计算技术研究所，微处理器中心

论文阅读报告　其2
IBM Power 7 Multicore server processor

谭弘泽
201828013229048

October 18, 2018

Contents

1 摘要 **1**

2 简介 **1**

3 POWER7核 **2**

4 指令获取 **3**
4.1 分支预测 3

5 指令重排单元 **3**

6 数据获取 **4**
6.1 访存执行 4
6.2 访存重排 4
6.3 地址翻译 4
6.4 L1数据高速缓存结构 4
6.5 把握加载缺失（load miss） 4

7 定点单元 **5**
7.1 向量和标量指令执行 5

8 高速缓存层次 **7**
8.1 L2高速缓存 7
8.2 L3高速缓存 8

9 内存子系统 **9**

10 I/O子系统 **9**
10.1 片上互联 9
10.2 多片互联 9
10.3 集群互联 10

11 可靠性，可用性和服务能力 **11**

12 总结 **11**

13 附录A-单词翻译对照表 **11**

1 摘要

IBM POWER处理器是当今世界上主流的RISC微处理器，有着超过20年创新和实践的丰富历史积淀。在这篇论文中，作者描述了POWER7处理器的关键特征。在这个芯片上的是一个8核处理器，每核都支持4路并发多线程操作。通过11层金属的IBM的45nm SOI（Silicon-on-Insulator，绝缘衬底上的硅）技术制造，芯片包含超过十亿个晶体管。为提高无论是单线程还是多线程情形下的响应时间性能或者吞吐率，处理器核和高速缓存受到了显著加强。内存子系统包含三级片上高速缓存，其中嵌入SOI的动态随机访问存储器（DRAM）设备作为最后一级高速缓存。文章将会讨论。一种使用带缓冲的double-data-rate 3(DDR3)DRAM的新内存接口以及可靠性，可用性和服务能力。

2 简介

使用了嵌入式动态随机访问存储器（eDRAM,embedded dynamic random access memory）来做高速缓存。8核每核4线程。宣传性能相比上一代有较大提升（吸引别人来买新产品）。
下图是大致空间分布，每个核有独立的L2Cache，并且L3Cache会有一块离自身较近可以相对高速地进行访问的区域。最中间是L3Cache和片上互联，两侧有内存控制器，中间有本地SMP连接和远程SMP与I/O连接。

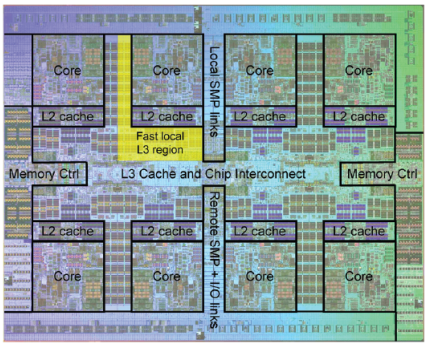

Figure 1

Die photo of the IBM POWER7 chip.

3 POWER7核

每个POWER7处理器核的平面图如下：

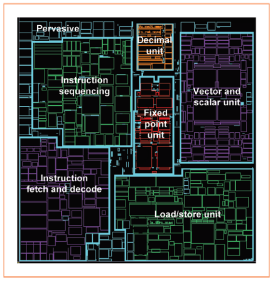

Figure 2

IBM POWER7 processor core floorplan.

其中几大块分别是取值译码、指令重拍、定点单元、浮点单元、向量和标量单元、访存（Load/Store）单元。
POWER7处理器核的流水线流程。

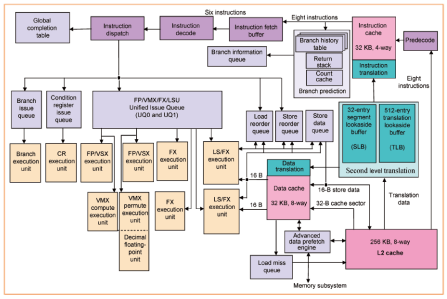

Figure 3

IBM POWER7 processor core pipeline flow.

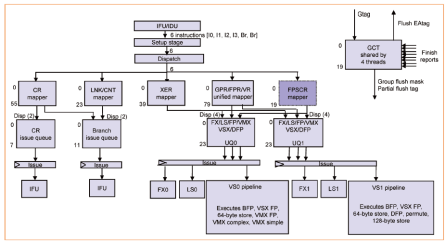

Figure 4

IBM POWER7 ISU overview.

会去支持一些媒体指令和向量指令。POWER7也会进行面积功耗优化等。

4 指令获取

取指和译码的流水线，很详细

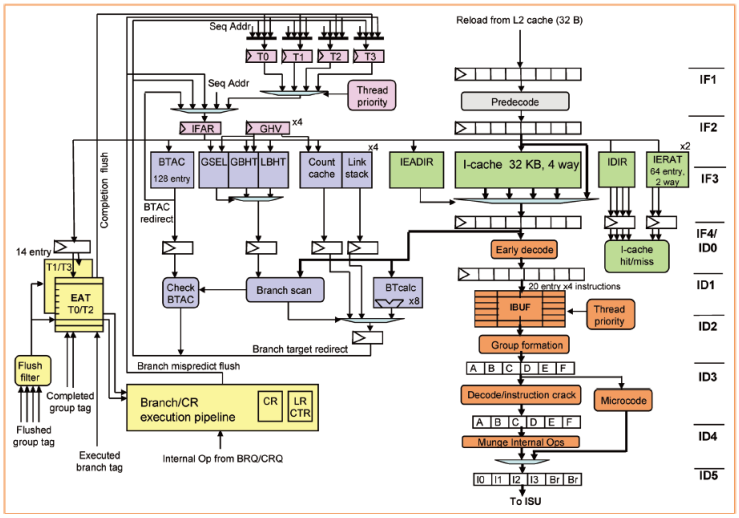

Figure 5

IBM POWER7 instruction fetch and decode pipe stages.

4.1 分支预测

分开进行跳转和不跳转的预测。
对于间接跳转使用缓存来预测跳转地址。

5 指令重排单元

同一个线程对多译码2个跳转指令和4给非跳转指令。

6 数据获取

LS级流水线的流程图，也很详细

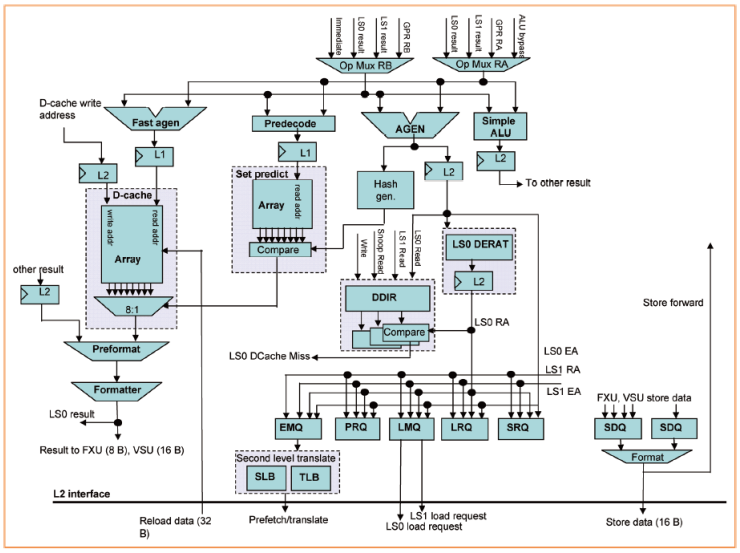

Figure 6

LSU microarchitecture (LS0 pipe shown).

6.1 访存执行

L1 数据高速缓存大小是32KB

6.2 访存重排

LSU有两个主要队列SRQ和LRQ分别负责写和读的请求。

6.3 地址翻译

每线程有32项的SLB和512项的TLB。

6.4 L1数据高速缓存结构

32KB 8路组相联。有两个读口和一个写口

6.5 把握加载缺失（load miss）

当L2Cache返回结果时，优先处理。
LMQ最多能合并两个加载指令，最多能支持8个缺失同时等待。

7 定点单元

图也很详细。

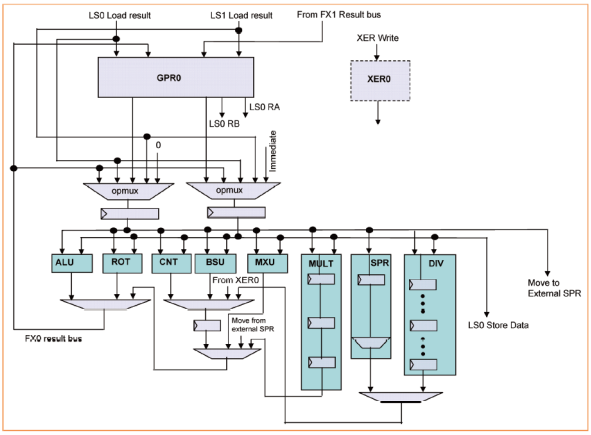

Figure 7

IBM POWER7 FXU overview (FX0 pipe shown).

7.1 向量和标量指令执行

图同样很详细

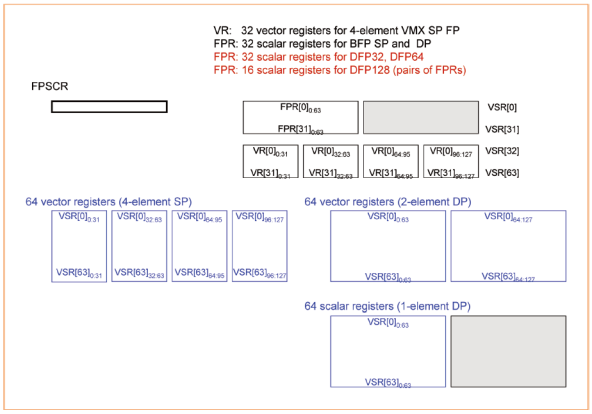

Figure 8

FPRs and VMX VRs as part of the new VSRs.

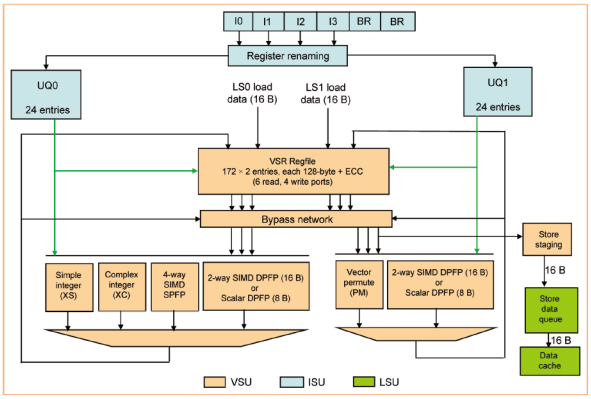

Figure 9

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

各个模块的框图都很详细。但是除非已经记着很多框图，否则直接看这些框图几乎也不会有很多收获。这些图最大的用途或许是用来对比上一代的结构。例如当读者试图从POWER 6升级成POWER 7时，把两篇论文拿出来，用找不同一样的方式对比地看两个图，能对变化有所体会。

此外，虽然已经很详细了，但是当然地（即使不算版权问题），只看这些图离着能把这些系统复现出来也很有距离。

8 高速缓存层次

高速缓存层次的参数如下表：

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

展示了一些延迟，带宽，容量，组相联路数等信息，并且和上一代进行了对比。
下表介绍高速缓存状态，这些状态组成状态机共同维护一致性协议。

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

8.1 L2高速缓存

主要在降低延时和降低能耗上下了一些功夫。

8.2 L3高速缓存

L3高速缓存的容量和工作模式如下图:

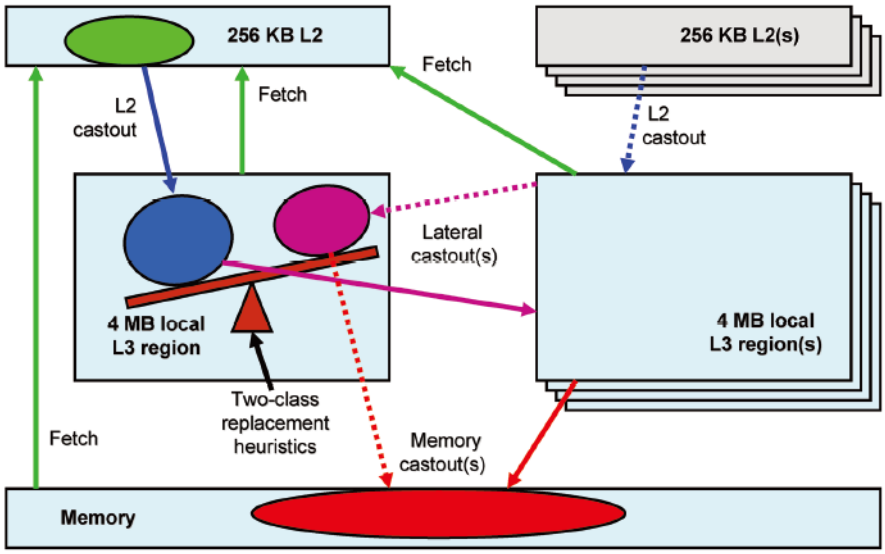

Figure 10

L3 region behavior.

L3的妹4MB的区域会优先只让每个L2高速缓存使用。

9 内存子系统

有基于循环冗余校验(CRC，cyclic redundancy check)的错误检测。
使用错误矫正码（ECC，Error Correction Code）来降低错误率。
下图是处理器到缓存通道和缓存到DRAM通道在不同速度等级下的原始峰值带宽。

Table 3 IBM POWER7 memory configurations.

DDR3 DRAM frequency	*Channel frequency*	*Speed ratio*	*Raw channel bandwidth*	*Raw DRAM data bandwidth*
800 MHz	4.8 GHz	6:1	135 GB/s	102 GB/s
1,066 MHz	6.4 GHz	6:1	180 GB/s	137 GB/s
1,333 MHz	5.3 GHz	4:1	149 GB/s	171 GB/s
1,600 MHz	6.4 GHz	4:1	180 GB/s	205 GB/s

10 I/O子系统

10.1 片上互联

片内作为一个SMP系统有着较高的带宽，和较低的延迟。

10.2 多片互联

下图描绘第一级结点结构，最多结合4枚POWER7芯片。

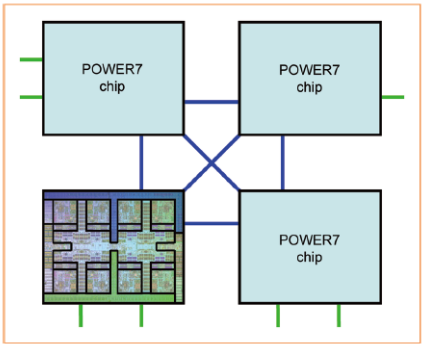

Figure 11
IBM POWER7 first-level nodal topology.

在第一级结点结构的基础上，可以组建更复杂，结点数目更多的网络，如下图:

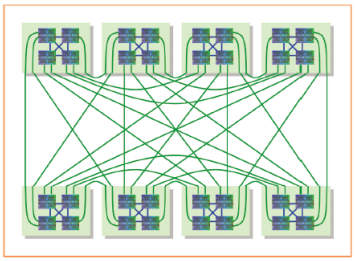

Figure 12

IBM POWER7 second-level system topology.

最多连接32枚POWER7芯片并且组成一个256路的SMP系统。

10.3 集群互联

如下图，每四个POWER7芯片互联以后借助于集群芯片可以组件更大的网络。

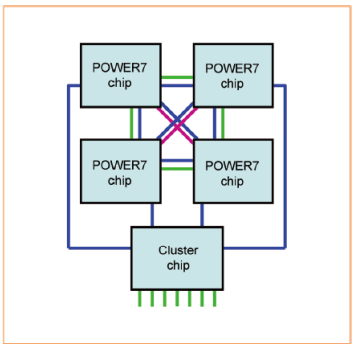

Figure 13

Cluster system first-level coherent nodal topology.

不同的系统接口不太一样，如下表：

Table 4 IBM POWER7 standard system and cluster system interfaces.

Logical interface	*Standard system*	*Cluster system*
POWER7 to POWER7 nodal interconnect	Use three of four POWER7 on-node SMP links	Use three of four POWER7 on-node SMP links
POWER7 to POWER7 nodal interconnect 2× bandwidth expansion	n/a	Use two POWER7 off-node SMP links and both GX I/O links
POWER7 to cluster chip nodal interconnect	n/a	Use fourth POWER7 on-node SMP link
System interconnect	Use one or two POWER7 off-node SMP links	Use seventh cluster chip off-node SMP/cluster links
I/O subsystem attachment	Use POWER7 GX I/O links	Use cluster chip PCI interfaces
Memory channels	Use eight POWER7 channels	Use eight POWER7 channels

11 可靠性，可用性和服务能力

POWER7有上一代POWER6所有的RAS（Reliability,availably,and serviceability，即可靠性，可用性和服务能力）特性。特定系统还可以通过内存镜像来进一步提高可靠性。

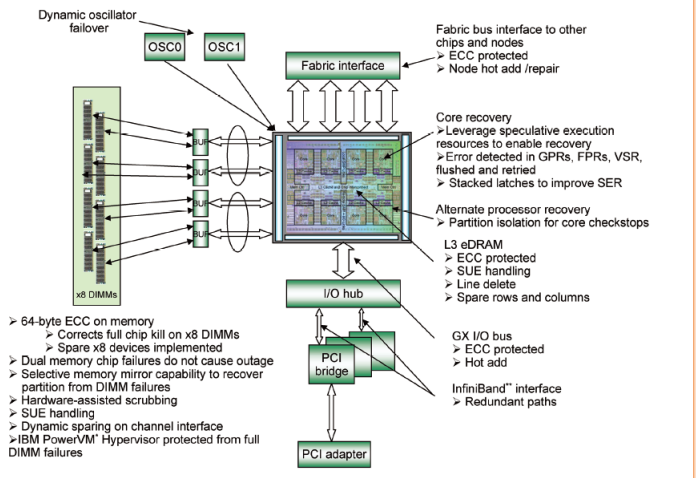

Figure 14

IBM POWER7 reliability and availability features.

12 总结

POWER7继承了POWER系列处理器的创新传统。这第7代芯片在POWER系列的创新中加入了平衡性多核设计，eDRAM技术，和SMT4。POWER7芯片与POWER6相比有4倍的核数和8倍的线程数，同时也有8倍的每周期浮点操作数。平衡的设计允许处理器适应从单插槽低端片到32插槽256核1024线程高端事业级系统的规模。这崭新的创新性设计相比于前代的POWER6处理器提供了超过4倍的单片性能。

13 附录A-单词翻译对照表

英文	中文
fabricate	制造，捏造，装配
insulator	绝热体，绝缘体，隔离者，隔离物
floorplan	平面图
pervasive	普遍的；n.遍布
sequencing	定序，重排
portfolio	公文包,代表作品集，证券投资组合
enterprise	事业，计划