中国移动 China Mobile | 中移软件 CMSoft

ARM/x86服务器的ceph性能对比

演讲人：黄小曼

公司：中移（苏州）软件技术有限公司

CONTENTS

PART 01

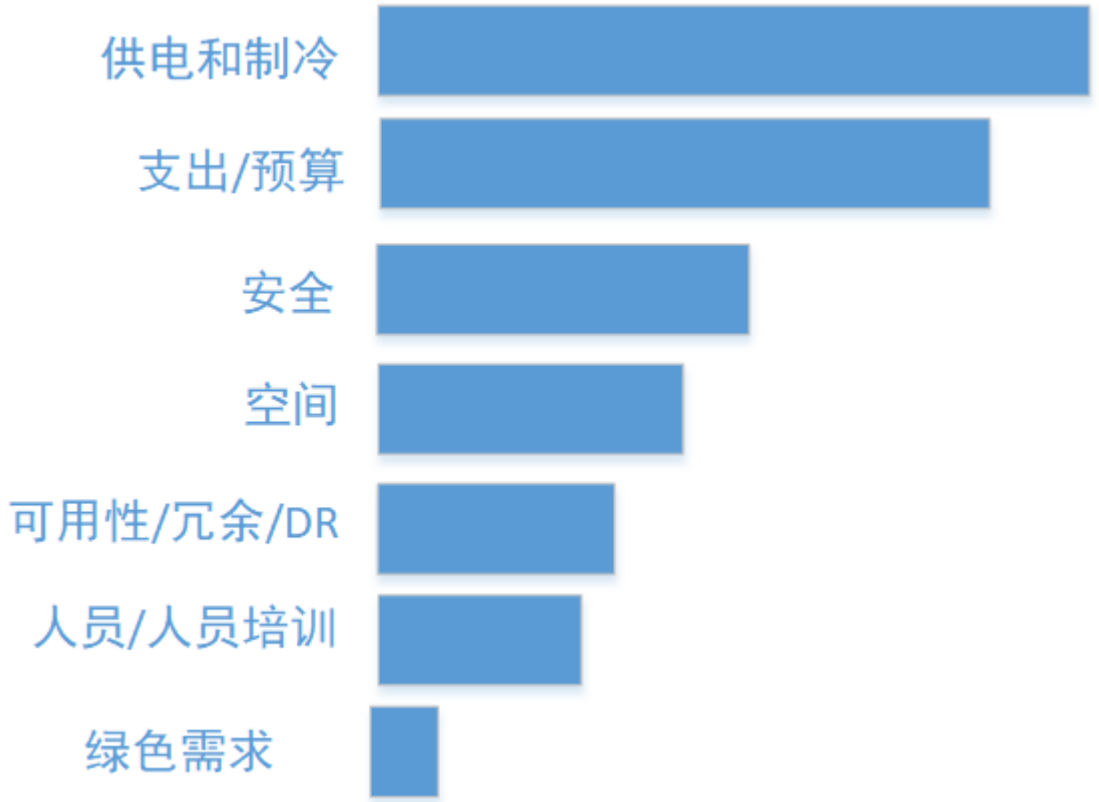
简介

背景介绍

业界动态：数据中心的挑战

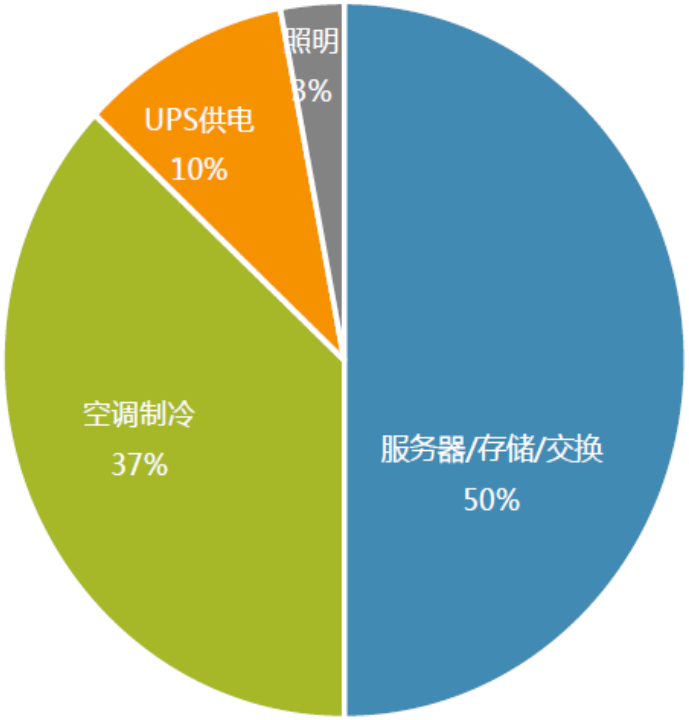
- 数据中心的平均寿命为10年
  有45%的数据中心已超过10年
- 65%的数据中心存在供电和制冷方面的问题
- 30%希望未来3-5年升级主要的数据中心基础设施
- 70%数据中心运行在65-75华氏度（18-24摄氏度）
- 90%表示高能效服务是重要的选择

数据中心所面临的挑战：



背景介绍

业界动态：降低服务器功耗



数据中心功耗占比

服务器/存储/交换电力占数据中心总电力50%以上

据IDC调查，65%的数据中心存在供电和制冷方面问题

数据中心平均四年的电费将超过数据中心基础设施投资

背景介绍

降耗方案：用ARM芯片做存储

在服务器CPU市场上，Intel占据超过90%市场份额。

而在移动芯片市场上，ARM架构的芯片几乎占据了所有市场份额。

ARM处理器的特点：体积小 | 低功耗 | 低成本

2013年百度和Marvell合作了一款ARM服务器，用于百度数据中心。

2016年12月，Cavium与中国联通签合作协议，ARM服务器走进中国联通CORD产业联盟。

2017年2月，阿里巴巴与ARM合作，逐步把数据中心的英特尔CPU换成ARM产品，以提高用电效率。

测试简介

测试工具：
前端：Cosbench
后端：Radosbench

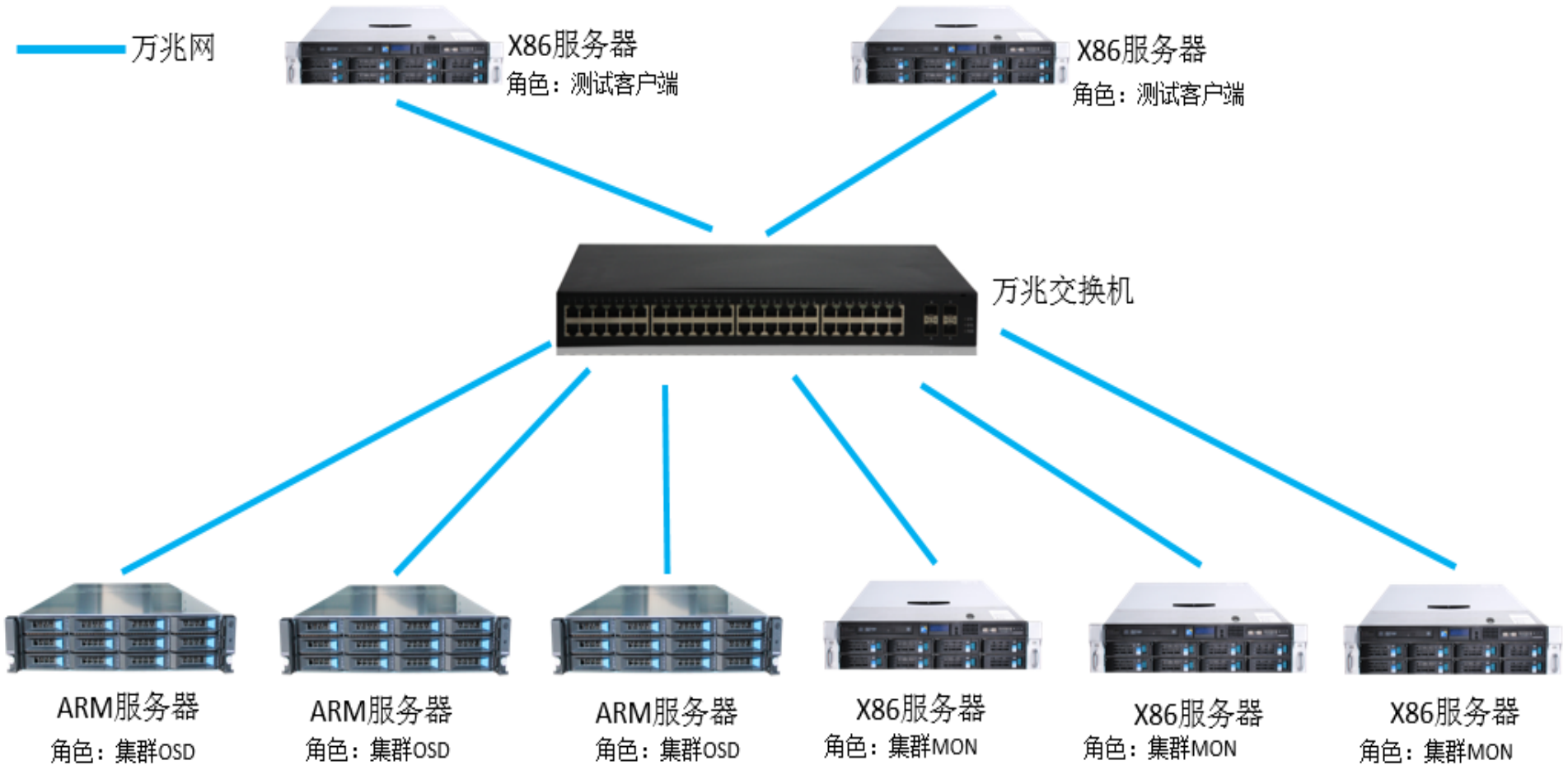
测试内容：
从前端/后端，64K/4M两种文件上，分别对ARM和x86进行读/写/删除性能测试（其中ARM和x86服务器集群均使用SSD盘来存放元数据信息）
Ceph：12.2.5

PART 02

测试环境

ARM测试环境

集群组网拓扑



ARM测试环境

单机箱外部结构

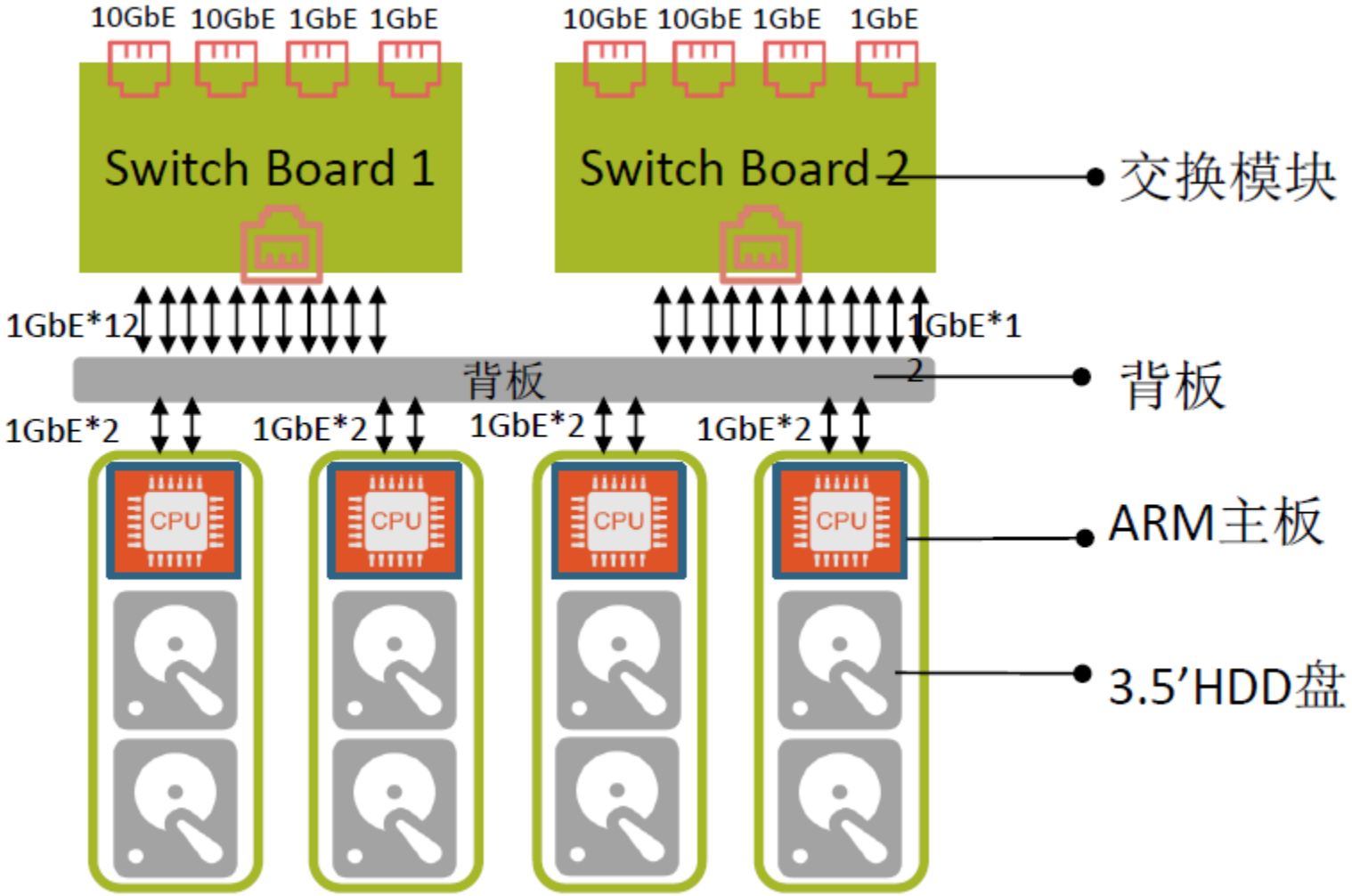
前面板图

后面板图

存储模块图

ARM测试环境

单机箱内部结构



高度：2U
功耗：270W左右
机柜：支持标准19英寸机柜
服务器尺寸：482.6mm x 88.9mm x 650mm (宽 x 高 x 深)

ARM测试环境

软硬件详细配置

主机数 / 硬件配置	3台集群ARM服务器(2U)（存储型节点）	3台集群X86服务器（管理型节点）	2台测试机
节点数	一台ARM服务器中12个OSD	每一台服务器上部署了mon和radosgw	部署cosbench和radosbench
处理器	单个存储模块配置一个ARM处理器 ARMv7 2核 32bit 1.6GHZ	Intel(R) Xeon(R) CPU E3-1231 v3 @ 3.40GHz	Intel(R) Xeon(R) CPU E3-1231 v3 @ 3.40GHz
内存	单个存储模块配置2G内存	32GB	32GB
硬盘	单个存储模块2块硬盘，每块6T，本次只使用一块	256G SSD	256G SSD
网络	一个ARM服务器上两块交换板，一个ARM服务器上有两个万兆网卡，但是节点之间是用千兆网络	1块双口万兆网卡	1块双口万兆网卡

操作系统： CentOS Linux release 7.3.1611 (Core)	Ceph：12.2.5

X86测试环境

软硬件详细配置

主 机数 硬件配置	3台x86服务器 （存储型节点+管理型节点）	1台测试机
节点数	一台x86服务器中12个存储节点，用于部署OSD，部署mon和radosgw	部署cosbench和radosbench
处理器	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
内存	128G	128G
硬盘	共12块硬盘，55T,10HDD+2SSD	共12块硬盘，55T,10HDD+2SSD
网络	每个服务器1张万兆网卡，节点间使用万兆网络	1张万兆网卡

操作系统： BigCloud Enterprise Linux release 7.3.1611 (Core)	Ceph：12.2.5

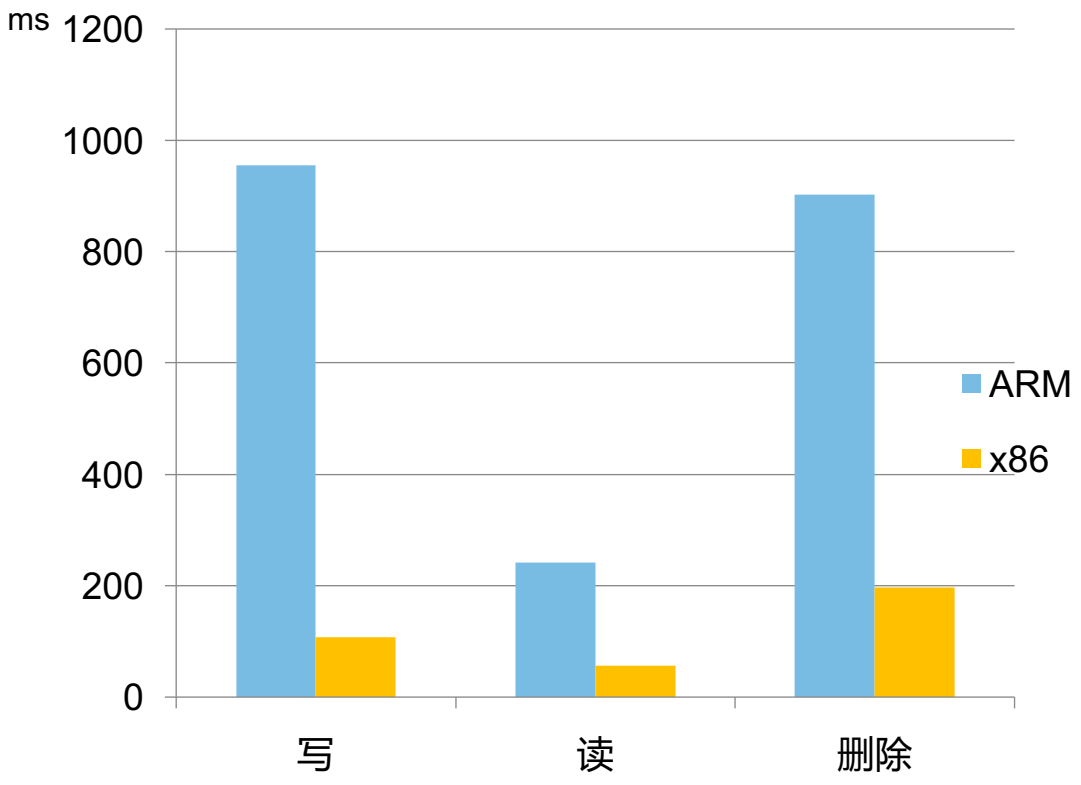
PART 03

测试结果

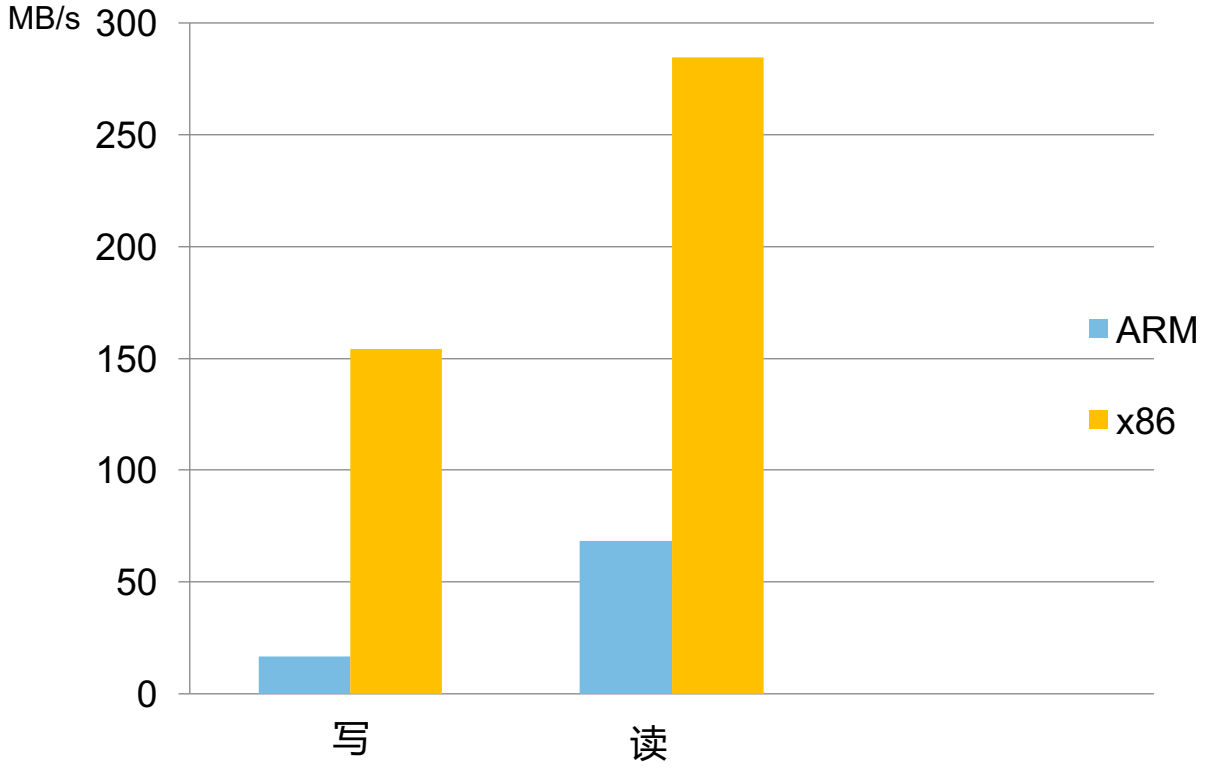
测试结果

前端-64K-250并发，300万对象

处理时间

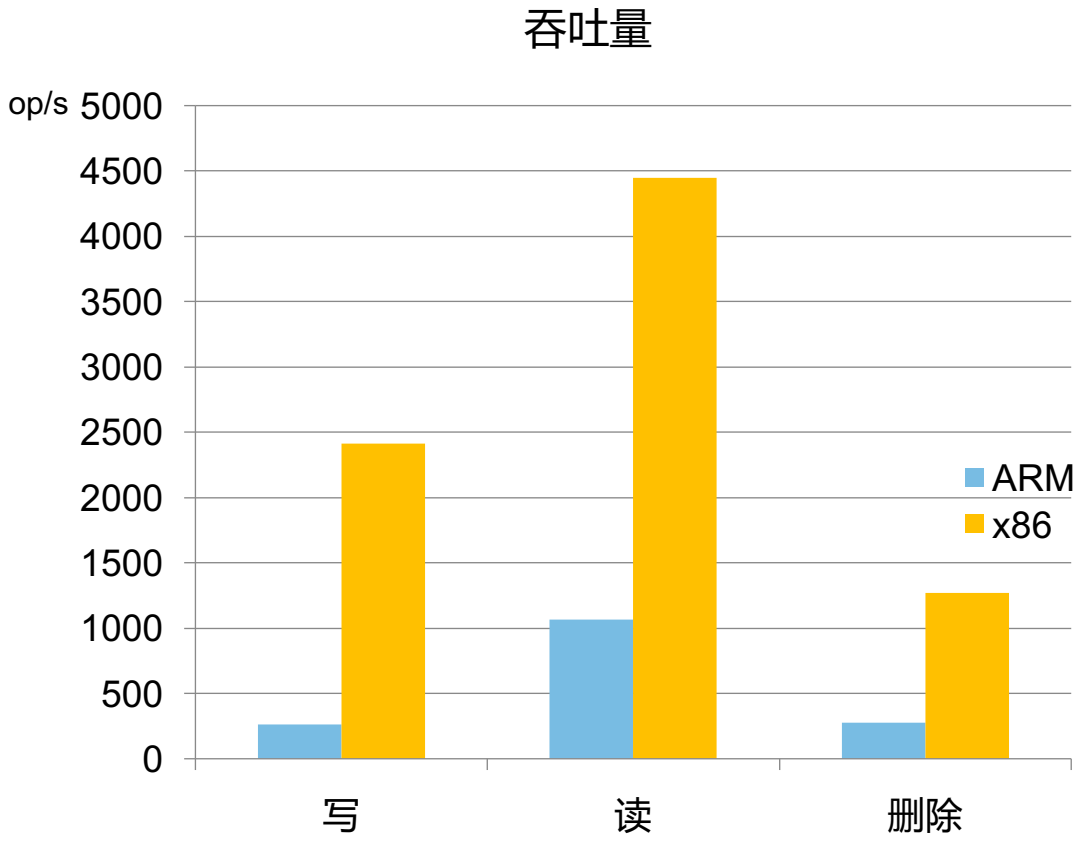


带宽



测试结果

前端-64K-250并发，300万对象



三副本配置下，从处理时间、IOPS和带宽综合看：

1，x86集群的写性能约是ARM的9倍；

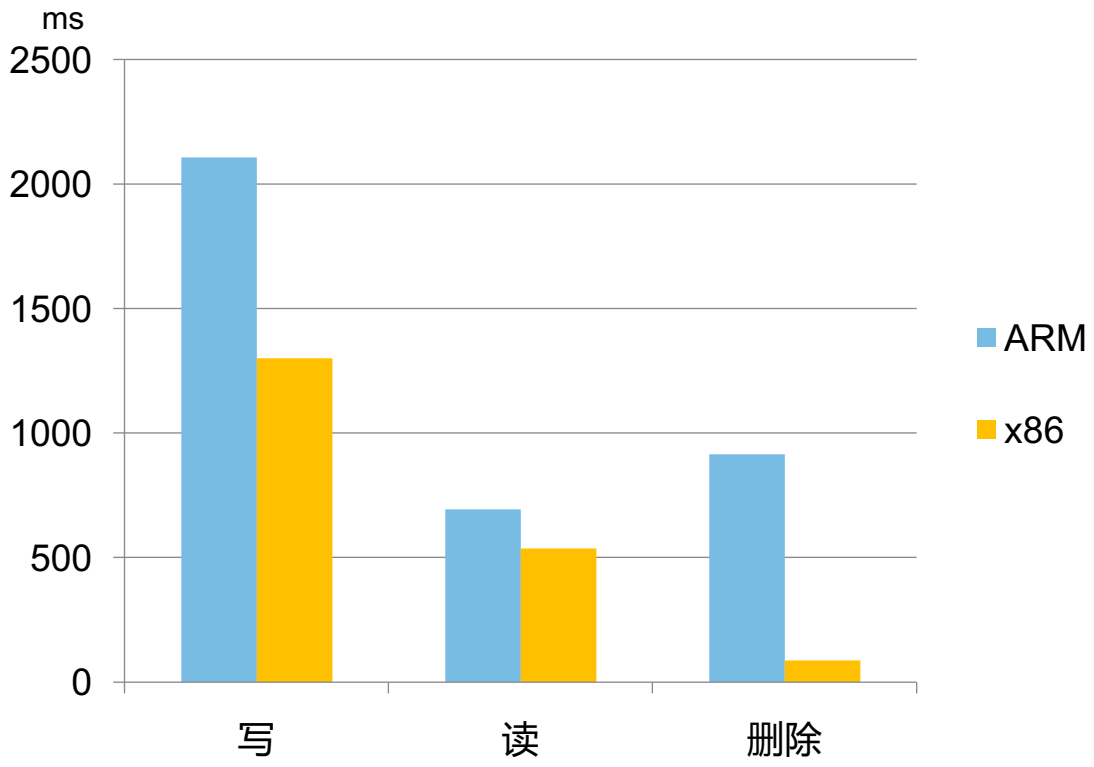
2，读性能是ARM的4倍；

3，删除是ARM的4.5倍

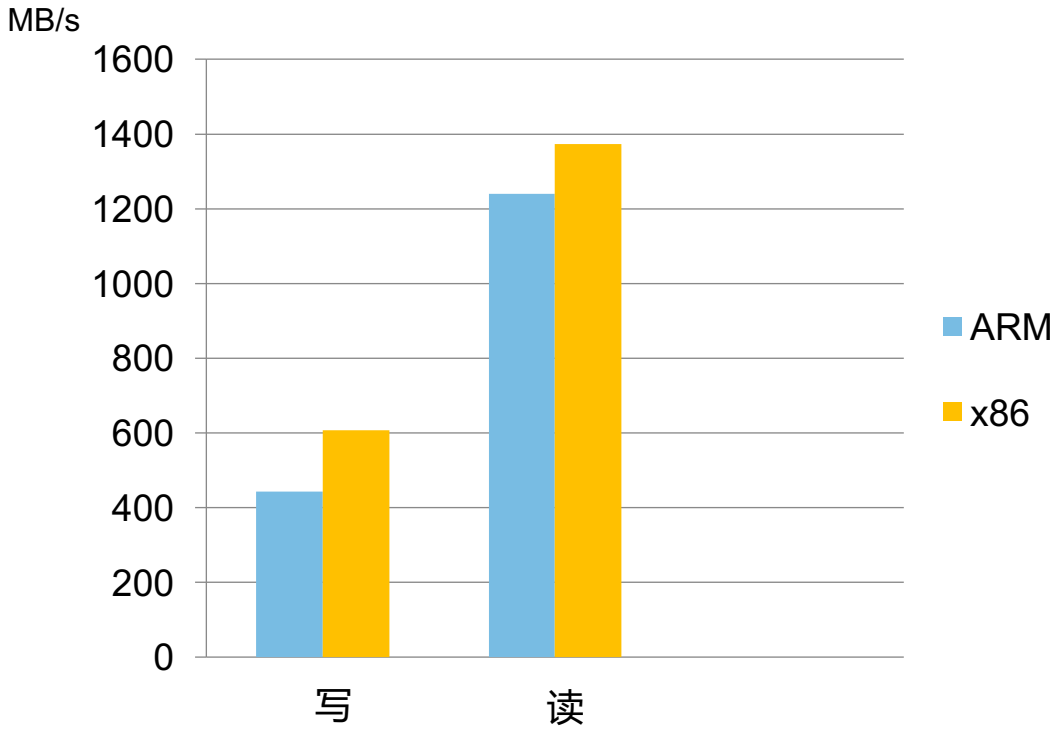
测试结果

前端-4M-200并发，30万对象

处理时间

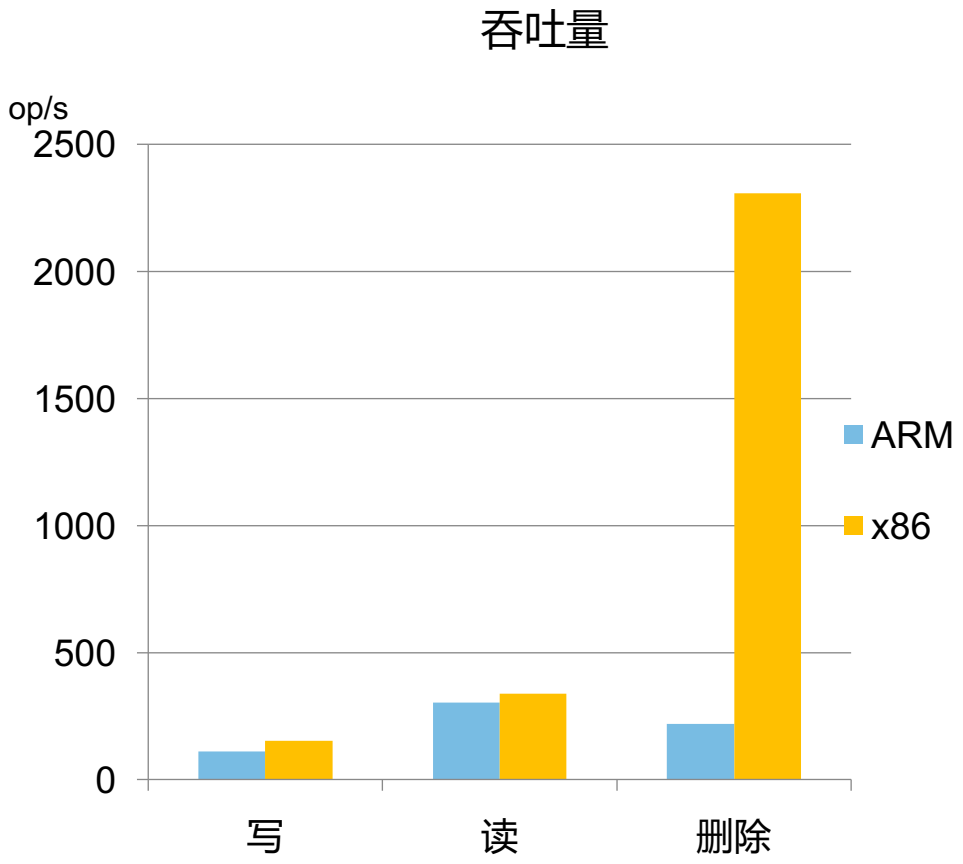


带宽



测试结果

前端-4M-200并发，30万对象



三副本配置下，从处理时间、IOPS和带宽综合看来：

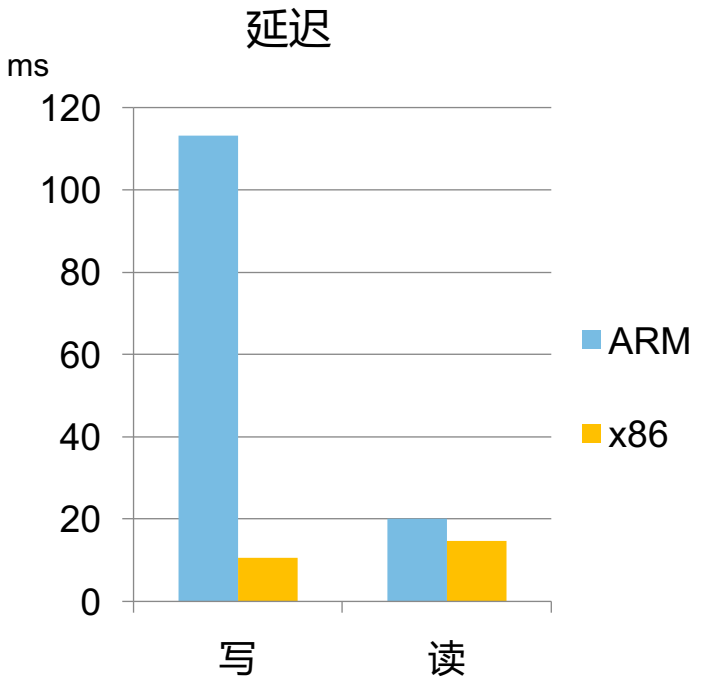
1，x86集群的写性能是ARM的1.3倍；

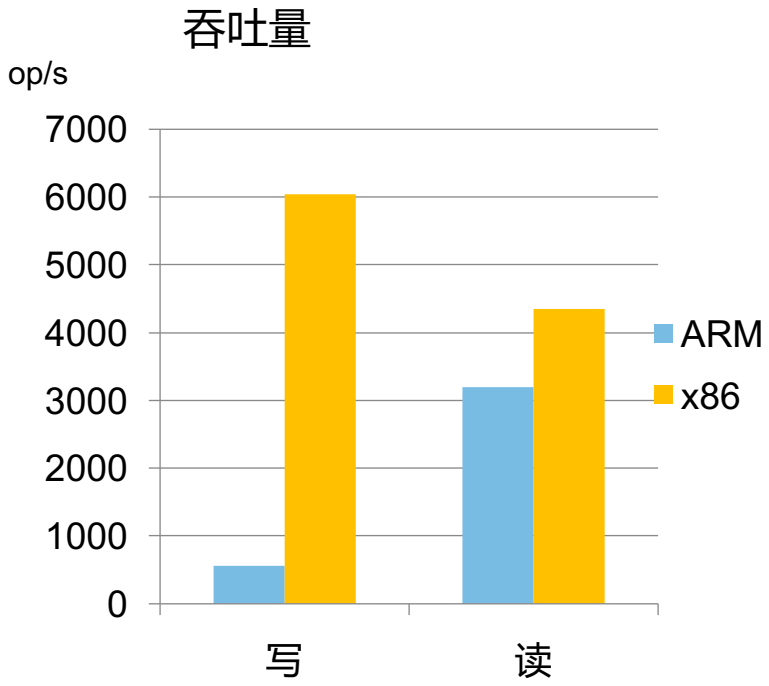
2，读性能略优于ARM；

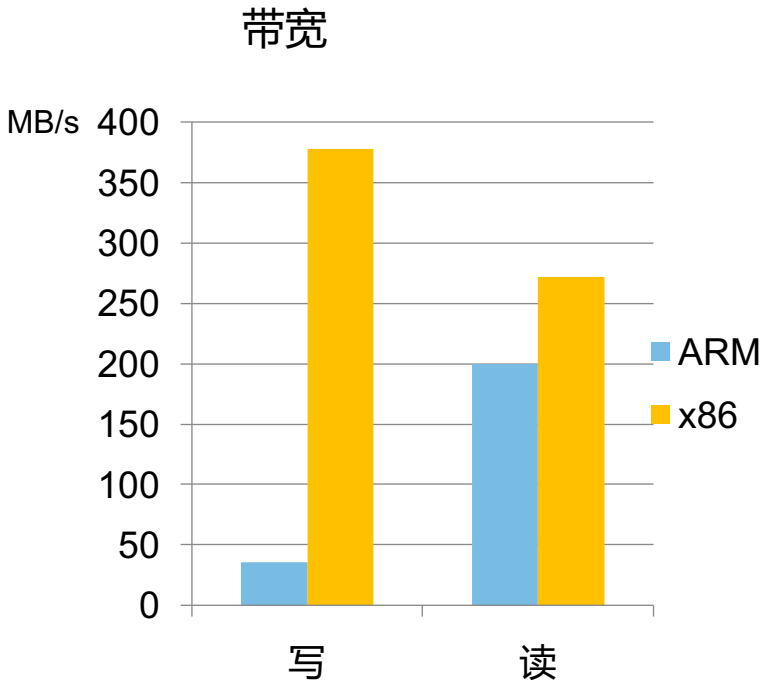
3，删除性能是ARM的10倍

测试结果

后端-64K-64并发，运行300s







三副本配置下，从处理时间、IOPS和带宽综合看来，64k小文件在64并发下的读写性能：

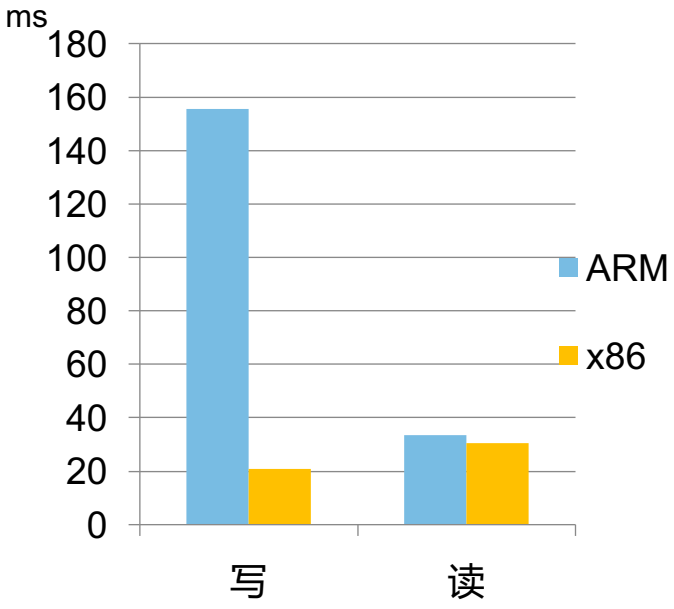
1，x86集群的写性能是ARM的10倍；

2，读性能略优于ARM；

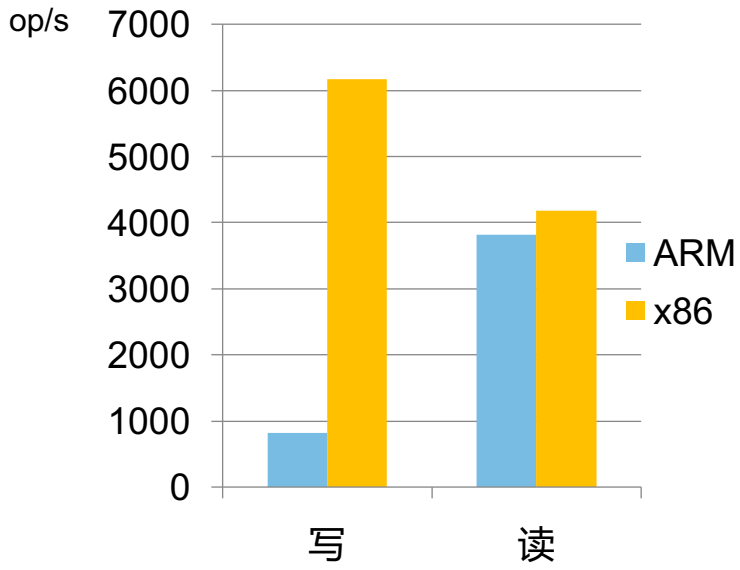
测试结果

后端-64K-128并发，运行300s

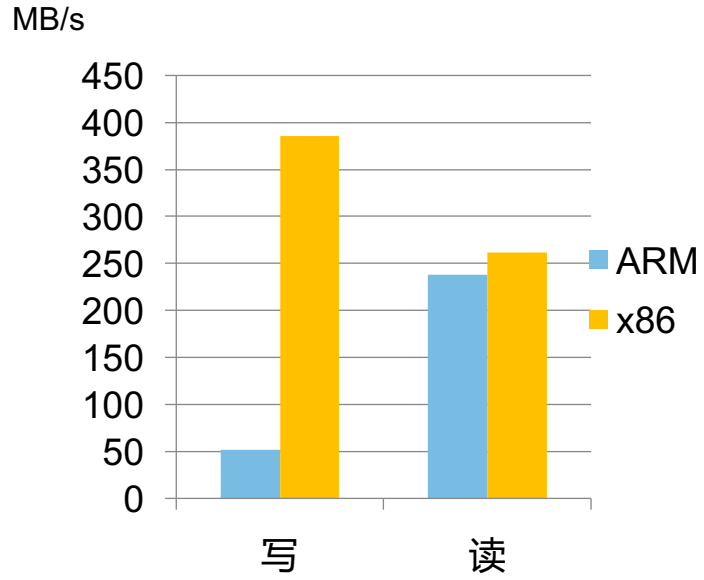
延迟



吞吐量



带宽



三副本配置下，从处理时间、IOPS和带宽综合看来，64k小文件在128并发下的读写性能：

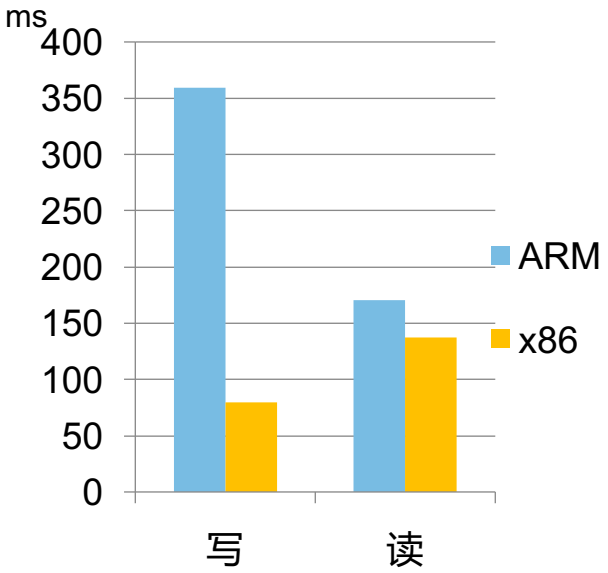
1，x86写性能是ARM的7.5倍

2，读性能差距不大

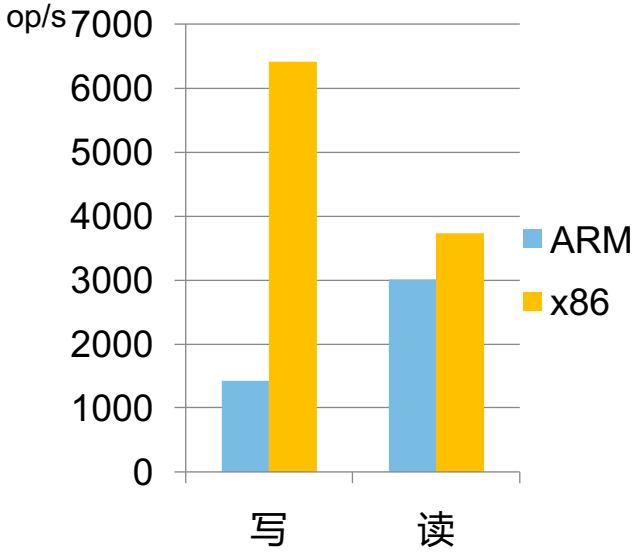
测试结果

后端-64K-512并发，运行300s

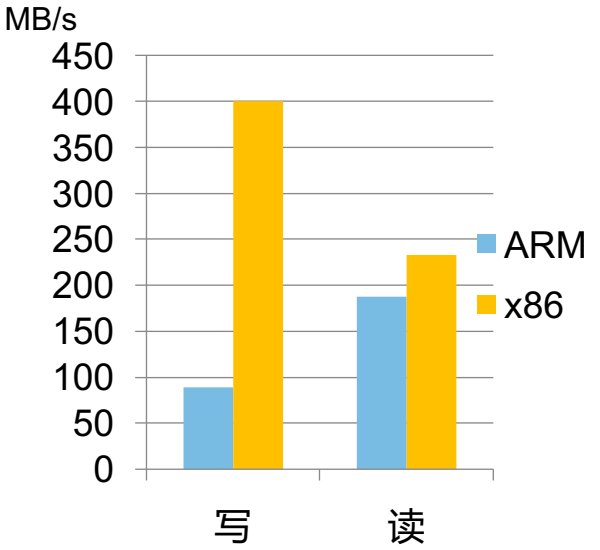
延迟



吞吐量



带宽



三副本配置下，从处理时间、IOPS和带宽综合看来，64k小文件在512并发下的读写性能：

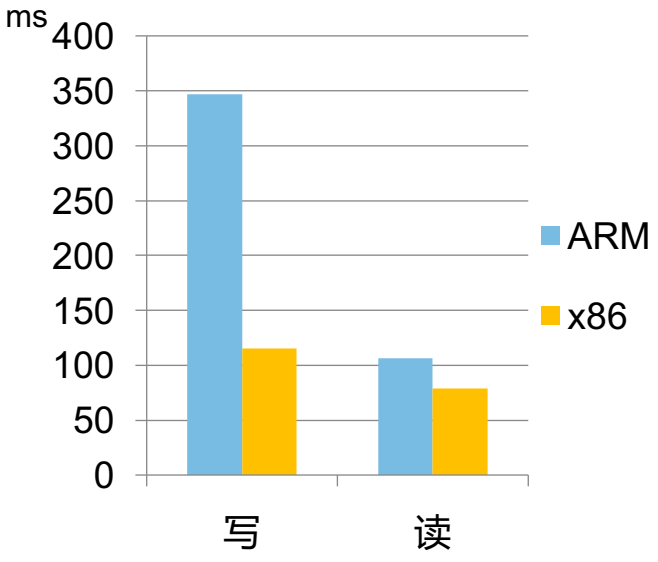
1，x86写性能是ARM的4.5倍

2，读性能是ARM的1.25倍。

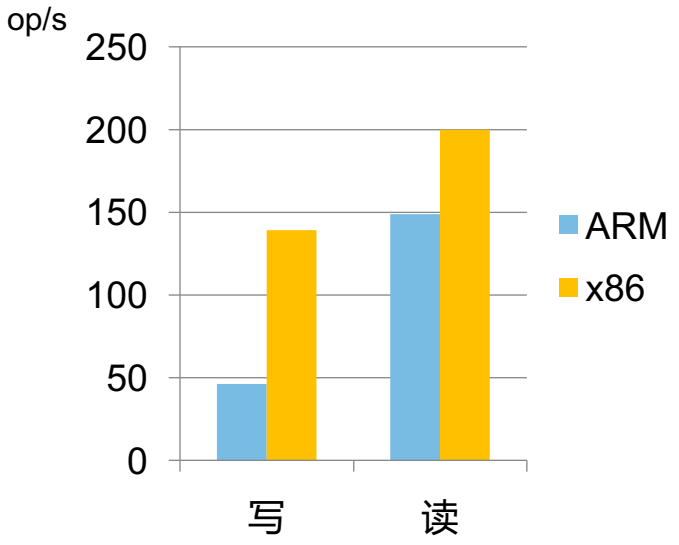
测试结果

后端-4M-16并发，运行300s

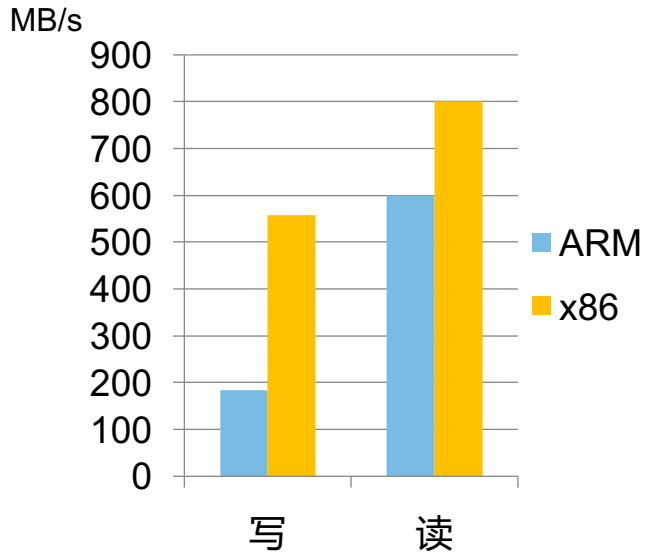
延迟



吞吐量



带宽



三副本配置下，从处理时间、IOPS和带宽综合看来，4M大文件在16并发下的读写性能：

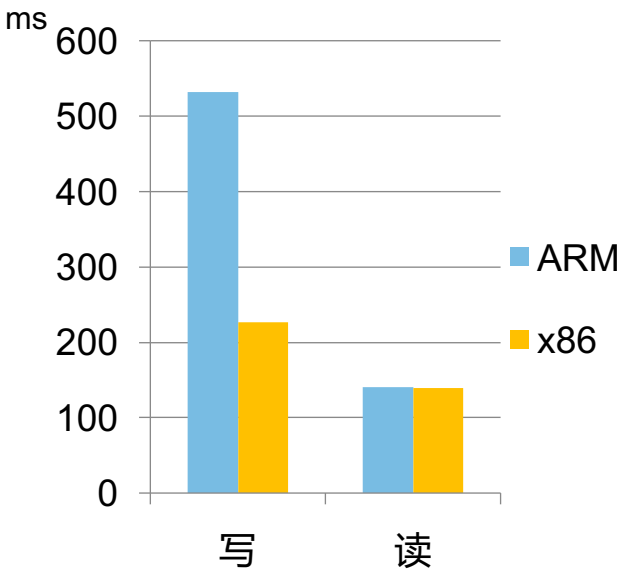
1，x86写性能是ARM的3倍

2，x86读性能是ARM的1.3倍。

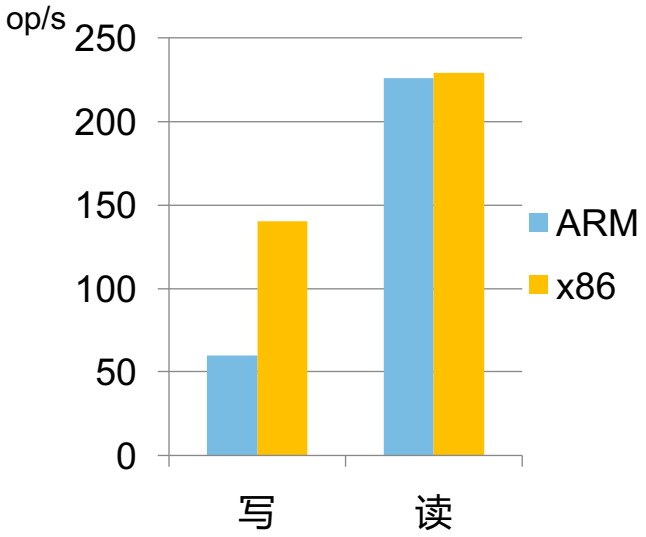
测试结果

后端-4M-32并发，运行300s

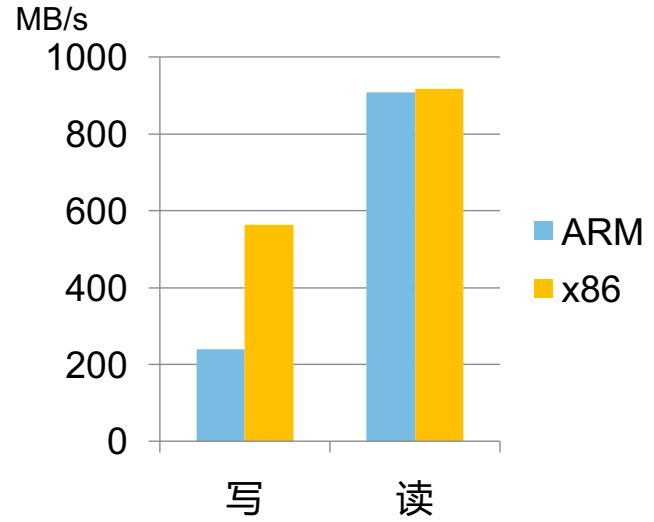
延迟



吞吐量



带宽



三副本配置下，从处理时间、IOPS和带宽综合看来，4M大文件在32并发下的读写性能：

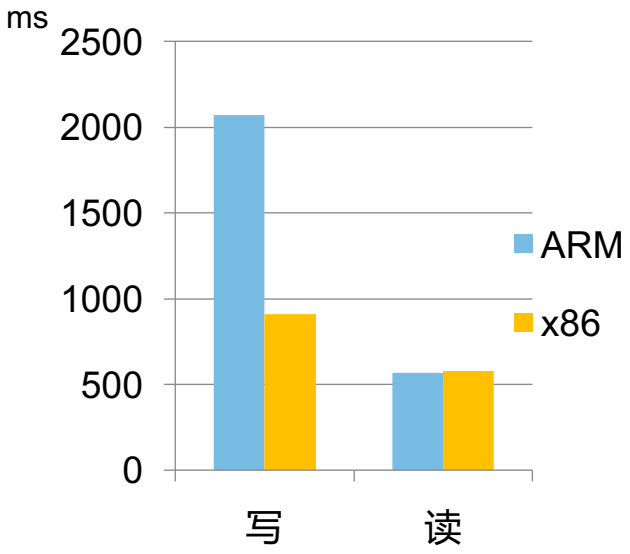
1，x86写性能是ARM的2.3倍

2，读性能差距不大。

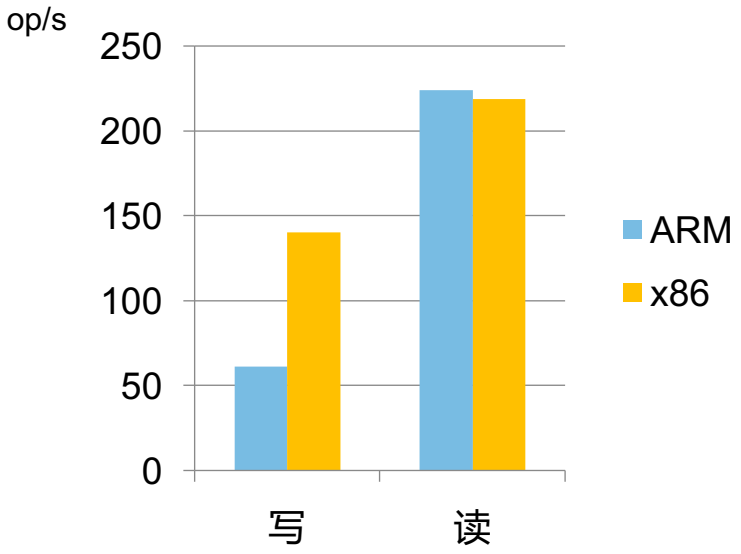
测试结果

后端-4M-128并发，运行300s

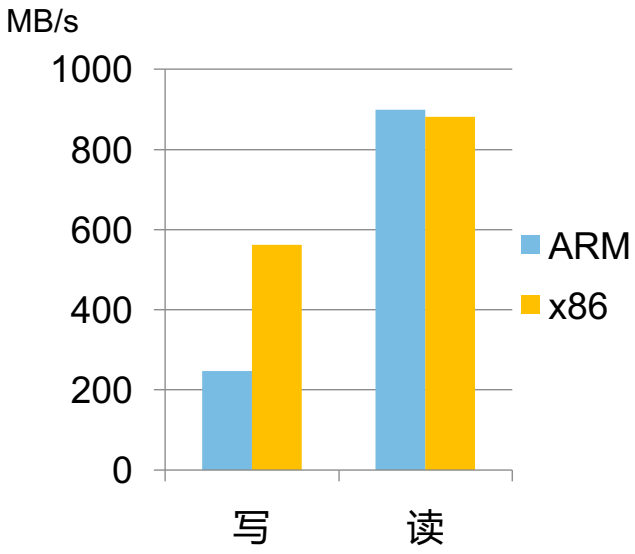
延迟



吞吐量



带宽



三副本配置下，从处理时间、IOPS和带宽综合看来，4M大文件在128并发下的读写性能：
1，x86写性能是ARM的2.2倍
2，ARM的读性能略优于x86

PART 04

总结

总结

从后端测试结果看，ARM的读性能与x86的读性能差别不是很大，而写性能差别较大。
从前端测试结果看，小文件下的ARM性能整体较差，大文件下，ARM的写性能和读性能差距不大，而删除性能较差。
从整体结果看，ARM服务器的读性能还不错，但是写性能和删除性能较差，故建议将其可存放只读数据或者冷数据，用于视频监控、影像数据、备份归档等场景下。

冷存储

视频监控

媒体

备份归档

测试遗留问题

1，三副本测试时，写2000万对象，osd出现down状态
2，纠删码测试时，正常写对象，也会出现osd down的情况，而且恢复较慢

怀疑原因：每个OSD节点分配的2G内存不足，osd peering时，容易出现tcmalloc内存分配失败情况，然后自杀；而纠删码测试中，需要更大的计算量和交互量，对内存的要求会更高。

目前状态：硬件厂商采用4G内存，三副本的问题基本可以解决，纠删码问题尚在优化。
后续ARM服务器会继续深入研究最佳方案，并采用64位处理器，以检测性能是否有所提升。

中国移动 China Mobile | 中移软件 CMSoft

感谢参与

演讲人：黄小曼

公司：中移（苏州）软件技术有限公司