CPU支持的指令集很大，我们之前说过有2000

多条指令。有些指令很简单，执行也很快，比如无条件跳转指令，不需要通过ALU进行任何计算，只要更新一下PC寄存器里面的内容就好了。而有些指令很复杂，比如浮点数的运算，需要进行指数位比较、对齐，然后对有效位进行移位，然后再进行计算。两者的执行时间相差二三十倍也很正常。

既然这样，Pentium 4的超长流水线看起来很合理呀，为什么Pentium 4最终成为Intel在技术架构层面的大失败呢？

第一个，自然是在第3讲里讲过的功耗问题。提升流水线深度，必须要和提升CPU主频同时进行。因为在单个Pipeline Stage能够执行的功能变简单了，也就意味着单个时钟周期内能够完成的事情变少了。所以，只有提升时钟周期，CPU在指令的响应时间这个指标上才能保持和原来相同的性能。

同时，由于流水线深度的增加，我们需要的电路数量变多了，也就是我们所使用的晶体管也就变多了。

主频的提升和晶体管数量的增加都使得我们CPU的功耗变大了。这个问题导致了Pentium 4在整个生命周期里，都成为了耗电和散热的大户。而Pentium 4是在2000 ~ 2004年作为Intel的主打CPU出现在市场上的。这个时间段，正是笔记本电脑市场快速发展的时间。在笔记本电脑上，功耗和散热比起台式机是一个更严重的问题了。即使性能更好，别人的笔记本可以用上2小时，你的只能用30分钟，那谁也不爱买啊！

更何况，Pentium 4的性能还更差一些。**这个就要我们说到第二点了，就是上面说的流水线技术带来的性能提升，是一个理想情况。在实际的程序执行中，并不一定能够做得到。**

还回到我们刚才举的三条指令的例子。如果这三条指令，是下面这样的三条代码，会发生什么情况呢？

```
int a = 10 + 5; // 指令1
int b = a * 2; // 指令2
float c = b * 1.0f; // 指令3
```

我们会发现，指令2，不能在指令1的第一个Stage执行完成之后进行。因为指令2，依赖指令1的计算结果。同样的，指令3也要依赖指令2的计算结果。这样，即使我们采用了流水线技术，这三条指令执行完成的时间，也是 $200 + 300 + 600 = 1100 \text{ ps}$ ，而不是之前说的 800 ps 。而如果指令1和2都是浮点数运算，需要 600 ps 。那这个依赖关系会导致我们需要的时间变成 1800 ps ，和单指令周期CPU所要花费的时间是一样的。

这个依赖问题，就是我们在计算机组成里面所说的**冒险**（Hazard）问题。这里我们只列举了在数据层面的依赖，也就是数据冒险。在实际应用中，还会有**结构冒险**、**控制冒险**等其他的依赖问题。

对应这些冒险问题，我们也有在**乱序执行**、**分支预测**等相应的解决方案。我们在后面的几讲里面，会详细讲解对应的知识。

但是，我们的流水线越长，这个冒险的问题就越难一解决。这是因为，同一时间同时在运行的指令太多了。如果我们只有3级流水线，我们可以把后面没有依赖关系的指令放到前面来执行。这个就是我们所说的乱序执行的技术。比方说，我们可以扩展一下上面的3行代码，再加上几行代码。

```
int a = 10 + 5; // 指令1
int b = a * 2; // 指令2
```

```
float c = b * 1.0f; // 指令3
int x = 10 + 5; // 指令4
int y = a * 2; // 指令5
float z = b * 1.0f; // 指令6
int o = 10 + 5; // 指令7
int p = a * 2; // 指令8
float q = b * 1.0f; // 指令9
```

我们可以不先执行1、2、3这三条指令，而是在流水线里，先执行1、4、7三条指令。这三条指令之间是没有依赖关系的。然后再执行2、5、8以及3、6、9。这样，我们又能够充分利用CPU的计算能力了。

但是，如果我们有20级流水线，意味着我们要确保这20条指令之间没有依赖关系。这个挑战一下子就变大了很多。毕竟我们平时撰写程序，通常前后的代码都是有一定的依赖关系的，几十条没有依赖关系的指令可不好找。这也是为什么，超长流水线的执行效率反而降低了一个重要原因。

总结延伸

相信到这里，你对CPU的流水线技术，有了一个更加深入的了解。你会发现，流水线技术和其他技术一样，都讲究一个“折衷”（Trade-Off）。一个合理的流水线深度，会提升我们CPU执行计算机指令的吞吐率。我们一般用IPC（Instruction Per Cycle）来衡量CPU执行指令的效率。

IPC呢，其实就是我们之前在第3讲讲CPI（Cycle Per Instruction）的倒数。也就是说， $IPC = 3$ 对应着 $CPI = 0.33$ 。Pentium 4和Pentium D的IPC都远低于自己上一代的Pentium III以及竞争对手AMD的Athlon CPU。

过深的流水线，不仅不能提升计算机指令的吞吐率，更会加大计算的功耗和散热问题。Intel自己在笔记本电脑市场，也很快放弃了Pentium 4，而是主推了使用Pentium III架构的图拉丁CPU。

而流水线带来的吞吐率提升，只是一个理想情况下的理论值。在实践的应用过程中，还需要解决指令之间的依赖问题。这个使得我们的流水线，特别是超长的流水线的执行效率变得很低。要想解决好冒险的依赖关系问题，我们需要引入乱序执行、分支预测等技术，这也是我在后面几讲里面要详细讲解的内容。

推荐阅读

除了之前的教科书之外，我推荐你读一读[Modern Microprocessors, A 90-Minute Guide!](#)这篇文章。这篇文章用比较浅显的方式，介绍了现代CPU设计的多个方面，很适合作为一个周末读物，快速理解现代CPU的设计。

课后思考

除了我们这里提到的数据层面的依赖，你能找找我们在程序的执行过程中，其他的依赖情况么？这些依赖情况又属于我们说的哪一种冒险呢？


欢迎留言和我分享你的疑惑和见解。你也可以把今天的内容，分享给你的朋友，和他一起学习和进步。

深入浅出计算机组成原理

带你掌握计算机体系全貌

徐文浩 bothub 创始人



新版升级：点击「 请朋友读」，20位好友免费读，邀请订阅更有**现金**奖励。

精选留言：

- 免费的人 2019-06-12 13:41:25
比如下一条该取哪一条指令决定于上一条指令的结果，if...else...分支
- 陈华应 2019-06-12 12:48:33
老师，为什么没有依赖关系的指令的流水级可以并行执行？
- Linuxer 2019-06-12 09:13:50
条件分枝也是一种依赖吧
- 殷勤的匠人 2019-06-12 08:50:34
...单个时钟周期内能够完成的事情变少了。所以，只有提升时钟周期，CPU 在指令的响应时间...

此處筆誤。應是降低週期，提升「主頻」。