PAUSEOPT 指令说明

版本号 1.0.0

发布日期 2024.08.13

上海兆芯集成电路股份有限公司

目录

1. 简介

PAUSEOPT 指令是一种全新的 x86 架构指令，旨在优化 CPU 的功耗。该指令通过 CPU 内部机制，使处理器在一段时间内保持低功耗状态，从而提升整体系统的能效表现。PAUSEOPT 指令能够在系统忙等待事件发生时有效地减少功耗，同时保证响应速度，为现代计算需求提供了一个强有力的解决方案。

2. 背景信息

在现代计算环境中，CPU 的功耗优化一直是重要的研究方向。随着处理器核心数量的增加和计算任务的复杂化，如何在满足高性能计算需求的同时有效控制功耗成为一个关键问题。传统的功耗管理方法通常涉及在空闲状态下降低频率或关闭部分核心，但这些方法可能会导致性能下降或响应延迟。

PAUSEOPT 指令的设计初衷是为了解决上述问题。该指令通过在合适的时机让 CPU 进入一种特殊的优化状态，使得 CPU 能够在保持较高性能的同时显著降低功耗。具体来说，当操作系统需要处理一些忙等待的事件时，可以选择执行 PAUSEOPT 来替代原有的忙等待指令，PAUSEOPT 立即触发优化机制。这种机制不仅能够降低电源消耗，还能减少系统的热量产生，延长硬件的使用寿命。

PAUSEOPT 指令的引入为硬件制造商和系统设计者提供了更多的选择，使他们能够在设计产品时更好地平衡性能与功耗之间的关系。对于用户而言，这意味着在日常使用中可以享受更长的电池续航时间和更安静的系统运行环境。

3. 技术细节

3.1 硬件部分

3.1.1 指令支持检测

PAUSEOPT 用 CPUID，leaf=0xC0000006,sub-leaf=0, 得到的 EAX 的 bit0 来做 PAUSEOPT 指令支持的枚举进行枚举，1 表示支持，0 表示不支持。

在 CPUID 显示不支持 PAUSEOPT 时，执行 PAUSEOPT 指令会触发 UD 异常。

3.1.2 指令格式

prefix	opcode	modrm(reg 型)
0xF2	0x0F_A6	0b11 010 - - -

指令有三个源操作数：EAX，EDX，ECX

EDX:EAX 共同组合成一个 64 位的数，单位是 TSC 周期，用来指定 CPU 进入优化状态的时间。

PAUSEOPT 在执行时隐式地使用 ECX 寄存器作为优化状态控制寄存器。

ECX 的第 0 位预留给后续保留功能的支持，当前产品还未用到，设置该 bit 不会有影响。其余 bit 不可设置，设置 1 会报 GP 异常。

3.1.3 非 VMX 环境

PAUSEOPT 可以在任意 ring level 和任意模式下执行，在 ring0 下执行无限制，在非 ring0 下执行时，CR4.TSD 必须是 0，否则会触发 GP 异常。

在执行 PAUSEOPT 之前，OS 可以指定允许 CPU 挂起(进入优化状态)的最长时间。这可以通过写入 PAUSEOPT_CONTROL(0x187F) MSR 来实现。

PAUSEOPT_CONTROL[31:2] 决定了 CPU 可以最长在 P0.1 状态停留的时间，如果写 0，表示没有限制。该`最长时间`是一个 32-bit 的数据，其中的高 30bit 来自该 MSR 的 [31:2]，最低的 2bit 始终为 0。

PAUSEOPT_CONTROL[1] 为保留位。

PAUSEOPT_CONTROL[0]预留给后续保留功能的支持，当前产品还未用到，设置该 bit 不会有影响。

。

如果 PAUSEOPT 指令因到达 OS 指定的最长时间而使 CPU 状态恢复到正常状态，PAUSEOPT 会设置 RFLAGS.CF,如果不是因为到达 OS 指定的最长时间导致的 CPU 恢复到正常状态，PAUSEOPT 会清除 RFLAGS.CF.

以下事件会导致处理器从优化状态中退出：

1. NMI/SMI

2. DEBUG Exception

3. #MC

4. BINIT/INIT/RESET 信号

5. 普通 INTR

退出后，处理器会继续执行 PAUSEOPT 后面的指令。此外,不论 IF 是否为 1，外部中断均会导致 CPU 从优化状态退出。如果 IF=1，外部中断导致 CPU 从优化状态退出，会去执行中断处理程序中的指令；如果 IF=0，外部中断导致 CPU 从优化状态退出，会去执行 PAUSEOPT 后面的指令。

3.1.4 VMX 环境

新增 VMCS 域 VMCS_PROCBASED_CNTL3，用来表示 Guest 环境下是否支持 PAUSEOPT 指令的执行。如果该 bit 是 0，在 VMX Guest 下执行 PAUSEOPT 指令会导致#UD（invalid opcode）。

Index 0x4200 vmcs_procbased_cntl3	Description
Bit 0	bit 0 means "enable PAUSEOPT execution in vmx non-root opration".
Bit 1 – Bit 31	Reserved bits. Set these bits as 0.

VMCS_PROCBASED_CNTL3 中的所有控制，均需要通过 只读 MSR-VMX_PROCBASED_CNTL3_MSR(index:0x12A7)确认当前 CPU 是否支持该控制。

VMX_PROCBASED_CNTL3_MSR	Description
Bit 31:0	**Can Be 0:** Bit X of this field indicates whether bit X of VMCS_PROCBASED_CNTL3 can be set to 0 on vmentry. If bit X of this field is 1, and bit X of VMCS_PROCBASED_CNTL3 is 0 on vmentry，then vmentry will fail.
Bit 63 – Bit 32	**Can Be 1:** Bit X of this field indicates whether bit X of VMCS_PROCBASED_CNTL3 can be set to 1 on vmentry. If bit X of this field is 0, and bit X of VMCS_PROCBASED_CNTL3 is 1 on vmentry，then vmentry will fail.

如果 Guest 在执行 PAUSEOPT 时，“RDTSC exiting”和“VMCS 第三级执行控制域的 bit0”，这两个 control 均为 1，会导致 PAUSEOPT vmexit。

PAUSEOPT 指令的行为在 VMX Non-Root 环境下会发生变化。

如果 VMCS 的第三级执行控制域的 bit0 为 0，PAUSEOPT 的执行会导致#UD.该异常的优先级高于任何指令边缘可能会发生的异常。

PAUSEOPT 指令说明

如果 VMCS 的第三级执行控制域的 bit0 为 1，PAUSEOPT 的具体行为取决于"RDTSC exiting" control 的值。

- 如果"RDTSC exiting" control 为 0，PAUSEOPT 会延迟一段物理时间。该时间由虚拟 delay 时间转换而来，转换依据是 Guest 的 TSC。
- 如果 PAUSEOPT_CONTROL [31:2]是 0(表明 Guest OS 没有限制最大 delay 时间)，**那么虚拟 delay 的虚拟 tsc 数由** EDX:EAX 中指定的 TSC 目标数确定。如果 PAUSEOPT_CONTROL [31:2]不为 0，那么虚拟 delay 的虚拟 tsc 数是 EDX:EAX 与 PAUSEOPT_CONTROL_MSR[bit31:2]<<2 中的较小值指定的 TSC 目标数。
- 具体 delay 时间取决于“use TSC offsetting” and “use TSC scaling”的控制：
  - 如果这两个 control 均为 0，物理 delay 就是虚拟 delay。
  - 如果这两个 control 均为 1，虚拟 delay 会左移 48bit，产生一个 128-bit 的整数，然后除以 tsc multiplier，产生一个 64-bit 的整数，该整数就是实际的物理 delay.
- 如果“RDTSC exiting”control 为 1，PAUSEOPT 会产生一个 PAUSEOPT vmexit.

VMCS 的 vmexit reason 会被填充为 68，以此数值来记录 PAUSEOPT vmexit 事件。

VMCS 的的 vmexit intruction information 域，将会和在 vmx guest mode 下执行 RDRAND / RDSEED 一样，做同一种类型的记录，即：

a. VMCS 的 vmexit intruction length 会在 PAUSEOPT 发生 VMEXIT 时被填充。填充的数据为 PAUSEOPT 指令的长度。
b. VMCS 的 VMEXIT instruction information 会在 PAUSEOPT 发生 vmexit 时被填充。具体的填充内容为：

Bit 定义	内容描述
2:0	undefined
6:3	操作数寄存器 – PAUSEOPT 的源操作数 0 = RAX 1 = RCX 2 = RDX 3 = RBX 4 = RSP 5 = RBP 6 = RSI 7 = RDI 8–15 分别代表 R8–R15(在支持 64 位架构的兆芯 CPU 上)
10:7	undefined
12:11	操作数 size 0:16bit 1:32bit 2:64bit Value 3 未被使用到
31:13	undefined

PAUSEOPT 还引入了一个新的 VMCS 域,即 PAUSEOPT_TARGET_TSC（index 0x2200）。

- 这个域的 default 是 0，由每次 vmlaunch 前配置。

- 这个域在 Guest 下由处理器在 VMEXIT 时自行填写修改，主要是用来记录 guest 下 PAUSEOPT 指令执行所期望达到的目标虚拟 TSC 值。
- 每次在 Guest 下正常执行完 PAUSEOPT 后，处理器会把 PAUSEOPT_TARGET_TSC 清 0。
- vmresume 前，如果 guest_rip 指向的是 PAUSEOPT 的下一条指令，Software（VMM）需要把 PAUSEOPT_TARGET_TSC 清 0

3.2 软件部分

兆芯基于 Linux-6.6 提供了两套补丁，分别用于支持 PAUSEOPT 在 Host 操作系统和 Guest 操作系统中的使用。这里我们称为 Native OS Support 和 Virtualization Support.此外，在虚拟化支持部分，为了在虚拟机中使用 PAUSEOPT，还需要另一个 VMM 补丁。

请注意，这些补丁是对 PAUSEOPT 功能进行枚举和简单使用的初始版本。在未来，可能会出现针对 PAUSEOPT 的不同使用方法。在将 PAUSEOPT 应用于生产环境之前，用户应确保充分了解其各项功能，以避免潜在的影响。

本文档中提及的所有补丁，请联系兆芯客户支持部门获取。

3.2.1 Native OS Support

当 CPUID. C0000006.0:EAX.PAUSEOPT[bit 0] = 1 时，表示支持 PAUSEOPT 指令，PAUSEOPT Host Patch 检测到支持 PAUSEOPT 指令后，会使用 PAUSEOPT 指令实现的 delay 来替换 Linux 内核默认的 delay；从而在 Linux 内核执行 mdelay/udelay/ndelay 函数时触发 PAUSEOPT 指令的执行。

3.2.2 Virtualization OS Support

该功能支持包含两个 Patch，基于 Linux-6.6 的 Patch 和基于 QEMU-5.2.0 的 patch。同时，为了在虚拟机中使用 PAUSEOPT，还需要在 Host OS 中打上 Navtive OS Support 的 patch。

该补丁集用于支持在虚拟机中使用 PAUSEOPT。

QEMU PAUSEOPT Patch Based on QEMU-5.2.0

该补丁向 QEMU 中添加了必要的代码，以提供以下功能：

1. 使虚拟机能够检测到用于枚举 PAUSEOPT 功能的 CPUID
2. 使虚拟机能够使用 PAUSEOPT 相关的 MSR，如 PAUSEOPT_CONTROL

Linux PAUSEOPT Patch Based on Linux-6.6

该补丁向 Linux 内核中添加了必要的代码，以提供以下功能：

1. 根据 Host OS 是否支持 PAUSEOPT 确认是否要在启动虚拟机时修改相关控制，以支持在虚拟机中使用 PAUSEOPT
2. 当虚拟机中执行 PAUSEOPT 时，对可能发生的情况（如产生#UD 异常、发生 PAUSEOPT vmexit、修改 PAUSEOPT_CONTROL MSR 等），进行相应的处理

4. 测试方法

本节描述在打了 PAUSEOPT 相关 patch 后，如何确定 PAUSEOPT 正常执行。

4.1 Native OS

测试 PAUSEOPT：

（1） 添加 PAUSEOPT Host Patch；
（2） 启动并进入 Linux 系统，通过/proc/cpuinfo 查看是否支持 PAUSEOPT 指令：在 flags 中看到 PAUSEOPT 即表示当前平台支持 PAUSEOPT 指令；
（3） 通过（2）确定支持 PAUSEOPT 指令后，Linux 内核多处会执行 mdelay/udelay/ndelay，进而会执行 PAUSEOPT；可以观察系统是否存在异常来判断 PAUSEOPT 指令是否正确执行；
（4） 也可以编写内核 Module，在每个 Core 上创建线程，在线程中循环执行 udelay，进行 Burn In 测试：

线程代码：

```
while true
do
   udelay(100);
   touch_nmi_watchdog();
done
```

同样通过观察系统是否有异常来判断 PAUSEOPT 指令执行是否正确；

4.2 Virtualization OS

应用 QEMU 和 Linux 的 PAUSEOPT patch 后，理论上虚拟机的“硬件环境”已经包含了 PAUSEOPT 支持，即使虚拟机 OS 中没有任何 PAUSEOPT 的指令支持，也可以通过手动编写代码执行 PAUSEOPT，但这并不符合大多数用户的使用场景，因此，仍使用 Native OS 中测试 PAUSEOPT 的方式，测试虚拟机中 PAUSEOPT 的执行是否正常。

（1） 添加 PAUSEOPT Host Patch；
（2） 添加 PAUSEOPT KVM Patch；

（3） 添加 PAUSEOPT QEMU Patch;

（4） 添加 PAUSEOPT Guest Patch(与 PAUSEOPT Host Patch 相同)；

（5） 用 QEMU 启动并进入虚拟机 OS，通过/proc/cpuinfo 查看是否支持 PAUSEOPT 指令：在 flags 中看到 PAUSEOPT 即表示当前平台支持 PAUSEOPT 指令；

（6） 通过（5）确定支持 PAUSEOPT 指令后，Guest OS 内核多处会执行 mdelay/udelay/ndelay，进而会执行 PAUSEOPT；可以观察系统是否存在异常来判断 PAUSEOPT 指令是否正确执行；

（7） 也可以编写内核 Module，在每个 vCPU 上创建线程，在线程中循环执行 udelay，进行 Burn In 测试：

线程代码如 4.1 节中所示。

5. 法律声明

6. 版本记录

版本号	日期	描述
V 0.0.1	2024-05-22	初始草稿
V 1.0.0	2024-08-13	发布 1.0.0 版本