3.3 SMP 负载均衡

在阅读本章之前请思考如下几个小问题：

1. *一个 4 核处理器里每个物理 CPU 核并且有独立 L1 cache 并且只有一个线程，分成 2 个族 cluster0 和 cluster1，每个族包含 2 个物理 CPU 核，族里的 CPU 核共享 L2 cache。请画出该处理器在 Linux 内核里的拓扑关系图。*
2. *调度组能力系数 capacity 做什么用途？如果一个调度组里的正在运行的进程数量大于调度组里的调度能力因子 group_capacity_factor 说明什么情况，反之又什么情况？*
3. *如果查找出一个调度域里那个调度组最繁忙？*
4. *一个调度域如果负载不均衡，那么如何计算需要迁移多少负载量呢？*

1 CPU 域初始化

物理 CPU 域可以根据实际物理属性分成如下几类。

CPU 分类	Linux 内核分类	说明
超线程（SMT）	CONFIG_SCHED_SMT	超线程使用相同 CPU 资源并且共享 L1 cache，迁移进程不会影响 Cache 利用率。
多核（MC）	CONFIG_SCHED_MC	单个物理 CPU 有多个物理核心，每个核心独享 L1 cache，同一个 CPU 的多个核心共享 L2/L3 cache
处理器（CPU）	内核称为 DIE	SoC 级别

内核有一个数据结构来描述这种 CPU 的层次关系 struct sched_domain_topology_level，通常简称为 tl。

[include/linux/sched.h]

```
struct sched_domain_topology_level {
	sched_domain_mask_f mask;
	sched_domain_flags_f sd_flags;
	int		flags;
	int		numa_level;
	struct sd_data	data;
#ifdef CONFIG_SCHED_DEBUG
	char		*name;
#endif
};
```

另外内核默认定义了一个数组 default_topology[]来概括 CPU 域的层次结构。

[kernel/sched/core.c]

```
/*
 * Topology list, bottom-up.
 */
static struct sched_domain_topology_level default_topology[] = {
#ifdef CONFIG_SCHED_SMT
	{ cpu_smt_mask, cpu_smt_flags, SD_INIT_NAME(SMT) },
```