IBM POWER7 multicore server processor

邹旭 201828013229112

2018 年 10 月 19 日

IBM 研制 POWER7，采用 45nm 工艺，它拥有大容量片上嵌入式动态随机存取 cache（eDRAM）和 8 个高性能四路多线程核。相较于 POWER6，它在一下四个方面有所突出：
1）减少了核面积和功耗，这是通过微结构的提升，指令和线程并行度等提高实现的。
2）内存总线的频率更高，带宽更大。
3）去掉了之前的外部三级 Cache，有效降低了功耗。
4）POWER7 的每个核在浮点计算方面性能提升了一倍。

POWER7 core

power7 的核大致由 7 个模块组成：取指和译码模块，指令排序模块，定点单元，十进制单元，向量和标量单元，存取单元以及其他。
它有先进的预测和指令预取功能，同时也能高效地进行乱序执行。核可以动态地改变单线程模式，支持单线程，双线程和四线程。POWER7 的优化主要针对于减小核的面积和功耗。它采用了一种方法重命名寄存器能支持比 POWER5 多一倍的线程数。早期的乱序执行处理器中，GPR，FPR，VR 的重命名结构是分开的，而 POWER7 中，它们合并成了一个统一的有 80 个入口的结构。早期的浮点和定点发射队列也是分开的，POWER7 中也统一了。L1Cache 由很多的 bank 组成，每个 bank 同时只能两个读操作或者一个写操作，而高度 bank 化之后可以支持高并行的读写操作。

取指

取指单元包含了高精确度的分支预测机制。它有一个 32KB 四路组相联的指令 Cache，并且使用一个 IEADIR 对 Cache 快速进行路选择。IFU 单元将指令从 L2Cache 取到 L1Cache，每次读取 4 个 32 字节的块。当指令从内存子单元取出时，两个时钟周期被用来做预译码操作，预译码后的编码用于做分支预测，分组，以及报告可能的异常等。取值地址寄存器用于追踪每个线程的程序计数器地址，每个周期有一个线程的指令地址用来取值，每个周期取 8 条指令。
分支预测通过以下两种机制来预测：
1）非子程序返回的间接分支使用一个共享的 128 入口的计数 Cache 来预测。
2）子程序返回使用一个链接栈来预测，每个线程一个栈。执行每一条分支链接指令的时候，它的下一条指令的地址被压到相应进程的链接栈中。

指令排队单元

指令排队单元完成的工作有：调度指令，重命名寄存器，发射指令，完成指令，处理异常条件等。

POWER7 以为 group 基础调度指令，每次将一个线程的一个 group 指令分派到 ISU 里面。在这个 group 分配之前，所有的资源必须准备好。每个 group 最多包含两条分支指令和四条非分支指令。发射队列包括 48 入口的 UQ，12 入口的分支发射队列（BRQ），8 入口的 CRQ，整个发射队列每个时钟周期最多可以发送八条指令。指令的追踪和完成也都由 ISU 负责，当一条指令成功执行完之后，它被标记为完成，当一个 group 里面的所有指令执行完之后，整个 group 才算完成，这个 group 的所有操作的结果对整个结构可见，并且它要释放所有的资源。指令流水线的刷新也由 ISU 控制，比如分支预测错误等。

取数据

取数操作由 LSU 完成，它包含两个对称的 LS 流水线。取数单元包含的子单元有 LS 地址生成和执行单元，存储数据队列，加载数据队列，加载缺失处理，地址转换机制，L1 数据 Cache 等。单线程和双线程模式下，每条流水线执行一个线程，四个线程下，每条流水线分配两个线程。LSU 的流水线可以执行定点加和逻辑操作，这样可以支持更多的定点操作。LSU 实现了两个队列：存储重排序队列 SRQ 和加载重排序队列 LRQ。队列由 32 入口的实地址 CAM 构成。L1 数据 Cache 有三个端口，两个写一个读。它又分为四个 macro，每一个由 16 个 bank 组成，只要对 bank 的访问不冲突，L1 可以支持多个指令的访问。L1Cache 是写穿透的，写操作直接写入 L2。L1 里面有一个预测数组，当 L1 数据缺失时，预测机制根据有效地址做一个哈希来进行数据预测。加载缺失

定点单元

定点单元包含两条流水线，每个定点流水线包含多端口 GPR 文件，一个算术逻辑单元，一个轮转单元，一个计数器，一个比特选择单元，一个除法器，一个乘法器。一个混合执行单元。大多数指令的执行只需要一个时钟周期。每个 FXU 的中心是一个 GPR 文件，两个写端口每个时钟周期进行两次写操作，因而逻辑上提供了四个写端口，节约了功耗。POWER7 实现了一个 64 位的乘法流水线，流水线延迟是 4 周期，并且每个周一能完成一条指令。在一条比较指令和条件分支指令之间的延迟一般对性能的影响较大，FXU 通过一个快速的客户比较宏实现了比 ALU 更快的比较。为了满足几个重要应用的性能需求，新的面向比特的定点指令被加入指令集中。

向量和标量执行

为了加速浮点操作，POWER7 实现了四路乘加二进制浮点流水线。

Cache 级

Cache 级在以下几个方面进行了优化：

1）将流水线的高频设计改为功耗性能优化设计；

2）将指令有序为主改为指令无序为主；

3）每个核由两个线程变成四个线程；

4）讲 L1 数据 Cache 的大小由 64KB 减为 32KB，同时降低延迟；

5）每片由两核变成八核。

L3 的替换策略有所改进，栅栏同步寄存器 BSR 被虚拟化。

内存子系统

每块 POWER7 集成了两个内存控制器，每个控制器支持四个内存通道。POWER7 为了给 8 个核提供充足的访存带宽以及 8 核平衡的访寻操作，它采用了以下几种策略：

1）通过给内存提供更多的线外信号资源来消除片外 L3Cache；

2）开发一种不同的高频信号技术来提供更高的带宽，同时降低功耗；

3）实现一种低开销的循环冗余检错协议；

4）开发一种分级的缓存方案来提供从每个缓存直接访问内存借口的功能；

5）利用 DDR3 的内存技术来提供更快的访存频率，但是会引入更复杂的方案规则和更复杂的错误检测代码；

6）开发一种先进的时序安排算法来实现更高的接口利用。

7）分配更多的缓存资源来提供更大的读写操作池来满足现金的时序安排。

内存子系统

POWER7 芯片支持两个集成的 I/O 控制器。每个控制器提供专有的四比特片外度端口和四比特片外写端口。

片上互联：为了延续早期 POWER 系列芯片的高可扩展性和低延迟的特点，POWER7 芯片利用一种基于非阻塞广播一致性传输机制，该机制基于 POWER6 平台提供的分布式处理。

片间互联：POWER7 系统拓扑是由 POWER6 系统发展过来的。标准的商业系统可以将 32 个 POWER7 芯片连接起来形成一个简单的 256 路 SMP 系统。一级节结构组合了 4 个 POWER7 芯片，二级的节结构组合了八片。

集群互联：这种互联用于大宗的多浮点操作高带宽集群系统。