IBM POWER7 multicore server processor

IBM POWERA 处理器是当今世界上占主导地位的简化指令集计算微处理器，在过去 20 年里有着丰富的实现和创新历史。本文介绍了 POWER7A 处理器芯片的主要特点。在芯片上是一个八核处理器，每个核心能够同时进行四路多线程操作。这种芯片采用 IBM 公司 45 纳米绝缘体硅(SOI)技术，由 11 个级别的金属制成，包含了超过 10 亿个晶体管。处理器内核和缓存得到了显著增强，以提高单线程、面向响应时间和多线程、面向吞吐量的应用程序的性能。内存子系统包含三级片上缓存，SOI 嵌入式动态随机存取存储器(DRAM)设备作为最后一级缓存。讨论了一种使用缓冲双数据速率的新内存接口——三种 DRAM 和可靠性、可用性和可服务性方面的改进。

1. 介绍

多年来，IBM POWER*处理器[1 4]引入了简化的指令集计算体系结构、高级分支预测、无序执行、数据预取、多线程、双核芯片、核加速器[5]、片上高带宽内存控制器和高度可伸缩的对称多处理(SMP)互连。在这个第 7 代 POWER 处理器中，IBM 引入了一种平衡的八核多芯片设计，具有大型片上嵌入式动态随机存取存储器(eDRAM)缓存和高性能的四向多线程内核，实现了 IBM PowerPC*体系结构版本 2.06。

图 1 显示了 POWER7 芯片，它有 8 个处理器核心，每个核心都有 12 个能够同时运行 4 个线程的执行单元。为了提供这 8 个高性能内核，POWER7 在芯片的每一侧都有两个内存控制器。每个内存控制器支持四个双数据速率- 3 (DDR3)内存通道。这 8 个通道总共提供了 100 GB/s 的持续内存带宽。在芯片的顶部和底部是 SMP 链接，提供 360 GB/s 的相干带宽，允许 POWER7 有效地扩展到 32 个插槽。

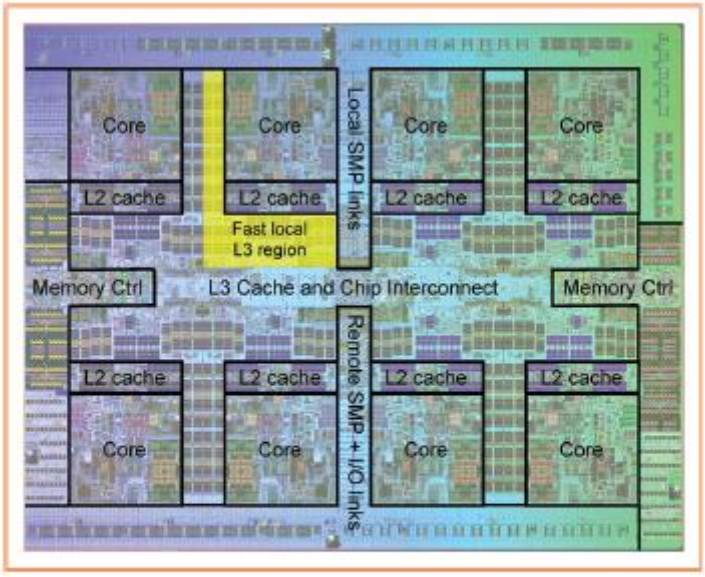

Figure 1

Die photo of the IBM POWER7 chip.

2. power 7 core

图 2 显示了处理器核心的平面图。POWER7 具有高级的分支预测和预取功能，以及深度无序执行功能，可以显著提高 ST 的性能。同时，它有足够的资源来有效地支持每个核心的 4 个线程。核心可以在 ST、双向 SMT(或 SMT2)和 SMT4 模式之间动态改变模式。

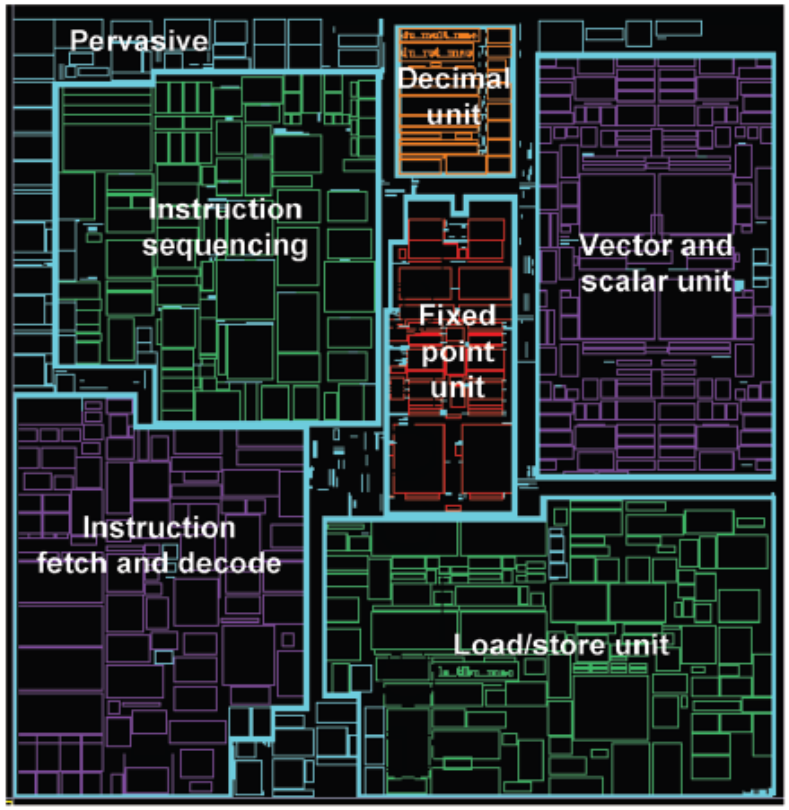

Figure 2

IBM POWER7 processor core floorplan.

图 3 显示了处理器核心中的指令流。

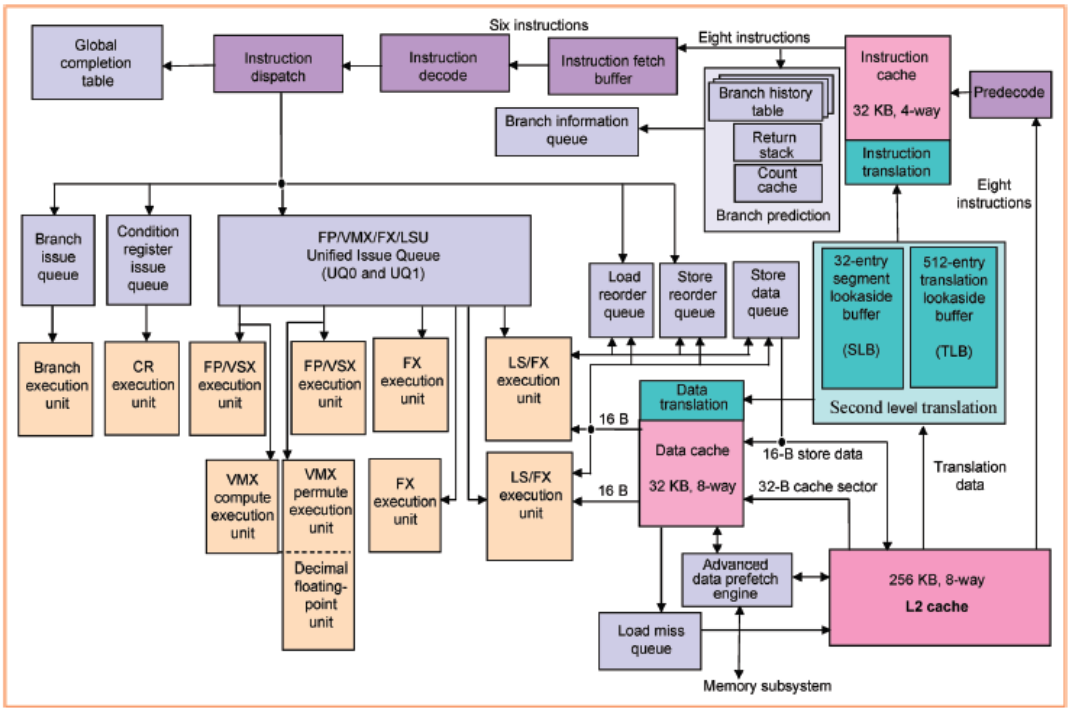

Figure 3

IBM POWER7 processor core pipeline flow.

图 4 显示了指令如何流向各种问题队列，然后被发送到功能单元以执行。

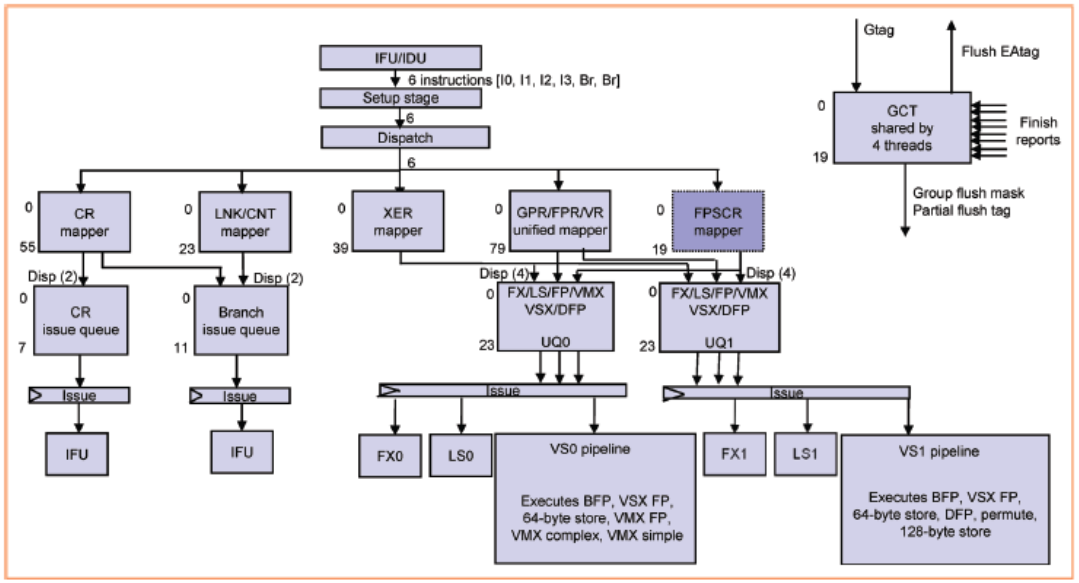

Figure 4

IBM POWER7 ISU overview.

此外，POWER7 还实现了健壮的 RAS 特性。它可以检测大多数软错误。在软错误检测时，内核会自动刷新管道中的指令，并返回并重新执行它们，这样就不会丢失数据完整性。有关 RAS 的部分提供了更多细节。

3. 取指（Instruction fetching）

POWER7 中的 IFU 负责向指令管道提供最可能的指令流，这些指令流基于高度精确的分支预测机制，远早于执行点，来自每个活动线程。IFU 还负责维护基于软件指定的线程优先级的活动线程的指令执行速率的平衡，对指令管道其余部分的指令进行解码和形成指令组，以及执行分支指令。

图 5 显示了 POWER7 中的指令获取和解码管道阶段。

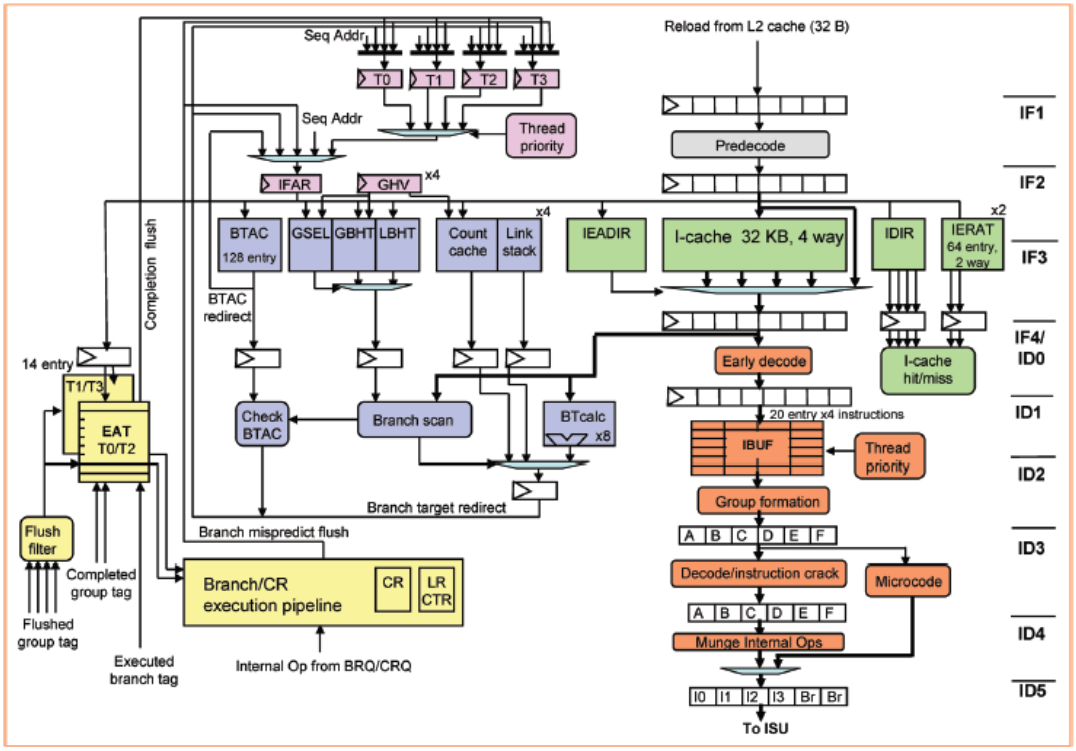

Figure 5

IBM POWER7 instruction fetch and decode pipe stages.

4. 分支预测（Branch Prediction）

条件分支的方向是通过一个复杂的分支历史表(BHTs)来预测的，该表由一个 8-K 条目本地 BHT (LBHT)数组、一个 16-K 条目全局 BHT (GBHT)数组和一个 8-K 条目全局选择(GSEL)数组组成。这些数组一起为每个循环中 fetch 组中的所有指令提供分支方向预测。fetch 组最多可以有 8 条指令，所有指令都可以是分支。这些数组由所有活动线程共享。本地数组直接由指令取回地址的 10 位索引。GBHT 和 GSEL 数组通过指令获取地址进行索引，该地址通过 21 位全局历史向量(GHV)折叠成 11 位，每个线程 1 位。GSEL 条目中的值用于在

LBHT 和 GBHT 之间进行选择，以预测每个分支的方向。所有的 BHT 条目由 2 位组成，高阶位决定方向(取或不取)，低阶位提供滞后。

分支目标地址的预测使用以下两种机制:

1)非子例程返回的间接分支的预测使用 128 个条目计数缓存，由所有活动线程共享。计数缓存使用一个地址进行索引，该地址是通过执行 7 位 XOR(分别来自指令获取地址和 GHV)获得的。计数缓存中的每个条目包含一个 62 位的预测地址和两个置信位。如果间接分支预测不正确，则使用置信位来确定何时替换条目。

2)子例程返回是使用一个链接堆栈来预测的，每个线程一个。每当扫描一个分支和链接指令时，下一条指令的地址就会被下推到该线程的链接堆栈中。当扫描到分支到链接指令时，将弹出链接堆栈。POWER7 链接堆栈允许在扫描分支和链接指令然后由于程序顺序中较早出现的错误预测分支而刷新的情况下保存一个推测条目。在 ST 和 SMT2 模式中，每个线程使用一个 16 个条目的链接堆栈。在 SMT4 模式中，每个线程使用一个八项链接堆栈。

5. 获取数据

数据获取由 LSU 执行，LSU 包含两个对称的 LS 执行管道(LS0 和 LS1)，每个管道都能够在一个循环中执行加载或存储操作。

图 6 显示了 LSU 管道的微体系结构，它包含几个子单元，即 LS 亚仁和执行,SRQ 和存储数据队列(SDQ)LRQ,加载小姐队列(LMQ),地址转换机制,包括 D-ERAT,ERAT 小姐队列,段后备缓冲区(SLB)和 TLB 和 L1 D-cache 数组的支持组预测和数据目录(DDIR)数组,和数据预取请求队列(PRQ)引擎。

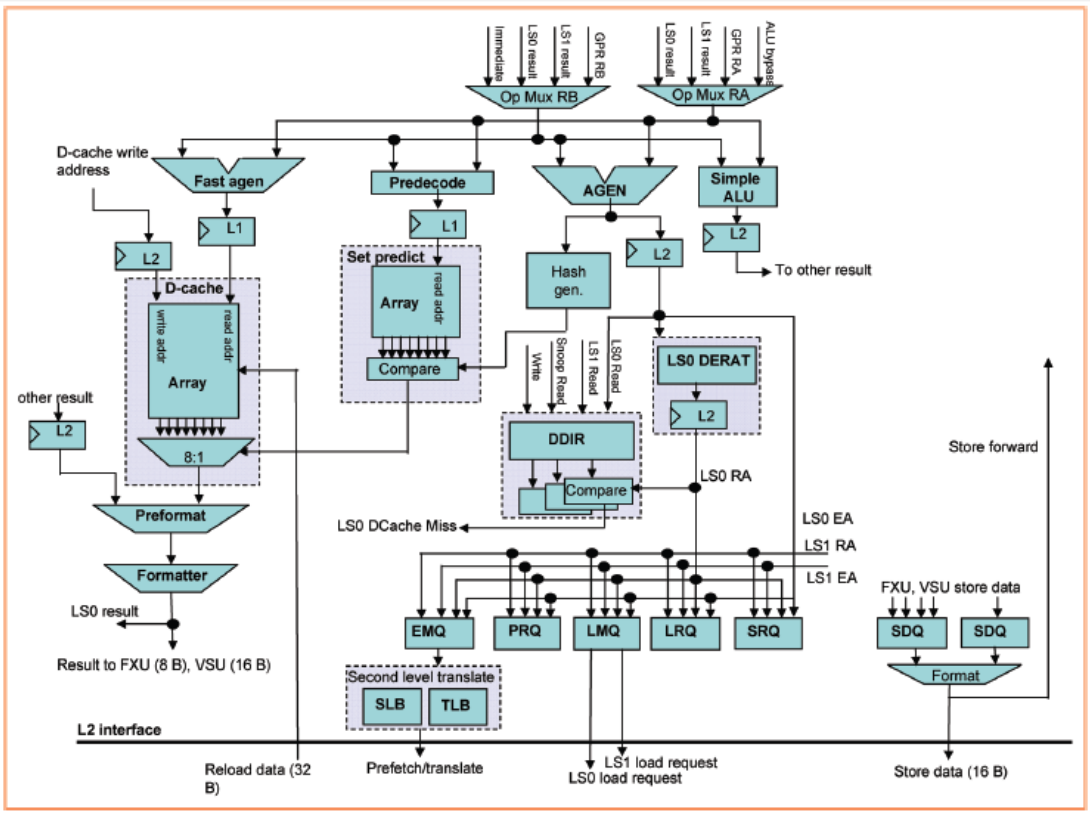

Figure 6

LSU microarchitecture (LS0 pipe shown).

6. 地址翻译

在 PowerPC 架构中，程序是使用 64 位有效地址(EAs)(32 位寻址模式下的 32 位)编写的。在程序执行期间，EAs 通过第一级转换转换为 46 位的实际地址，这些地址用于缓存和内存子系统中的所有寻址。第一级转换由两个 64 入口的 D-ERAT 缓存和一个 64 入口的 IERAT 组成。如果 ERAT 缓存(数据或指令)出现错误，将调用第二级转换来生成转换。第二级转换由一个 32-entry-per-thread 的 SLB 和一个 512-entry 的 TLB 组成，所有活动线程共享它。首先使用段表将有效地址转换为 68 位虚拟地址，然后使用页框表将 68 位虚拟地址转换为 46 位真实地址。而段表和页框表驻留在内存,一个 32-entry-per-thread SLB 是防止条目段表维护翻译有效的虚拟地址,和 512 条目 TLB 维护保持最近使用从页面表条目翻译从虚拟到真实的地址。POWER7 支持两个段大小，即， 256mb 和 1tb，以及四个页面大小，即， 4 KB, 64 KB, 16 MB 和 16gb。

7. L1 Data Cache

POWER7 包含一个专用的 32 kb 的 8 路集关联存储 L1 D-cache。高速缓存线的大小是 128 字节，包含 4 个扇区，每个扇区 32 字节。

L1 D-cache 有三个端口，两个读端口(用于两个加载指令)和一个写端口(用于存储指令或缓存线路重载)。

L1 D-cache 由 4 个按数据字节组织的物理宏组成，每个宏根据 EA 位划分为 16 个块，总共 64 个块。

L1 D-cache 是一种直通式设计;所有的存储都被发送到 L2 缓存，不需要 L1 退出。

L1 D-cache 使用 EA 位进行索引。L1 D-cache 目录使用二叉树 LRU 替换策略。

8. 定点运算单元（Fixed-point Unit）

FXU 由两个相同的管道(FX0 和 FX1)组成。如图 7 所示，每个 FXU 管道由一个多端口 GPR 文件组成;一种算术和逻辑单元(ALU)，用于执行加、减、比较和陷阱指令;一个旋转器来执行旋转、移位和选择指令;一个计数(CNT)领先零单位;位选择单元(BSU)用于执行位 PM 指令;一个分配器(DIV);一个乘数(乘);和一个杂项执行单元(MXU)来执行人口计数、奇偶校验和二进制编码的十进制辅助指令。FXU 管道中本地的所有 SPRs 都存储在 SPR 中。某些资源(如 FX XER 文件)在两个管道之间共享。

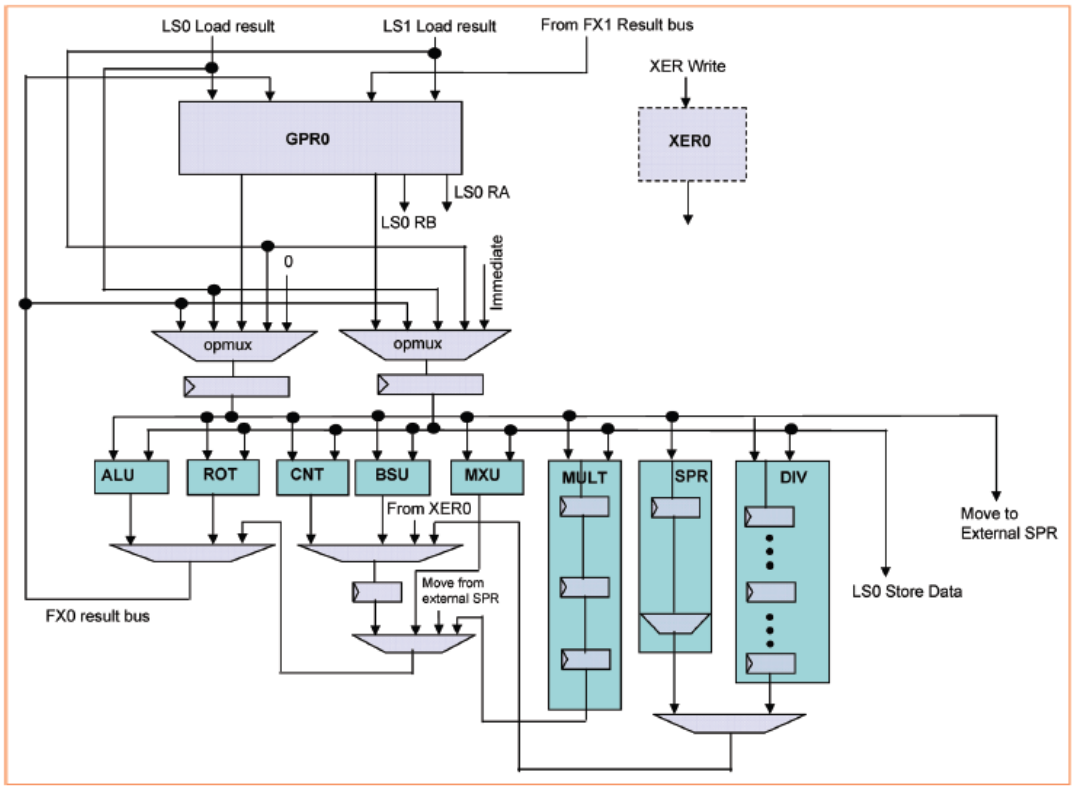

Figure 7

IBM POWER7 FXU overview (FX0 pipe shown).

9. 向量和标量指令执行

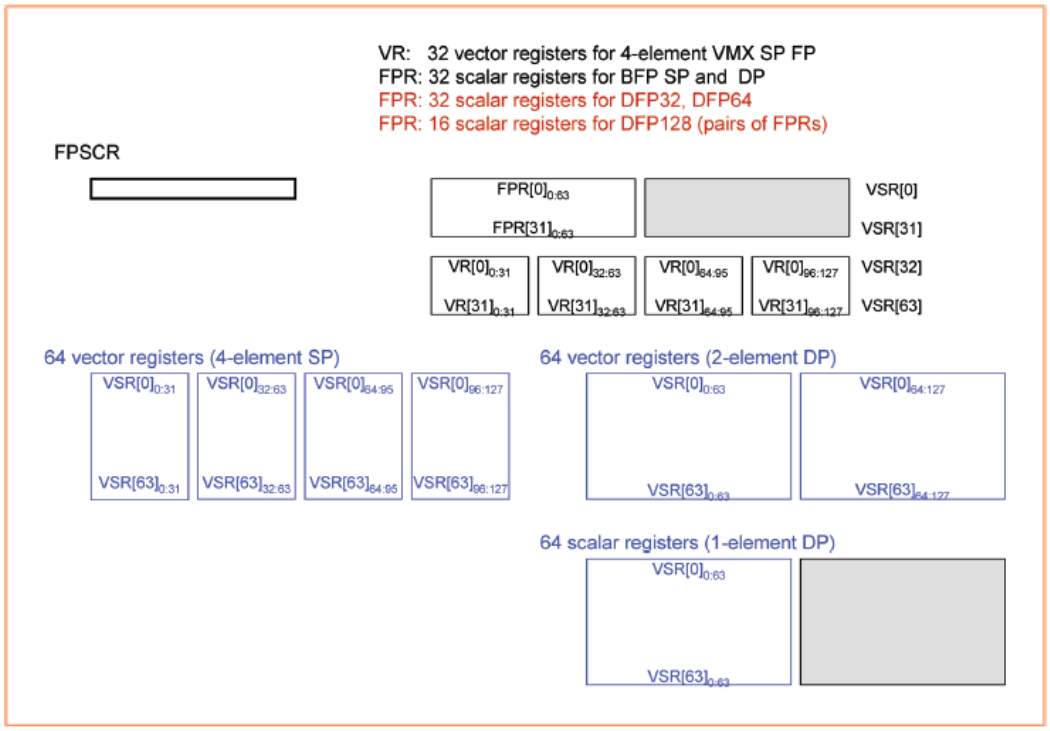

Figure 8

FPRs and VMX VRs as part of the new VSRs.

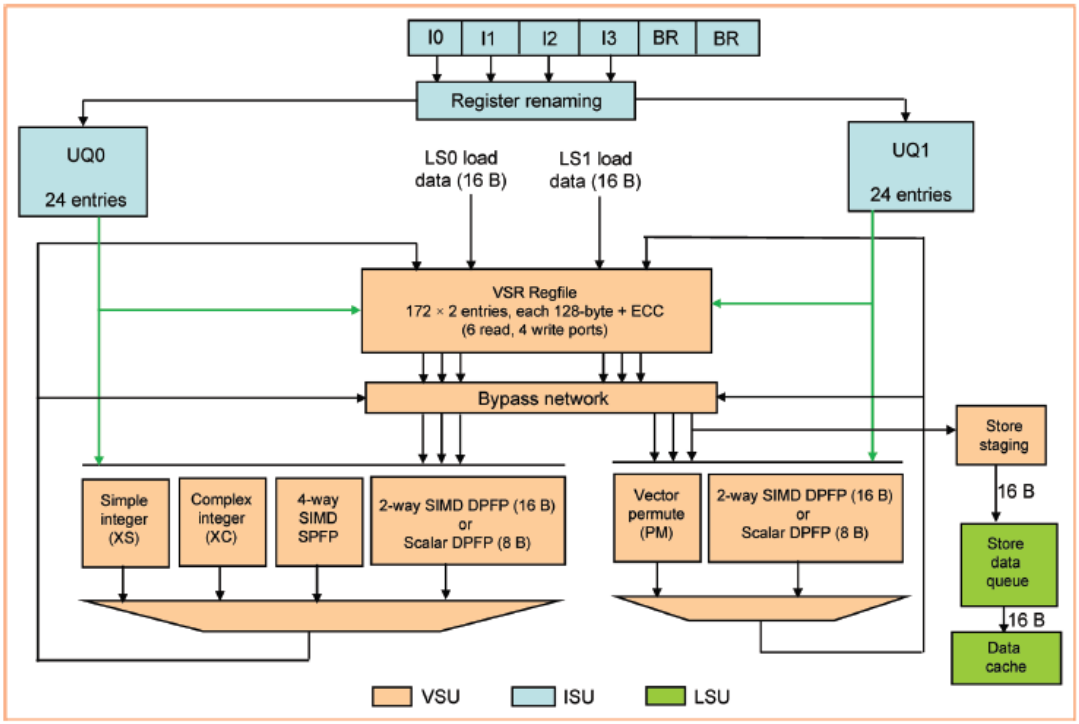

Figure 9

VSU pipeline diagram. (DPFP: double-precision floating point.)

10. 缓存层次结构

POWER7 的缓存层次结构针对上一代 POWER6 处理器进行了重新优化，以适应内核和芯片的以下更改:

1） 核心从高频设计重新流水线到功率/性能优化设计点。

2） 从主有序指令调度策略更改为主无序指令调度策略。

3） 每个核心从 2 个线程增长到 4 个线程。

4） L1 d 缓存大小从 64 KB 减少到 32 KB, L1 缓存访问时间也有所减少。

5） 从两个核心到八个核心。

如表 1 所示，POWER6 处理器中的 64-KB L1 D-cache 和 4- mb L2 cache 已经被 32-KB L1 D-cache、256-KB L2 cache 和 4- mb 本地 L3 区域所取代。

Table 1 Comparison of POWER6 and POWER7 cache hierarchies.

POWER6 (assuming 5-GHz core)	*POWER7 (assuming 4-GHz core)*
	32 KB store-through L1 D-cache 0.5ns latency, 192 GB/s private
64 KB store-through L1 D-cache 0.8ns latency, 80 GB/s private	256 KB store-in L2 cache 2.0-ns latency, 256 GB/s private
4 MB store-in L2 cache ~5.0-ns latency, 160 GB/s private	4 MB partial victim local L3 region ~6.0-ns latency, 128 GB/s private
32 MB victim L3 cache ~35-ns latency, 80 GB/s shared by 2	32 MB adaptive victim L3 cache ~30-ns latency, 512 GB/s shared by 8

如表 2 所示，POWER7 L2 和 L3 缓存支持与 POWER6 设计点相同的 13-cache-state 协议。虽然 POWER7 没有添加新的缓存状态，但是支持新的一致操作。

Table 2 IBM POWER7 cache states.

State	*Description*	*Authority*	*Sharers and scope*	*Source data*	*Data cast-out*	*Scope cast-out*
I	Invalid	None	N/A	N/A	N/A	None
ID	Deleted, do not allocate	None	N/A	N/A	N/A	None
S	Shared	Read	Yes, scope unknown	No	No	None
SL	Shared, local data source	Read	Yes, scope unknown	At request	No	None
T	Formerly MU, now shared	Update	Yes, probably global	If notified	Yes	Required, global
TE	Formerly ME, now shared	Update	Yes, probably global	If notified	No	Required, global
M	Modified, avoid sharing	Update	No	At request	Yes	Optional, local
ME	Exclusive	Update	No	At request	No	None
MU	Modified, bias toward sharing	Update	No	At request	Yes	Optional, local
IG	Invalid, cached scope-state	None	N/A, probably global copies	N/A	N/A	Required, global
IN	Invalid, scope predictor	None	N/A, probably local copies	N/A	N/A	None
TN	Formerly MU, now shared	Update	Yes, local	If notified	Yes	Optional, local
TEN	Formerly ME, now shared	Update	Yes, local	If notified	No	None

L2 cache

私有的 256kb POWER7 L2 缓存使用 128 字节行，是八路集关联。它由一个集中的控制器组成，有两个地址散列缓存数据数组。L2 高速缓存中的一些结构以核心频率运行，而其他一些结构以核心频率的一半运行。此后，我们将使用术语“核心循环”和“2:1 循环”来区分它们。进入核心的接口、数据流和缓存目录 SRAM 单元以核心频率运行，而地址流、控制逻辑和缓存数据数组 SRAM 以核心频率的一半运行。这种独特的核心循环结构和 2:1 循环结构的混合优化了一致性和数据带宽以及延迟，同时减少了能源、面积和线路拥塞。

L3 cache

共享的 32 mb POWER7 L3 缓存由 8 个 4-MB L3 区域组成。每个 L3 区域使用 128 字节的行，是 8 路集关联。给定的 L3 区域包括一个集中控制器，其中有四个地址散列的 eDRAM 缓存数据库和四个地址散列的 SRAM 目录块。本地 L3 区域与与给定核心相关联的 L2 缓存紧密耦合。所有 L3 结构都在 2:1 循环中运行(在 L2 缓存部分中定义)。

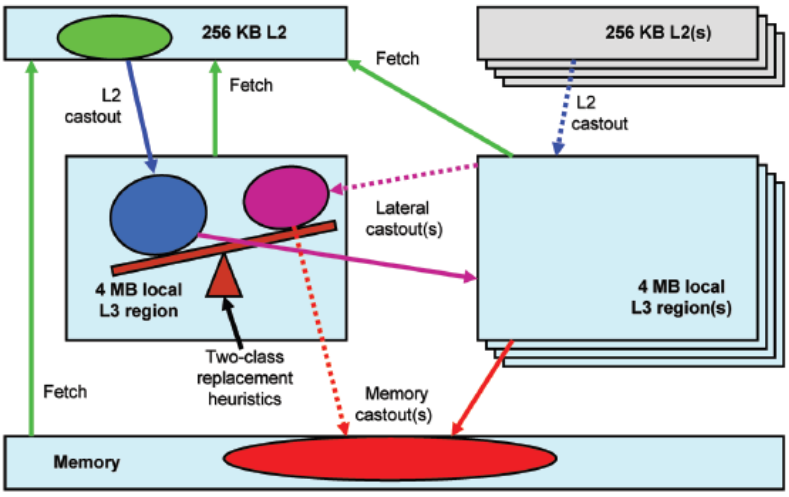

Figure 10

L3 region behavior.

总结

POWER7 延续了 POWER 处理器的创新传统。这款第七代芯片将平衡多核设计、eDRAM 技术和 SMT4 加入到电力创新组合中。POWER7 芯片的核心数是 POWER6 芯片的 4 倍，线程数是 POWER6 芯片的 8 倍，每个周期的失败次数是 POWER7 芯片的 8 倍。这种平衡的设计允许处理器从单个套接字低端刀片服务器扩展到具有 32 个套接字、256 个内核和 1024 个线程的高端企业系统。与上一代 POWER6 处理器相比，这种新的创新设计为每片芯片提供了 4 个以上的性能提升。