

基于独占式访存调度的片上系统电源门控方法^①

刘 苏^{②*} 苏孟豪^{*} 苏 文^{****}

(* 中国科学院计算技术研究所 北京 100190)

(** 中国科学院大学 北京 100049)

(*** 龙芯中科技术有限公司 北京 100190)

(**** 中国石油北京油气调控中心 北京 100007)

摘 要 为减少片上系统(SoC)中访存模块的漏电功耗,在分析访存行为的基础上,提出了一种基于独占式访存调度的电源门控方法。该方法根据访存模块对延迟容忍度较高和功耗优化空间很大的特点,利用访存调度器将访存模块的请求集中于窗口中,并在窗口关闭时对访存模块及其访存通路使用电源门控。相对于已有方法,该方法不但增加了电源门控的可用范围,同时可更多地降低漏电功耗。实验结果表明,该方法可以在不影响访存性能的情况下减少访存模块约 70% 的漏电功耗。

关键词 独占式,访存调度,电源门控,片上系统(SoC),漏电功耗

0 引 言

芯片漏电功耗会随着芯片工艺的升级大幅增加,已成为芯片总体功耗的重要部分。在 32 纳米和 22 纳米工艺节点下,漏电功耗占总功耗的 16.9% 至 52.7%^[1]。电源门控(power gating)^[2]是一种广泛使用的削减漏电功耗的技术,传统的电源门控方案在软件不需要使用一个电路模块时就将其电源关闭,但在一个模块的运行过程中,仍有结构和微电路层次的空闲,这些空闲时间也可以用来实现电源门控。为了进一步降低漏电功耗,一些运行时电源门控技术被提了出来。Mistry 等^[3]提出一种亚时钟电源门控技术,该技术在单个时钟周期内对组合逻辑电路进行关闭和打开,使组合逻辑电路只在需要工作时打开。Youssef 等^[4]提出一种基于预测的电源门控技术,当中央处理器(central processing unit, CPU)中某个功能部件的空闲时间超过阈值时,就将其关闭。Jeong 等^[5]提出一种适用于多核 CPU 的电源门控方法,当某个核心因较长访存延迟而停止工

作时,就对其进行电源门控。Wang 等^[6]提出一种运行时的电源门控方法,当图形处理单元高速缓存暂停时就将其关闭。Lin 等^[7]提出一种运行时的机制来预测每个独立访存模块的访存延迟周期数,并根据预测的访存延迟和阈值来决定是否使用电源门控。本文通过对片上系统(system on chip, SoC)访存行为的分析,发现访存模块对访存延迟有较大的容忍度,其功耗优化的空间也很大。据此,我们提出了一种基于独占式访存调度(exclusive memory access scheduling, EMS)的电源门控方法。访存调度器在满足每个访存模块的服务质量(quality of service, QoS)的前提下,将其分散的访存请求集中到调度周期中一定的窗口内,并使其在窗口中独占内存总线,从而将散乱的访存延迟和空闲时间整合为大段的空闲时间。在窗口关闭时对访存设备及其访存通路进行电源门控,从而达到降低漏电功耗的目的。实验结果表明,该方法可以减少访存模块约 70% 的漏电功耗且不会带来带宽损失。

① 国家“核高基”科技重大专项课题(2009ZX01028-002-003, 2009ZX01029-001-003, 2010ZX01036-001-002), 国家自然科学基金(60921002, 61003064, 61050002, 61070025, 61100163, 61133004, 61173001, 61232009)和 863 计划(2012AA010901, 2012AA011002, 2012AA012202, 2013AA014301)资助项目。

② 男, 1989 年生, 博士生; 研究方向: 计算机体系结构; 联系人, E-mail: liusu - cpu@ict. ac. cn
(收稿日期: 2013-05-14)

1 背景与动机

1.1 片上系统访存行为分析

片上系统中除 CPU 外的主要访存模块包括显示控制器、图形处理单元、高清视频解码单元等,这些模块在访存敏感类型和访存具体行为上都有一些相似性。

由于多媒体应用的特点,图像信息和声音信息等多媒体数据总是按照一定的帧数进行输出。由于引入预取逻辑和内部存储对数据进行缓存,这些模块对访存延迟要求不高,只需在缓存使用完之前返回足够数据即可^[8]。某些模块并不内建缓存^[7],当访存数据未返回时,该模块就不会工作,这类模块带宽和延迟需求均较低,但仍需要保证一定带宽。总的来说,片上系统的功能特点决定了主要访存模块都属于带宽敏感、延迟不敏感的访存类型。

从访存的具体行为来看,多媒体数据具有较强的局部性。图像和声音信息通常按照顺序存储在一块连续的内存空间中,在对其进行访问时,可以在同一个页中连续命中。单个模块通常按顺序访问图像或声音,体现出较好的局部性;但多个模块同时进行

访存时会产生大量访存冲突。访存冲突的原因是访问同一个 bank 的两个访存请求的页地址不同,导致该 bank 需要频繁地打开和关闭页,带来大量的预充电和行选代价,影响了访存的性能。

图 1(a)展示了来自两个不同访存模块的对同一 bank 但行地址不同的读请求的访存波形,每个访存模块按照连续地址发出两个访存请求独占一段时间的内存总线。图 1(b)展示了与图 1(a)相同个数的读请求的访存波形,区别在于单个访存模块的请求被错开。与图 1(a)的显著区别在于每一次访存请求都需要进行预充电和行地址选择,在 5/5/5/15 的内存时序参数下,对各自模块来说相同的访存序列因为调度方式的不同而分别消耗了 41 和 73 个时钟周期。

因独占或轮流两种不同调度方式产生了明显的区别,在 5/5/5/15 的内存时序参数下,当独占时间较长时,访存带宽最多会相差接近 3 倍。

综上所述,多媒体 SoC 的访存行为属于带宽敏感、延迟不敏感的访存类型^[9],虽然单个模块都具有良好的访存局部性,但多个模块同时访存会导致访存冲突而降低了访存带宽。

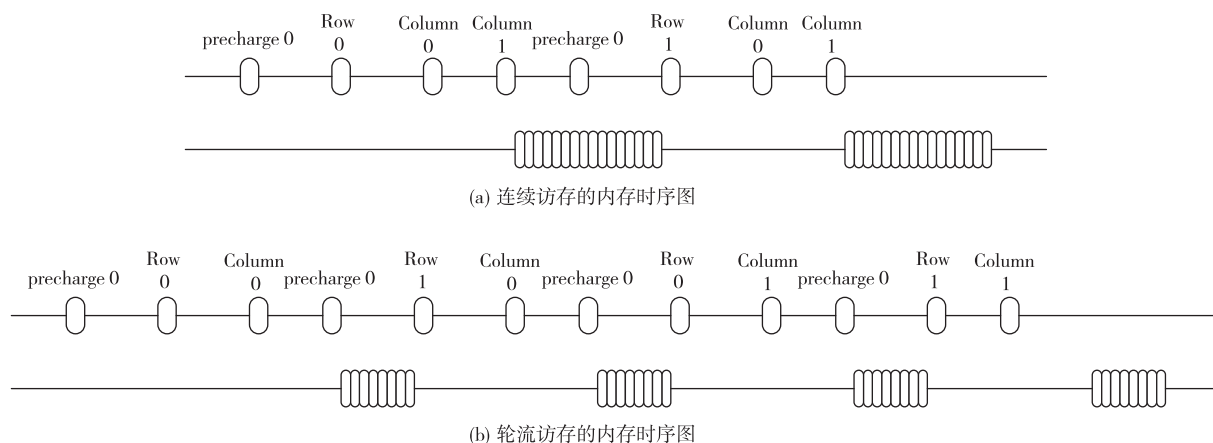


图 1 连续访存与轮流访存的内存时序对比

1.2 访存模块的功耗优化空间

共享访存的特点决定了访存模块不会一直占据访存总线,其真实工作时间应与其带宽需求相对应,访存延迟和访存之间的间隔时间都属于空闲时间,所以虽然访存模块在持续工作,其中仍包含可观的空闲时间可用于功耗优化。

Lin 等^[7]提出了一种运行时的访存感知电源门控(memory access aware power gating, MAPG)机制来预测每个访存模块的访存延迟周期数,并根据预测

的访存延迟和阈值来决定是否使用电源门控,当访存延迟大于阈值时,就将对模块整体使用电源门控技术。

MAPG 可以节省访存模块由于等待数据而产生的空闲时间内的漏电功耗,但也有些显著的不足之处:一是只能用于访存请求未返回就无法工作的模块,类似于显示控制器的需要一直工作的模块则并不能适用这种方法;二是假设内存控制器使用先来先服务的策略,当访存控制器使用其他调度策略

时难以得到准确的访存延迟预测值;三是对预测访存延迟的需求导致模块不能发出多个突发(out-standing)的访存请求,只能收到数据后再发出下一个请求。

为了能够适应变化的访存状况,使用先进先出(first in first out, FIFO)队列作为缓存的模块总是希望填满 FIFO,一旦 FIFO 空出了一次访存请求的空间,就发出一次请求。由于功能模块消耗或产生数据的带宽低于访存的突发带宽,因 FIFO 行为而导致的访存行为将是零散的,零散的访存请求不利于访存调度。另一方面,当 FIFO 中数据较少时,访存模块将发出连续的访存请求,连续的访存请求可以更好地利用访存带宽。

带宽敏感设备的实际访存时间与其带宽需求有正比关系。当访存局部性很好时,访存模块的实际工作时间比例应接近于带宽需求比例。但在真实芯片中,由于访存延迟和访存冲突等情况的存在,访存模块的实际工作时间往往远超过其理论工作时间,从而浪费了潜在的功耗优化空间。

2 基于独占式访存调度的电源门控

2.1 独占式访存调度方法

本文提出了一种独占式的访存调度方法,该方法通过在一定的调度周期内使每个访存模块独占总线以提高访存带宽,通过集中调度的方式降低芯片访存的不确定性并最大程度利用模块内缓存空间。

独占式访存调度利用多媒体 SoC 中访存模块对延迟的容忍性,在一定的调度周期内为每个访存模块开启一个窗口时间,只有在这个窗口之内该访存模块才能发出访存请求。这样便将来自同一个访存模块的访存请求集中到同一段时间之内,并在这一段时间内独占内存总线。

图 2 所示为独占式访存调度器的结构。访存请

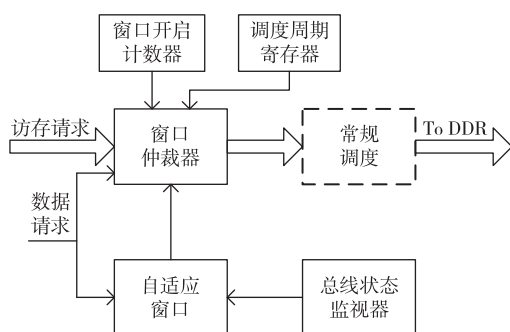


图 2 调度器结构

求进入窗口仲裁器,窗口仲裁器将决定是否响应请求,得到响应的请求再进入常规调度器中,进行轮转或其他方式的调度。窗口仲裁器通过自适应窗口和各个模块的数据请求来决定是否响应请求。

自适应窗口是独占式访存调度器的核心模块,可通过软件配置和硬件自适应两种方法对窗口开启时间进行调整。由于 CPU 的访存特性属于延迟敏感型,所以 CPU 的访存窗口将一直处于开启状态和最高优先级状态,以保证 CPU 的访存请求可以尽快返回。

当带宽敏感模块的配置变化时,软件可以根据新的带宽需求配置窗口开启时间。令调度周期为 T_s ,窗口开启时间为 T_w ,需求带宽为 B_n ,访存系统总带宽为 B_t ,则

$$T_w = T_s \times \frac{B_n}{B_t} \quad (1)$$

为软件配置的窗口开启时间。

为了提高带宽利用效率,我们实现了一个自适应的窗口机制,自适应窗口可以根据访存状况动态调整对不同 IP 开启的窗口时间,使窗口开启的时间接近于模块的需求带宽,减少带宽的浪费。

我们对访存延迟和访存带宽等总线状态进行监视,当发现某模块在独占访存时存在空闲周期时,说明分配给其的带宽大于需求带宽,就将其窗口缩小,达到整体上缩减带宽分配的目的。在实际应用中,窗口开启时间减小的量为该调度周期内的空闲周期数的一半,这样可以得到平滑的削减效果。

当产生数据请求时,说明长期的带宽分配偏小,FIFO 的容量逐渐消耗并无法补齐。所以有数据请求时,就将对应窗口扩大。窗口开启时间扩大的量可以选择整个空闲周期,或者每个调度周期扩大固定值。缩小和扩大窗口的机制使带宽分配动态调整,达到自适应的效果。

由于独占式访存调度是一个主动式的调度器,在达到更大调度力度的同时,也需要各访存模块反馈访存状态的信息。我们实现了一个数据请求的机制,当访存模块需求更多数据时即发出数据请求,数据请求通常可由模块内的计数器或 FIFO 的剩余容量得到。当某模块发出数据请求时,其对应窗口的使能信号打开,立即对访存请求进行响应。这样可能出现两个或多个窗口同时打开的情况,会导致访存冲突的增加,但仍将优于所有窗口同时打开的情况。数据请求机制可以实现访存模块与调度器的交互,防止出现数据饥饿。

调度周期反映了访存调度的粒度。由于每个模块在一个调度周期内的访存请求集中于一个较短的时间内,该模块的缓存(片内缓存或内存)大小需要保证其在调度周期内不产生数据饥饿。较大的调度周期可以减少访存冲突带来的开销,同时较大的连续工作和空闲时间可以给低功耗措施更大的施展空间。由于独占式访存调度中调度周期对所有模块都相同,所以必须满足缓存忍耐程度最低的模块的访存需求。在 SoC 设计中,设计者可以对各个模块的缓存大小进行调整,使其在最坏状况下的忍耐程序接近,这样可以最大程度利用缓存空间。

综上所述,独占式访存调度通过集中式的管理方法保证带宽服务质量,同时还可将访存模块内部原本为了保证服务质量的缓存用来提高访存性能。

2.2 电源门控方法

我们提出了一种基于独占式访存调度的电源门控(EMS based power gating, EMSPG)方法。该方法基于独占式访存调度,将访存调度器的窗口使能信号作为电源门控的开关信号,当窗口使能有效时处于上电状态,当窗口使能无效时处于下电状态。这种方法不需要预测精确的访存延迟,所以也不需要内存控制器的调度行为进行限制。

由于电源门控的开关管理信号都在访存调度器中,所以针对该 IP 的电源门控还可以扩展到 IP 到访存调度器的访存通路,通常访存通路上会包含一些同步或异步的缓存,这些缓存在没有访存请求时也会导致可观的漏电。与此同时,本文主动式的电源门控方法对于需要持续工作的 IP 也可以进行部分电源门控,电源门控划分的位置为缓存与访存接口的一侧,可以关电的逻辑包括直接内存访问(direct memory access, DMA)模块和总线接口模块等。电源门控方案如图 3 所示。

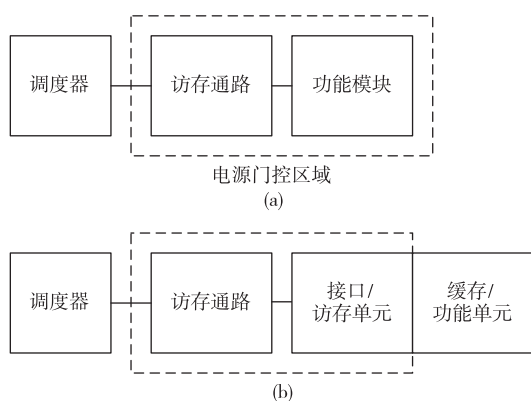


图 3 电源门控结构

图 3 展示了需要持续工作和无需持续工作两种 IP 的电源门控划分结构。需要持续工作的 IP 可以抽象为两个部分:访存控制(DMA)和总线接口(bus interface unit, BIU)、数据缓存(FIFO)和实际功能模块。FIFO 分隔了整个 IP, FIFO 前后的模块功能区别明显。FIFO 前的 DMA/BIU 只在需要访存时工作,有潜在的降低功耗的可能性,但如 1.2 节所述,只要 FIFO 不满就会发出访存请求, DMA/BIU 处于低吞吐、不连续的工作状态。当使用独占式访存调度时,我们就可以将这些模块的工作行为变为高吞吐连续工作一段时间后切换为低功耗状态,从而利用了原本分散的空闲时间。

访存感知电源门控(MAPG)方法的核心在于利用较大的访存延迟,这种访存延迟的主要来源是多个不同访存模块对内存资源的抢占。而独占式访存调度给予每个独立 IP 独享内存资源的机会,在访存局部性较好时,支持 outstanding 的访存延迟可以经常被连续访存结果掩盖,独占访存的方式在降低了访存延迟的同时可以减少甚至消除因为访存延迟而带来的空闲。图 4 直观地给出了两种方法的核心思想。

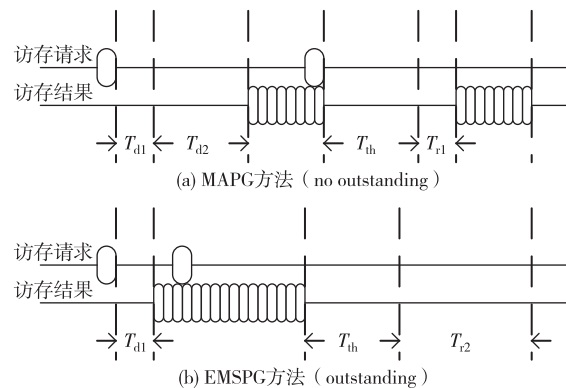


图 4 MAPG 和 EMSPG 方法

图 4 展示了 MAPG 方法与 EMSPG 方法的区别。图中 T_{d1} 表示独立访存时的访存延迟周期, T_{d2} 表示由于访存冲突和访存队列排队带来的额外访存延迟周期, T_{th} 表示电源门控阈值周期, T_{r1} 表示使用 MAPG 方法每个访存请求中可实际节约漏电的周期, T_{r2} 表示使用 EMSPG 方法可实际节约漏电的周期。由于在一定的物理条件下, 漏电功耗 P 是一个定值, 故存在 n 个访存请求时两种方法可节约的总漏电能量为

$$E_{\text{MAPG}} = P \times \sum_n \begin{cases} (T_{d1} + T_{d2} - T_{th}), & (T_{d1} + T_{d2} - T_{th}) > 0 \\ 0, & (T_{d1} + T_{d2} - T_{th}) < 0 \end{cases} \quad (2)$$

$$E_{\text{EMSPG}} = P \times \left(\sum_n (T_{d1} + T_{d2}) - T_{th} - T_{d1} \right) \quad (3)$$

$$E_{\text{OPT}} = P \times \sum_n (T_{d1} + T_{d2}) \quad (4)$$

对比上述公式,可见 EMSPG 方法相对于 MAPG 方法多节约了 $n-1$ 个电源门控阈值时间,实际节约漏功耗效果非常接近于理论最优值。

不支持 outstanding 请求将显著降低访存带宽,所以当今访存模块普遍使用 outstanding 方式访存。对于支持 outstanding 请求的模块,MAPG 方法只能在 outstanding 数量为 1 时进行电源门控,而本文提出的 EMSPG 方法在支持 outstanding 时可以得到更大的有效访存带宽,从而得到更多的空闲时间,达到更好的降低功耗的效果。

3 实验结果和分析

本文实现了 EMSPG 和 MAPG 方法,在高性能多媒体 SoC 的仿真环境下进行了漏功耗节省效果的比较,并比较了不同参数下 EMSPG 方法的效果。

3.1 实验环境

本文的实验平台基于一款高性能多媒体 SoC 的仿真环境,结构如图 5 所示。

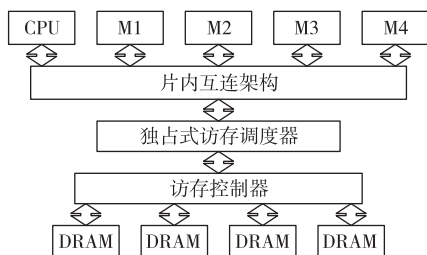


图 5 仿真环境结构

本文的实验环境将模拟多媒体 SoC 全速工作时的访存行为。图 5 中互联结构使用 AMBA AXI 协议;访存控制器使用 Denali 公司 Databahn 访存控制器 IP;DRAM 使用 DDR3-800 内存颗粒,时序参数为 5/5/5/15,数据总线宽度为 64bit,最大带宽为 6.4GB/s。

本文的实验环境包含 5 个访存模块,其中 1 个模拟 CPU 访存行为,在随机的访存间隔内发送地址随机的访存请求,4 个模拟主设备,每个设备均可配置访存地址、outstanding 数量、带宽需求等参数。每个访存模块的访存请求按照地址顺序发出访存请

求,用以模拟访存的局部性,使单个模块的请求最大程度在同一个打开页中命中。同时,每个模块的电源门控阈值可以进行配置,下文中若未特别提及,则电源门控阈值均设为 100 周期。

访存带宽压力的提高会使访存冲突的概率提高,也会影响基于访存的电源门控的效果。据此我们设置了 6 种不同的带宽压力来模拟不同的使用情景,每个模块的带宽需求如表 1 所示,bw1 至 bw6 表示 6 种不同带宽压力的测试用例。

表 1 各模块带宽需求 (MBps)

	bw1	bw2	bw3	bw4	bw5	bw6
M1	50	100	200	400	800	800
M2	50	100	200	400	800	800
M3	50	100	200	400	800	1600
M4	50	100	200	400	800	1600
总带宽	3.1%	6.2%	12%	25%	50%	75%

3.2 漏功耗节约量

图 6 展示了在地址冲突和突发长度为最大值 16 时不同带宽压力下各个主设备使用 EMSPG 方法得到的漏功耗节约比例。从图中可以看出,随着带宽压力越来越大,每个主设备的节约漏功耗比例都在变小。由于 4 个主设备存在优先级关系,优先级最低的 4 号主设备变小的幅度较大。需要指出的是,在总带宽需求为 75% 的测试中,4 号主设备已经出现了数据饥饿的情况,约 0.7% 的时间内 4 号主设备无数据可用。

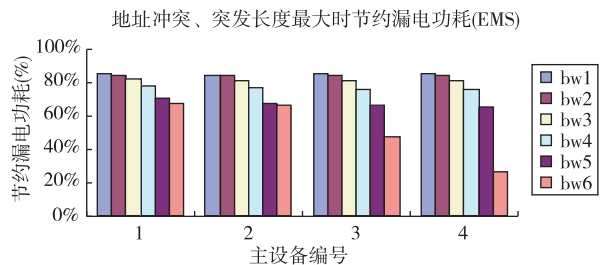


图 6 EMSPG 漏功耗节约比例

图 6 中 EMS 的调度周期为 1024,由于调度周期较小,即使在单个调度周期内只有一次访存,该次访存的延迟和电源门控阈值也占有可观的周期数,导致在最小带宽压力 bw1 下只能得到 85% 左右的功耗节约比例。

图 7 展示了与图 6 对应的采用 MAPG 方法得到

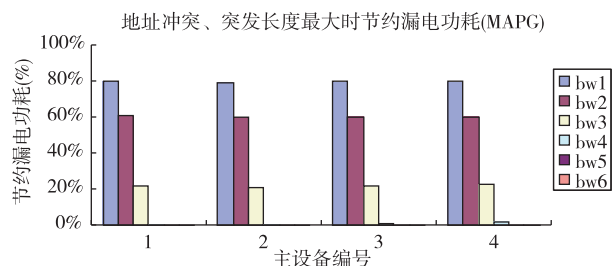


图7 MAPG 漏电功耗节约比例

的漏电功耗节约比例。如前一节所述,MAPG方法难以适用于突发长度大、内存控制器采用调度策略的情境,所以图7中的数据均通过静态方法获取设备在空闲和只有一个访存请求在外时的有效电源门控周期数。

从图7中可以看出,当访存带宽压力增大时,MAPG方法可节约的漏电功耗比例显著下降。当总带宽为25% (bw4测试)时,已经几乎无法节约任何功耗;当带宽压力为50%时,常规的QOS调度方法已经无法为4个主设备提供足够的带宽,产生了较为明显的饥饿;当带宽压力继续提高到75%时,4号主设备几乎完全无法工作。

对比图6和图7可以看出,在访存冲突较为严重的情境下,随着访存带宽需求的逐渐增加,EMSPG方法相对与MAPG方法的优势越来越明显。在中等的访存压力下,本文提出的EMSPG方法仍能达到60%~70%的漏电功耗节约比例,相对MAPG方法有极为显著的提升。

图8展示了在EMSPG方法下1号主设备采用不同配置参数时节约漏电功耗比例。图中OS16表示访存突发长度(outstanding)为16,C表示地址冲突,UC表示地址不冲突。从图中可以看出,无论采用哪一种配置参数,在带宽压力逐渐增大时,可节约的漏电功耗总是在下降。

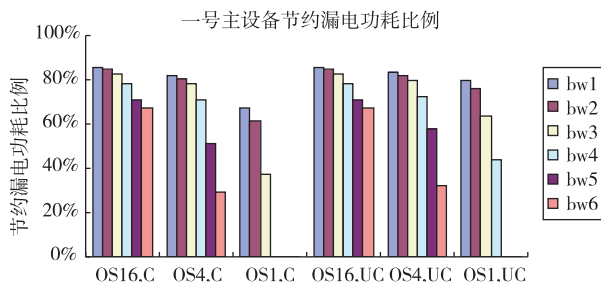


图8 一号主设备节约漏电功耗比例

当突发长度变小时,电源门控的效果也会随之

下降;当突发长度为1时,在25%带宽需求下就已经难以得到节约功耗的效果;带宽压力继续增加时甚至无法满足带宽需求。这是由于较小的突发长度无法充分利用内存控制器的性能,造成了大量的性能浪费。

对比地址冲突和地址不冲突两种情况,可以发现地址冲突时本方法节约功耗的效果只是略微降低;但在突发长度为16时,降低的效果不明显,这是由于本方法使每个单独的主设备都独享内存总线,地址不冲突时只是略微改善了主设备切换时的访存冲突状况。这也证明了本文提出的EMSPG方法可以适应复杂多变的访存状况。

3.3 自适应窗口效果

自适应窗口不但可以提升访存调度的效果,还可以对漏电功耗进行精细优化。

我们设计了一个场景来模拟需求带宽剧烈变化,从而检验自适应窗口扩大和缩小的效果。

图9展示了1号主设备窗口开启时间的变化图,当1号主设备的带宽需求从800MB/s突变为1.6GB/s时,窗口开启时间从预设的128周期变为可用的最大窗口640周期。但640周期超过了其需求,所以几个调度周期后窗口开启时间下降至与带宽对应的大小并保持平稳。当带宽需求从1.6GB/s回落至800MB/s时,窗口开启时间也在几个周期后回落至初始值附近。

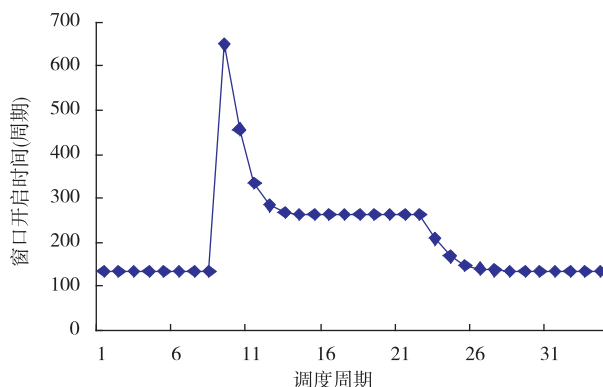


图9 自适应的窗口开启时间

由此可见,本文提出的自适应窗口能够根据带宽的变化对窗口开启时间进行自动调整。

4 结 论

本文提出了一种基于独占式访存调度的电源门控方法,针对多媒体片上系统中DMA模块的访存

行为具有局部性及带宽敏感、延迟不敏感的特点,我们利用访存调度器在一定的调度周期内使每个模块的访存请求集中于一段窗口中。访存模块在分配的窗口内独占访存总线,在最大程度利用访存局部性的同时,将零散访存请求压缩在窗口中,给出整合的空闲时间,并对这段空闲时间使用电源门控方法。实验表明,本文提出的方法可以将访存模块的漏电功耗减少 60% ~ 80%,且相对于已有方法更能适应较大的访存压力。

集中式的访存调度和电源门控方法要求片上系统设计者通盘考虑各个模块的配置对整体的影响,本文的未来工作可以探索在本文方法的基础上如何调整各访存设备的访存行为、缓存能力,使其性能达到较为平衡的状态。此外,还可以研究使独占式访存调度器能够适应带宽变化较为剧烈的设备,将被动的带宽调整改进为主动的方式。

参考文献

- [1] Li S, Ahn J, Strong R, et al. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In: Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture. 2009. 469-480
- [2] Shin Y, Seomun J, Choi K, et al. Power gating: circuits, design methodologies, and best practice for standard-cell VLSI designs. *ACM Trans on Design Automation of Electronic Systems*, 2010, 15(4): 1-37
- [3] Mistry J, Al-Hashimi B, Flynn D, et al. Sub-clock power-gating technique for minimising leakage power during active mode. In: Design, Automation Test in Europe Conference Exhibition (DATE), 2011. 1-6
- [4] Youssef A, Anis M, Elmasry M. Dynamic standby prediction for leakage tolerant microprocessor functional units. In: Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture. 2006. 371-384
- [5] Jeong K, Kahng A B, Kang S, et al. MAPG: memory access power gating. In: Proceedings of the Conference on Design, Automation and Test in Europe. EDA Consortium, 2012. 1054-1059
- [6] Wang Y, Roy S, Ranganathan N. Run-time power-gating in caches of gpus for leakage energy savings. In: Proceedings of the Conference on Design, Automation and Test in Europe. EDA Consortium, 2012. 300-303
- [7] Lin Y, Yang C, Huang J, et al. Memory access aware power gating for MPSoCs. In: Proceedings of the 17th Asia and South Pacific Design Automation Conference (ASP-DAC), 2012. 121-126
- [8] Su M, Gao X, Chen Y, et al. Efficiency-aware QoS DRAM scheduler. In: Proceedings of the IEEE International Conference on Networking, Architecture, and Storage, 2009. 223-226
- [9] Staschulat J, Bekooij M. Dataflow models for shared memory access latency analysis. In: Proceedings of the 7th ACM International Conference on Embedded Software. New York, USA, 2009. 275-284

A power gating technique based on exclusive memory access scheduling for SoCs

Liu Su * * * * *, Su Menghao * * * *, Su Wen * * * *

(* Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

(** University of Chinese Academy of Sciences, Beijing 100049)

(*** Loongson Technology Corporation Limited, Beijing 100190)

(**** Petrochina Oil and Gas Pipeline Control Center, Beijing 100007)

Abstract

Aiming at reducing the leakage power of memory access modules in a system on chip (SoC), a power gating technique based on exclusive memory access scheduling was proposed after the analysis of memory access behavior. Considering that the memory access modules have high tolerance of latency, the new technique uses a memory access scheduler to assemble the memory requests into windows, and performs the power gating for the module and its memory access path when the corresponding window is closed. The proposed technique can improve both the scope and effect of power gating. The experimental result shows that the technique can reduce about 70 percent of leakage power of the memory access modules with no performance penalty.

Key Words: exclusive, memory access scheduling, power gating, system on chip (SoC), leakage power