



# 浪潮信息云峦 KeyarchOS 基于 DSA 虚拟机加速启动最佳实践

浪潮电子信息产业股份有限公司

2023 年 10 月

## 目 录

1 测试概述 .....	1
1.1 应用场景 .....	1
1.2 技术背景 .....	1
1.3 测试内容 .....	1
1.4 术语解释 .....	1
2 软硬件环境 .....	2
2.1 硬件 .....	2
2.2 软件 .....	2
2.3 测试工具 .....	2
2.4 技术架构 .....	2
3 测试指导 .....	3
3.1 BIOS 配置及内核配置 .....	3
3.2 内核参数配置 .....	3
3.3 验证 DSA 是否正常 .....	4
4 测试用例及测试数据 .....	5
4.1 测试用例 .....	5
4.2 测试数据 .....	9
5 分析与结论 .....	9
6 可能存在的问题与解决方式 .....	10
7 附录 .....	10

# 1 测试概述

## 1.1 应用场景

虚拟机在使用透传设备，特别是为虚拟机配置大内存时，虚拟机的启动时间会变慢。虚拟机在启动时，会将虚拟地址和物理地址之间的关系全部固定下来，在这期间会触发“page fault”，并会对每一个内存页进行清零操作，这个过程相对来说耗时较长，是虚拟机启动时间慢的重要原因。

## 1.2 技术背景

为了解决内存页清零耗时较长的问题，KOS 基于 DSA 的”Memory Fill”能力，实现内存页的预清零，相比于其他的预清零方案，基于 DSA 的内存预清零方案可以将相关负载从 CPU 卸载到 DSA 中，有效提升应用程序和平台的性能。

## 1.3 测试内容

本文档是针对虚拟机设备透传，启动较慢的场景，使用 DSA 进行内存页预清零，加速虚拟机启动的测试报告，主要测试内容包括，DSA 加速内存页预清零，DSA 加速虚拟机启动，涉及到的测试用例总计 3 条。

## 1.4 术语解释

名词	描述
DSA	Intel 数据流加速器

## 2 软硬件环境

### 2.1 硬件

设备名称	部件	型号
NF5280M7*1 台	CPU	Intel(R) Xeon(R) Gold 6448H*2
	内存	内存 512G
	网卡	万兆网卡 Intel_F102IX710*1

### 2.2 软件

软件	版本
浪潮信息云峦服务器操作系统 KeyarchOS	V5.9
Kernel	5.10.134-15.2.8.x86_64.kos

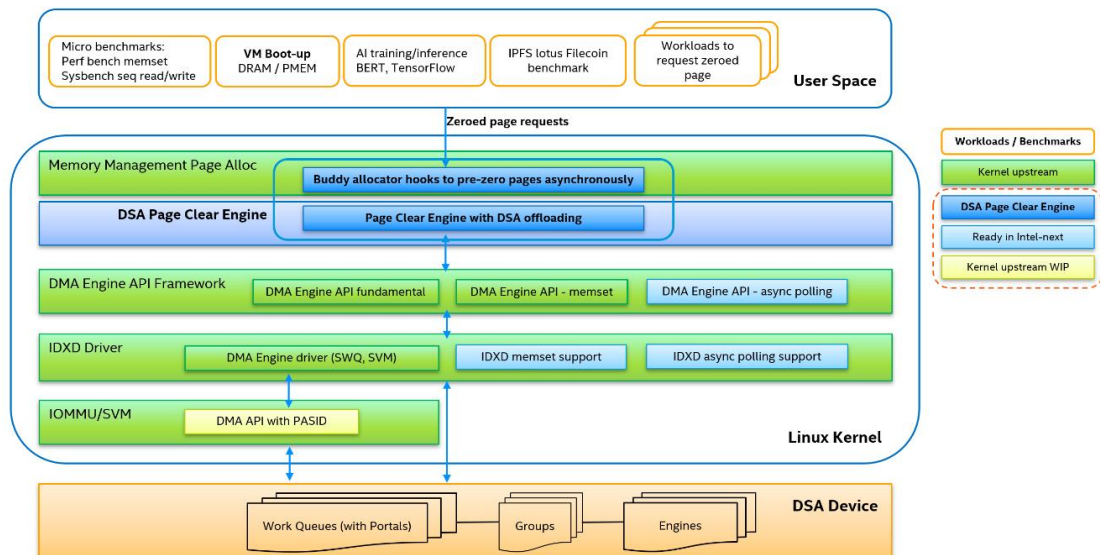
### 2.3 测试工具

测试工具	版本	测试内容	备注
Perf			

### 2.4 技术架构

虚拟机在使用透传设备，特别是为虚拟机配置大内存时，虚拟机的启动时间会变慢。虚拟机在启动时，会将虚拟地址和物理地址之间的关系全部固定下来，在这期间会触发“page fault”，并会对每一个内存页进行清零操作，这个过程相对来说耗时较长，是虚拟机启动时间慢的重要原因。为了解决内存页清零耗时较长的问题，KOS 基于 DSA 的”Memory Fill”

能力，实现内存页的预清零，相比于其他的预清零方案，基于 DSA 的内存预清零方案可以将相关负载从 CPU 卸载到 DSA 中，有效提升应用程序和平台的性能。该方案通过在每个 Node 节点上创建内核线程周期性的对空闲内存进行清零操作，使得在虚拟机启动发生“page fault”时可以跳过耗时较长的清零操作。



## 3 测试指导

### 3.1 BIOS 配置及内核配置

打开 M7 的隐藏项：

Platform --> Miscellaneous Configuration --> Advanced debug

function<Disable>

打开 DSA 设备：

Socket Configuration --> IIO Configuration --> IOAT Configuration--> Socket 0 (Sck0) IOAT Config --> DSA <Enable>

Socket Configuration --> IIO Configuration --> Intel VT for Direct I/O (VT-d) --> Intel VT for Direct I/O (VT-d) <Enable>

### 3.2 内核参数配置

Intel\_IOMMU 配置

必须在以下操作系统中启用具有可扩展模式支持的 Intel® IOMMU 驱动程序 (CONFIG\_INTEL\_IOMMU\_SVM)，内核配置如下图所示。如果未开启 CONFIG\_INTEL\_IOMMU\_DEFAULT\_ON 和 CONFIG\_INTEL\_IOMMU\_SCALABLE\_MODE\_DEFAULT\_ON 选项，则必须在内核启动参数中添加 “intel\_iommu=on,sm\_on”。

```
CONFIG_INTEL_IOMMU=y
CONFIG_INTEL_IOMMU_SVM=y
CONFIG_INTEL_IOMMU_DEFAULT_ON=y
CONFIG_INTEL_IOMMU_SCALABLE_MODE_DEFAULT_ON=y
```

## DSA 驱动配置

在构建/安装 Linux 内核时，启用所示的内核配置选项

```
CONFIG_INTEL_IDXD=m
CONFIG_INTEL_IDXD_SVM=y
CONFIG_INTEL_IDXD_PERFMON=y
```

### 3.3 验证 DSA 是否正常

Intel® DSA 是第四代英特尔® 至强® 可扩展处理器中集成的数据流加速器。在操作系统中可通过 lspci 查看其设备信息。

```
# lspci -d 8086:0b25
```

```
75:01.0 System peripheral: Intel Corporation Device 0b25
```

```
f2:01.0 System peripheral: Intel Corporation Device 0b25
```

```
# lsmod | grep -i idxd
```

```
idxd                118784  0
```

```
idxd_bus            20480  1 idxd (5.10 有的版本直接编进内核)
```

```
# dmesg | grep -i idxd
```




```
[ 14.629548] idxd 0000:75:01.0: enabling device (0144 -> 0146)
```

```
[ 14.631310] idxd 0000:75:01.0: Intel(R) Accelerator Device (v1S00)
```



## 4 测试用例及测试数据

### 4.1 测试用例

#### 测试用例 1

测试类型	DSA加速虚拟机启动	测试工具	Perf
测试目的	验证在未开启THP和DSA的场景下虚拟机启动时间		
前提条件			
<div>1. 硬件设备具体DSA加速器</div> <div>2. 宿主机内核为5. 10. 134-15. 2及以上版本，并且支持prezero功能，且合入如下patch</div> <div></div> <div>0001-anolis-mm-prezero-adapt-to-high-order-pcp.patch</div> <div>3. Qemu-kvm需要合入如下patch</div> <div></div> <div>0001-add-map-s ync-flags.patch</div> <div>4、虚拟机内核要合入如下patch</div> <div></div> <div>outb.diff</div>			
测试过程			
<div>1. 关闭THP，</div> <div># echo never &gt; /sys/kernel/mm/transparent_hugepage/enabled</div> <div># echo never &gt; /sys/kernel/mm/transparent_hugepage/defrag</div> <div>2. 启动虚拟机，设置虚拟机内存为16G，32G，64G，128G，分别获取虚拟机测试结果</div> <div># perf stat ./boot-time.sh 1 16G</div> <div># perf stat ./boot-time.sh 1 32G</div> <div># perf stat ./boot-time.sh 1 64G</div> <div># perf stat ./boot-time.sh 1 128G</div>			





<pre># perf stat ./boot-time.sh 1 200G</pre> <div> <b>boot-time.sh</b>  <b>qemu-boot-loop.sh</b></div>	
<p>预期目标</p> <p>能正确获取到启动时间。</p>	
<p>测试结果：<input type="checkbox"/> 通过      <input type="checkbox"/> 不通过</p>	
备注	

测试用例 2

测试类型	DSA加速虚拟机启动	测试工具	Perf
测试目的	验证在开启THP、未开启DSA的场景下虚拟机启动时间		
前提条件			
1. 硬件设备具体DSA加速器			
2. 宿主机内核为5.10.134-15.2及以上版本，并且支持prezero功能，且合入如下patch			
<div></div> <div>0001-anolis-mm-prezero-adapt-to-high-order-pcp.patch</div>			
3. Qemu-kvm需要合入如下patch			
<div></div> <div>0001-add-map-s yinc-flags.patch</div>			
4、虚拟机内核要合入如下patch			
<div></div> <div>outb.diff</div>			
测试过程			
1. 开启THP，开启透明大页THP			
# echo always > /sys/kernel/mm/transparent_hugepage/enabled			



<pre># echo always &gt; /sys/kernel/mm/transparent_hugepage/defrag</pre> <p>2. 释放缓存和压缩内存</p> <pre># sync</pre> <pre># echo 3 &gt; /proc/sys/vm/drop_caches</pre> <pre># echo 1 &gt; /proc/sys/vm/compact_memory</pre> <p>3. 启动虚拟机，设置虚拟机内存为16G，32G，64G，128G，分别获取虚拟机测试结果</p> <pre># perf stat ./boot-time.sh 1 16G</pre> <pre># perf stat ./boot-time.sh 1 32G</pre> <pre># perf stat ./boot-time.sh 1 64G</pre> <pre># perf stat ./boot-time.sh 1 128G</pre> <pre># perf stat ./boot-time.sh 1 200G</pre> <div><div> boot-time.sh</div><div> qemu-boot-loop.sh</div></div>	
<p>预期目标</p> <p>能正确获取到启动时间。</p>	
<p>测试结果：<input type="checkbox"/> 通过      <input type="checkbox"/> 不通过</p>	
备注	

测试用例 3

测试类型	DSA加速虚拟机启动	测试工具	Perf
测试目的	验证在开启THP、开启DSA的场景下虚拟机启动时间		
前提条件			
<div>1. 硬件设备具体DSA加速器</div> <div>2. 宿主机内核为5.10.134-15.2及以上版本，并且支持prezero功能，且合入如下patch</div> <div> 0001-anolis-mm-prezero-adapt-to-high-order-pcp.patch</div> <div>3. Qemu-kvm需要合入如下patch</div>			



0001-add-map-s  
ync-flags.patch

4、虚拟机内核要合入如下patch



outb.diff

测试过程

1. 开启透明大页THP

```
# echo always > /sys/kernel/mm/transparent_hugepage/enabled  
# echo always > /sys/kernel/mm/transparent_hugepage/defrag
```

2. 释放缓存和压缩内存

```
# sync  
# echo 3 > /proc/sys/vm/drop_caches  
# echo 1 > /proc/sys/vm/compact_memory
```

3. 配置和启动dsa

```
# ./enable-dsa.sh
```



enable-dsa.sh

4. 启用DSA页清零加速引擎



```
# echo 1 > /sys/kernel/mm/prezero/page_clear_engine/hw_enabled
```

5. 启用内存页清零功能并等待1~2分钟

```
# echo 1 > /sys/kernel/mm/prezero/sleep_msecs  
# echo 100 > /sys/kernel/mm/prezero/max_percent  
# echo 1 > /sys/kernel/mm/prezero/enabled
```

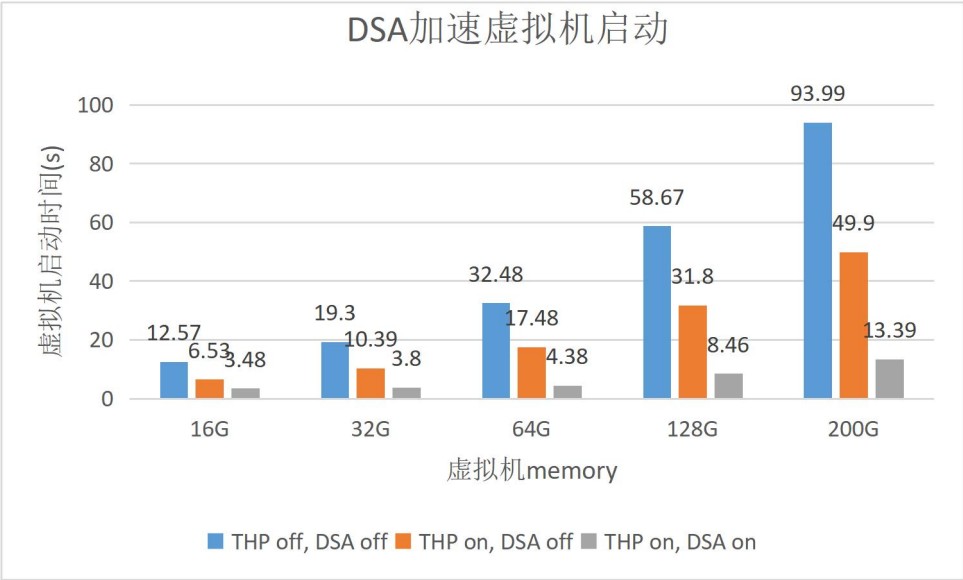
6. 启动虚拟机，设置虚拟机内存为16G，32G，64G，128G，分别获取虚拟机测试结果

```
# perf stat ./boot-time.sh 1 16G  
# perf stat ./boot-time.sh 1 32G  
# perf stat ./boot-time.sh 1 64G
```

<pre># perf stat ./boot-time.sh 1 128G # perf stat ./boot-time.sh 1 200G</pre> <div><div></div><div></div></div> <div>boot-time.sh    qemu-boot-loop.sh</div>	
预期目标 能正确获取到启动时间。	
测试结果： <input type="checkbox"/> 通过 <input type="checkbox"/> 不通过	
备注	

## 4.2 测试数据

测试数据 1：DSA 加速虚拟机启动（测试过程中会启动两个虚拟机，测试启动两个虚拟机的总的时间，启动时间可能会受硬件配置等影响）



## 5 分析与结论

综合上述测试数据来看，启动 DSA 之后，虚拟机启动时间节省 70%左右。

## 6 可能存在的问题与解决方式

---

无

## 7 附录

---

无