上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

SJTU

Mind the Gap: Broken Promises of CPU Reservations in Containerized Multi-tenant Clouds

王浩宇

2021年12月7日

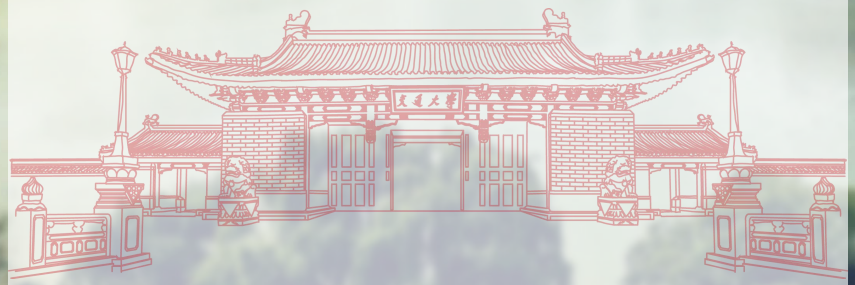
饮 水 思 源 • 爱 国 荣 校

目 录

01

论文简述

Brief Introduction



论文简述

简要内容：

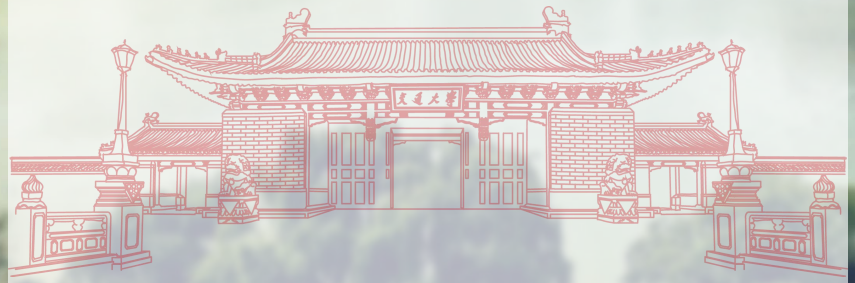
本篇论文首先证明了在容器共存的多租户环境中，性能差异和退化是显著存在的。本文认为与人们普遍认为的资源竞争和干扰导致这种退化的观点相反，主要问题出在容器请求的CPU数量和实际获得的CPU数量之间存在差距，根本原因在于目前的Linux调度机制的设计导致的两个重要问题：

1. 强制运行队列共享 *Forced Runqueue Sharing*
2. 幻影CPU时间 *Phantom CPU Time*

02

强制运行队列共享

Forced Runqueue Sharing



强制运行队列共享

目前Linux系统调度器常用完全公平调度，单核系统中，完全公平调度（CFS）是相当简单的，但是在多核环境下，调度决策将成为一个更加复杂的优化过程。

强制运行队列共享

多核系统中，每个物理CPU核心都有其单独的运行队列。进程将首先被分配给一个运行队列，然后在该CPU上运行。

理想情况：用户为容器请求X个cpu的时，在该容器内生成的所有进程都应该只调度在这X个CPU的运行队列中。

强制运行队列共享

完全公平调度（CFS）：

CFS 调度程序并不采用严格规则来为一个优先级分配某个长度的时间片，而是为每个任务分配一定比例的 CPU 处理时间。每个任务分配的具体比例是根据nice值来计算的。

CFS 调度程序没有直接分配优先级。相反，它通过每个任务的变量 vruntime 以便维护虚拟运行时间，进而记录每个任务运行多久。虚拟运行时间与基于任务优先级的衰减因子有关，更低优先级的任务比更高优先级的任务具有更高衰减速率。

强制运行队列共享

问题：

运行队列之间的负载平衡是影响整个系统性能和CPU利用率的关键，由于CFS的负载平衡活动，一个容器可能被迫与一个和多个相邻容器共享运行队列，从而减少其可用的CPU时间，增加潜在的干扰。

强制运行队列共享

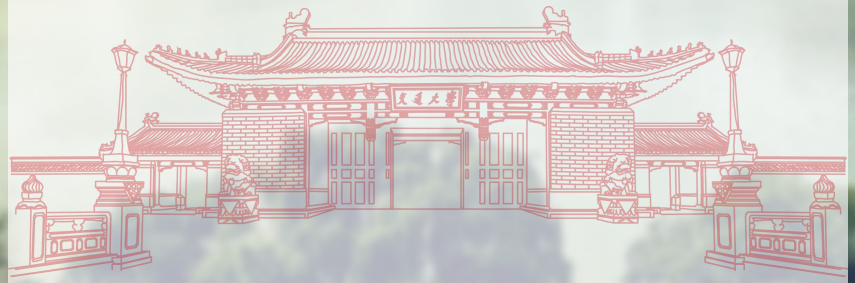
具体产生：

CFS根据多个因素和指标定期平衡每个运行队列的进程。其中一个指标是**负载**，它是由任务的权重（共享）和任务的性质派生出来的。CPU密集型优先级较低，请求大量CPU的容器中的进程不能保证唯一的占用运行队列。

03

幻影CPU时间

Phantom CPU Time



幻影CPU时间

目前Linux调度程序设计中，会为每个时间片保持“最小粒度”，它确定了进程在被抢占之前需要运行的最小时间。

幻影CPU时间

原因：

防止切片太短，频繁调用调度程序增加开销。

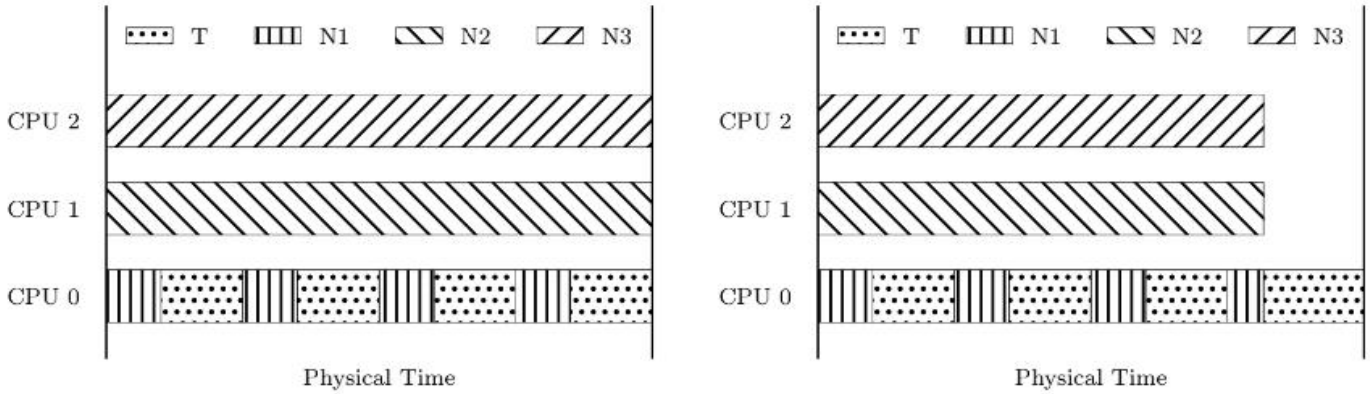
最小粒度由几个调度器参数决定：

sched_min_granularity_ns、sched_latency_ns 、sched_wakeup_granularity_ns

来自多个容器的进程共享同一运行队列的事实可能导致幻影CPU时间（Phantom CPU Time），即容器似乎能利用可用的CPU时间，但实际上不能，进一步导致了性能下降。

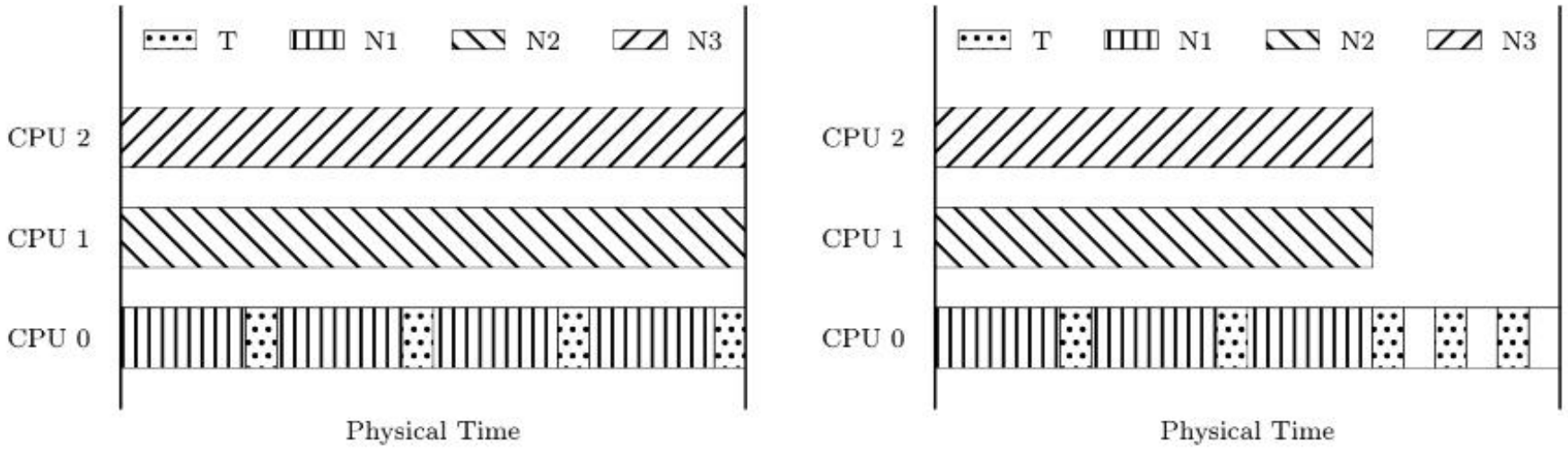
幻影CPU时间

Container	*requests.cpu*	*cpu.shares*	Task (thread)	Task weight
Target (T)	1	1024	T	1024
Neighbor (N)	2	2048	N1, N2, N3	683, 683, 683



(a) w/ burstable neighbors. **(b) w/ capped neighbors.**

幻影CPU时间


T
N1
N2
N3
CPU 2
CPU 1
CPU 0
Physical Time
T
N1
N2
N3
CPU 2
CPU 1
CPU 0
Physical Time

幻影CPU时间

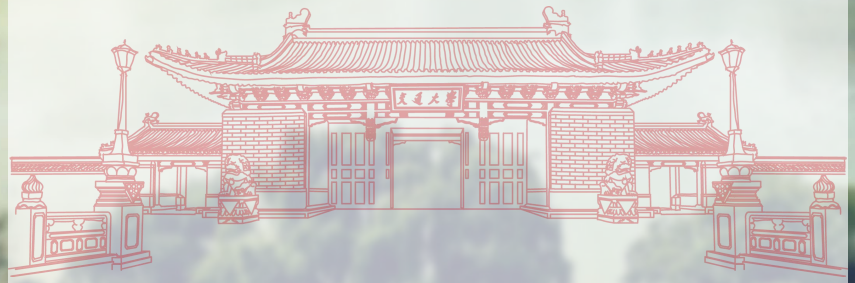
关键：

由于FRS和PCT，目标容器无法充分利用所有保留的CPU资源。这是现代调度器中各种设计考虑因素之间交互的结果，当涉及到运行容器化工作负载的系统时，来自容器的资源保留需求与现代调度器中所做的设计选择不兼容。

04

RKube

RKube



RKube

解决方案：

RKube，它通过引入当前缺失的参数来增强Kubernetes，允许用户向kubelet表达他们的需求，它依靠CPU预留(而不是动态权重调整)来实现资源分配和调度。

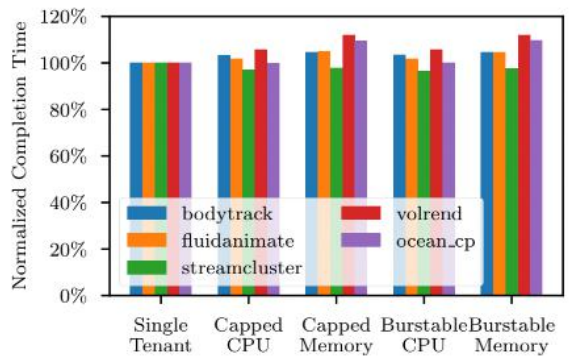
RKube

CPU保留：

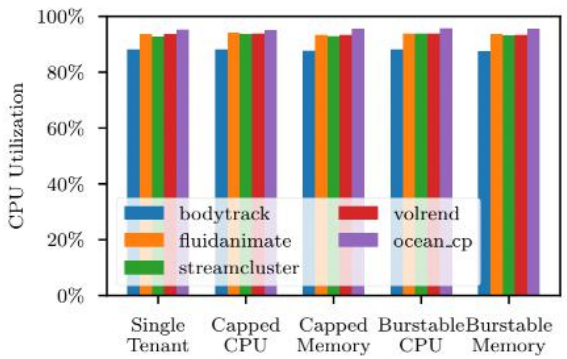
通过为单个容器设置CPU亲和性来实现CPU预留。为此，Linux提供了一个名为cgroups的控制特性，RKube可以通过设置cpuset.cpus来利用它。

cpu.shares提供了组之间的相对CPU共享，cpuset.cpus限制CPU使用的绝对值，这是独立于CPU速度和相邻容器的调度。

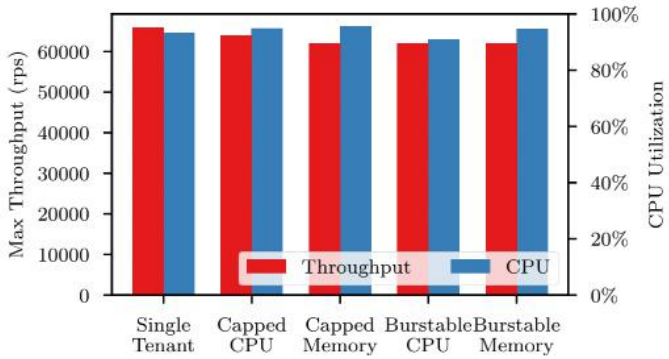
RKube



(a) Batch apps, performance.



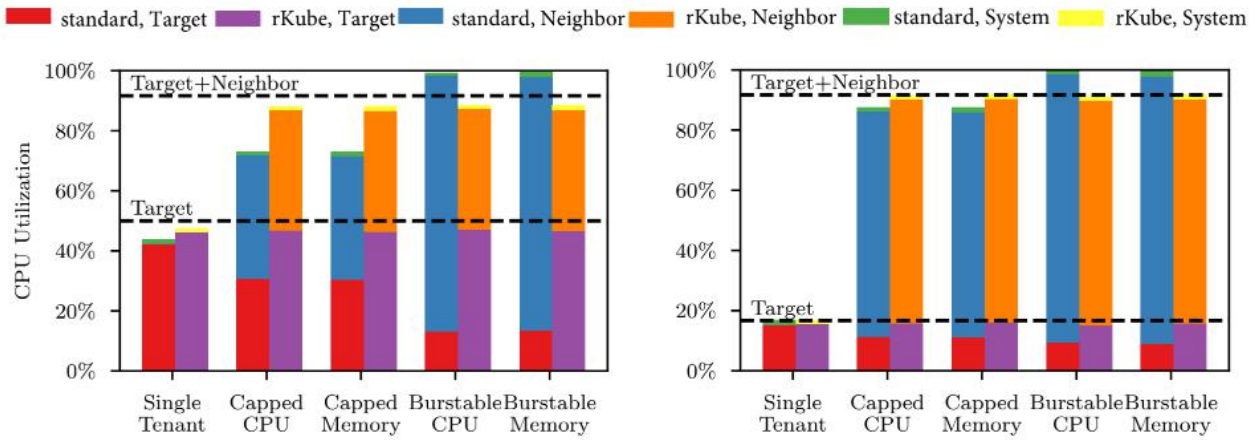
(b) Batch apps, CPU utilization.



(c) Memcached, performance & CPU util.

Figure 6: Performance and CPU utilization of target container with *rKube*. (*standard* result in Figure 2)

RKube



(a) Batch app (streamcluster). (b) Memcached.

Figure 7: Composition of overall host CPU utilization with *rKube* and *standard*. Dotted lines indicate the CPU allocated to target/neighbors.

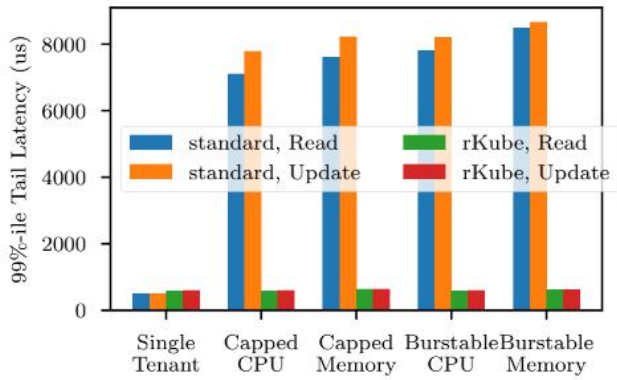
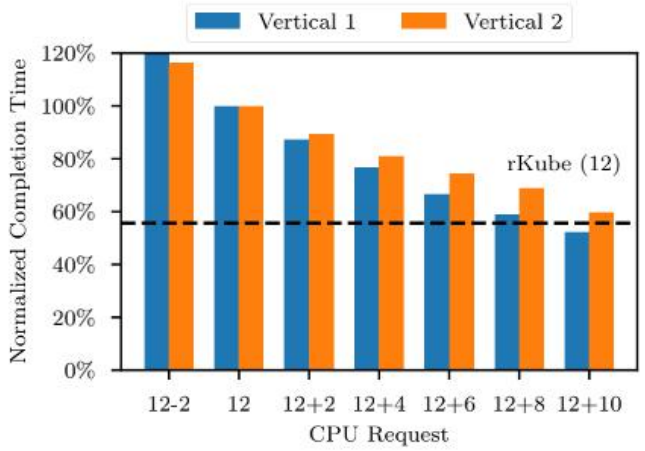


Figure 8: rKube vs. standard, Memcached tail latency.

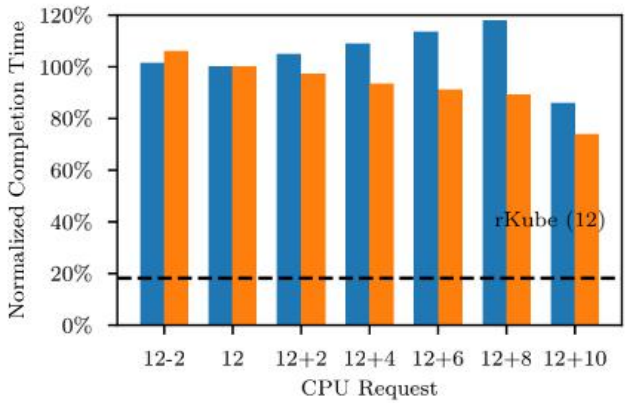
vs 垂直缩放：

当前实践中，当用户将应用程序部署到云中，而应用程序性能不令人满意时，用户通常需要投入更多的资源来提高其性能，这通常被称为垂直缩放。

RKube



(a) Capped & Mem-Intensive neighbors.



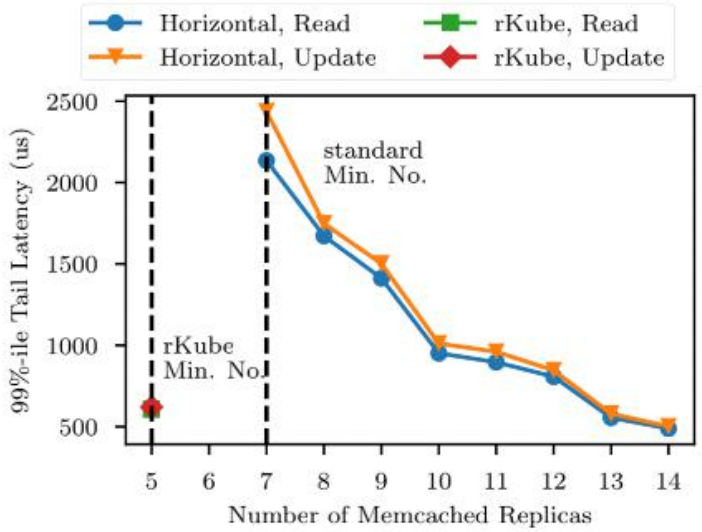
(b) Burstable & Mem-Intensive neighbors.

vs 水平缩放：

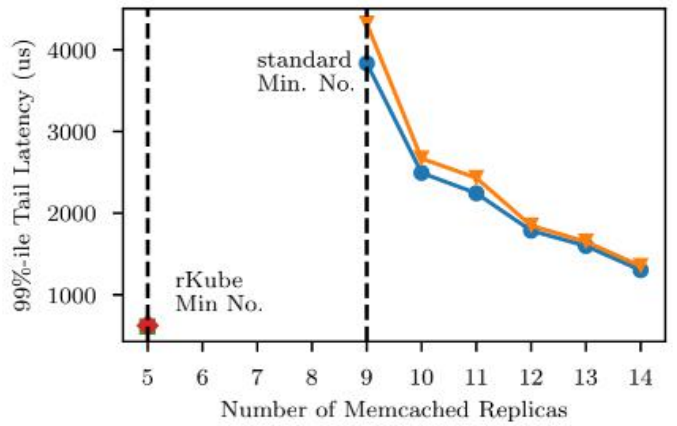
对于交互式应用程序，当应用程序遇到性能瓶颈时，通常会使用水平缩放。

水平缩放通常需要配置和提交额外的基础设施容量，以实现所需的性能。

RKube



(a) Capped & Mem-Intensive neighbors.



(b) Burstable & Mem-Intensive neighbors.

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

感谢聆听

饮水思源 爱国荣校

纯文字版式三

添加观点1：

请在此输入文字说明
请在此输入文字说明
请在此输入文字说明
请在此输入文字说明
请在此输入文字说明。

添加观点2：

请在此输入文字说明
请在此输入文字说明
请在此输入文字说明
请在此输入文字说明
请在此输入文字说明。

添加观点3：

请在此输入文字说明
请在此输入文字说明
请在此输入文字说明
请在此输入文字说明
请在此输入文字说明。

纯文字版式四

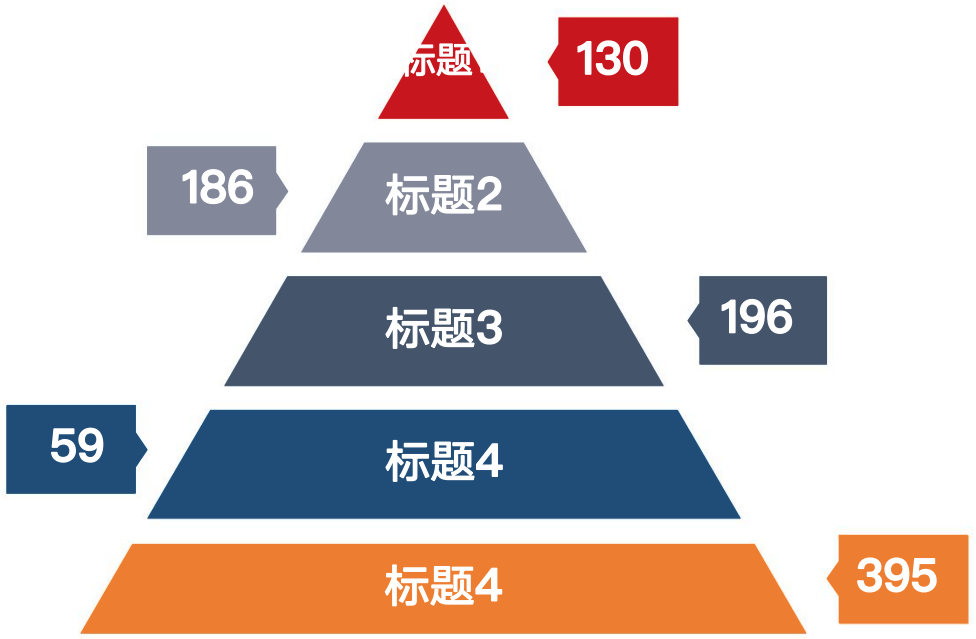
请在此输入文字说明
请在此输入文字说明

请在此输入文字说明
请在此输入文字说明

请在此输入文字说明
请在此输入文字说明

请在此输入文字说明
请在此输入文字说明

添加页面标题内容



添加观点1：

请在此输入文字说明请在此输入文字说明请在此输入文字说明请在此输入文字说明请在此输入文字说明。

此处可添加一段说明性文字：

请在此添加文字内容请在此添加文字内容请在此添加文字内容请在此添加文字内容请在此添加文字内容。

添加页面标题内容

- 请在此输入文字说明请在此输入文字说明请在此输入文字说明请在此输入文字说明请在此输入文字说明请在此输入文字说明。

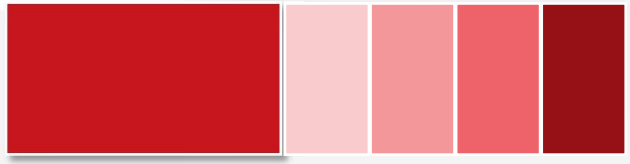
标题1	标题2	标题3	标题4	标题5

请在此输入文字说明

请在此输入文字说明

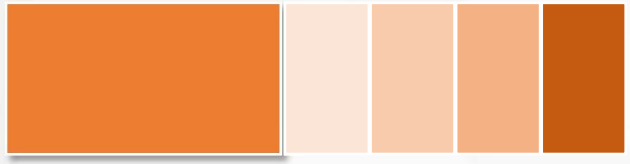
色彩规范

主色



对比色

平衡色



浅色与深色

建议尽量选择以上调色板中的颜色

字体规范

中文标题

微软雅黑

中文正文

微软雅黑

英文标题

Arial

英文正文

Arial

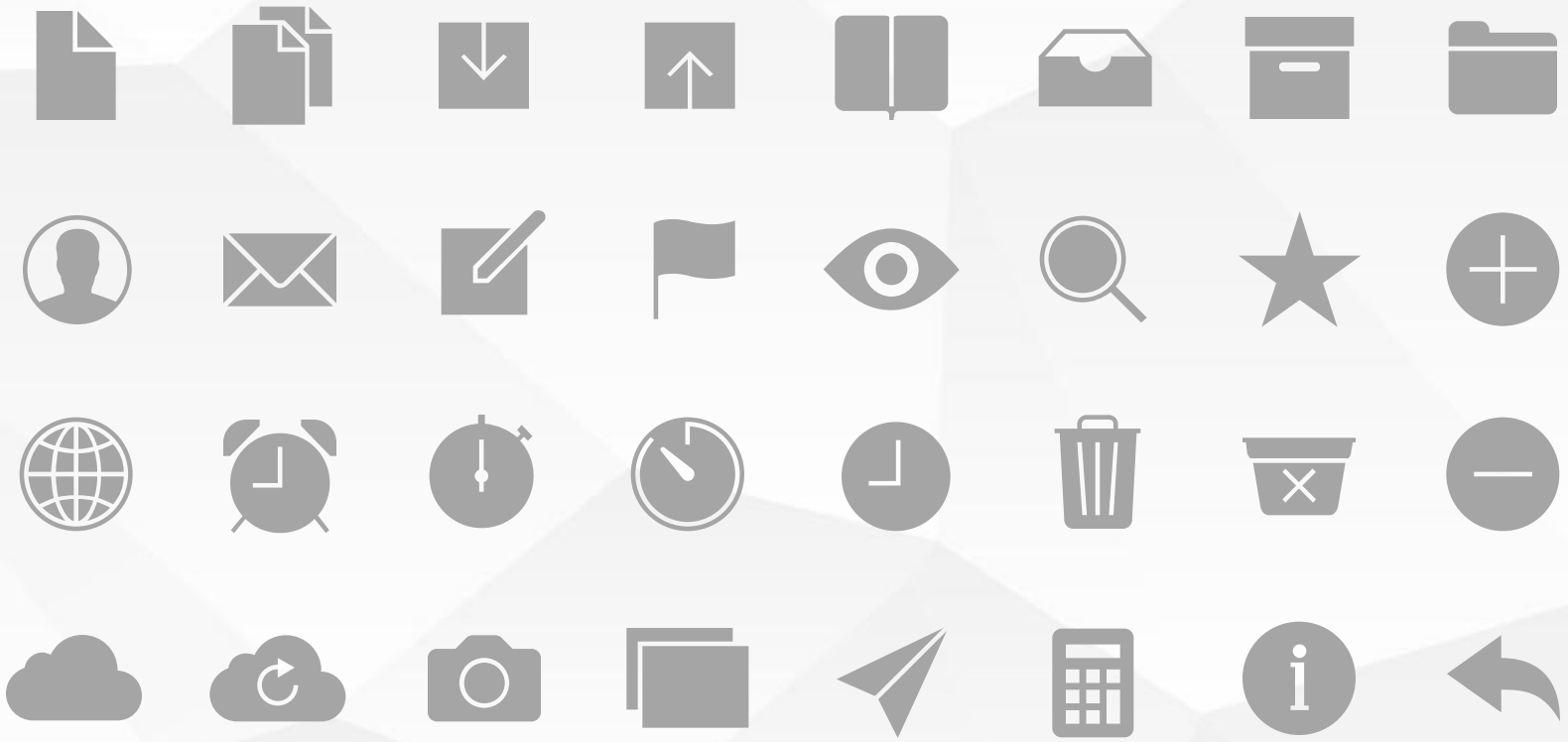
注意：

微软雅黑属于版权字体，商用请购买！

更多免费商用字体
https://jbox.sjtu.edu.cn/l/WuCIHQ

图标

图标



标注

使用说明

本PPT模板为作者原创，著作权归作者所有。
您仅可以个人非商业用途使用本PPT模板，未经权利人书面明确授权，不可将信息内容的全部或部分用于出售，或以出租、出借、转让、分销、发布等其他任何方式供他人使用，否则将承担法律责任。

声明

OfficePLUS尊重知识产权并注重保护用户享有的各项权利。
OfficePLUS拥有对本PPT模板进行展示、报道、宣传及用于市场活动的权利，若在比赛或商业应用过程中发生版权纠纷，其法律责任由作者本人承担。