Placement of Virtual Containers on NUMA Systems

Justin Funston*, Maxime Lorrillere†, and Alexandra Fedorova, University of British Columbia
Baptiste Lepers, EPFL
David Vengerov and Jean-Pierre Lozi, Oracle Labs
Vivien Quéma, IMAG

王昊 M201873144

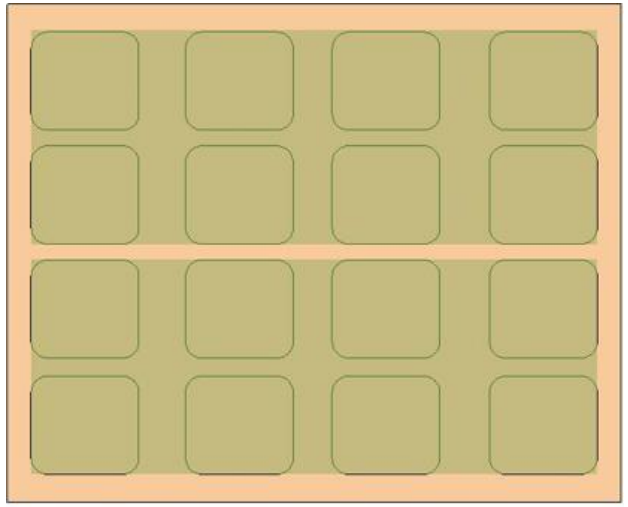
Motivation

Data Centers:

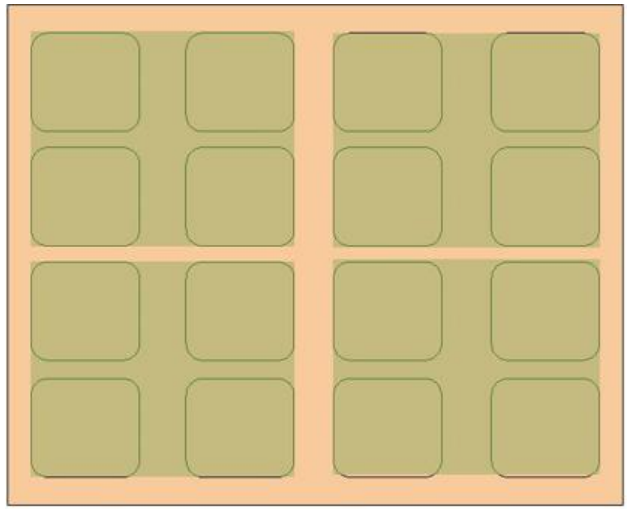
3% of global electricity usage

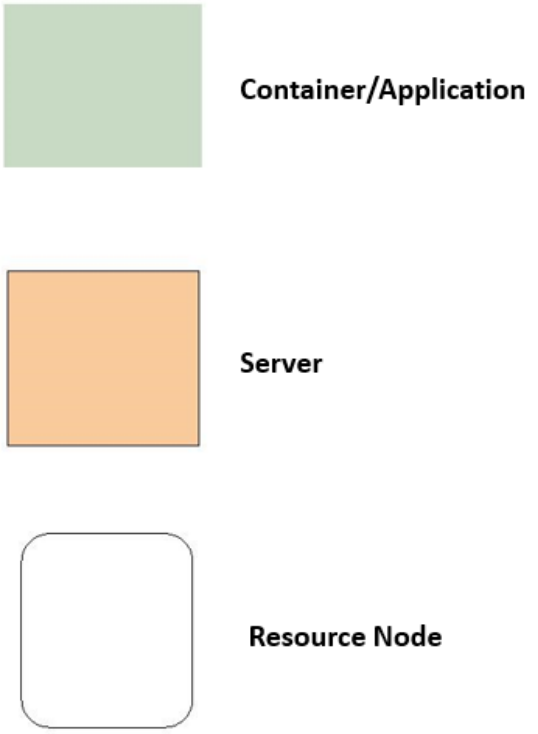
86% used by servers+cooling

Motivation – Server Packing



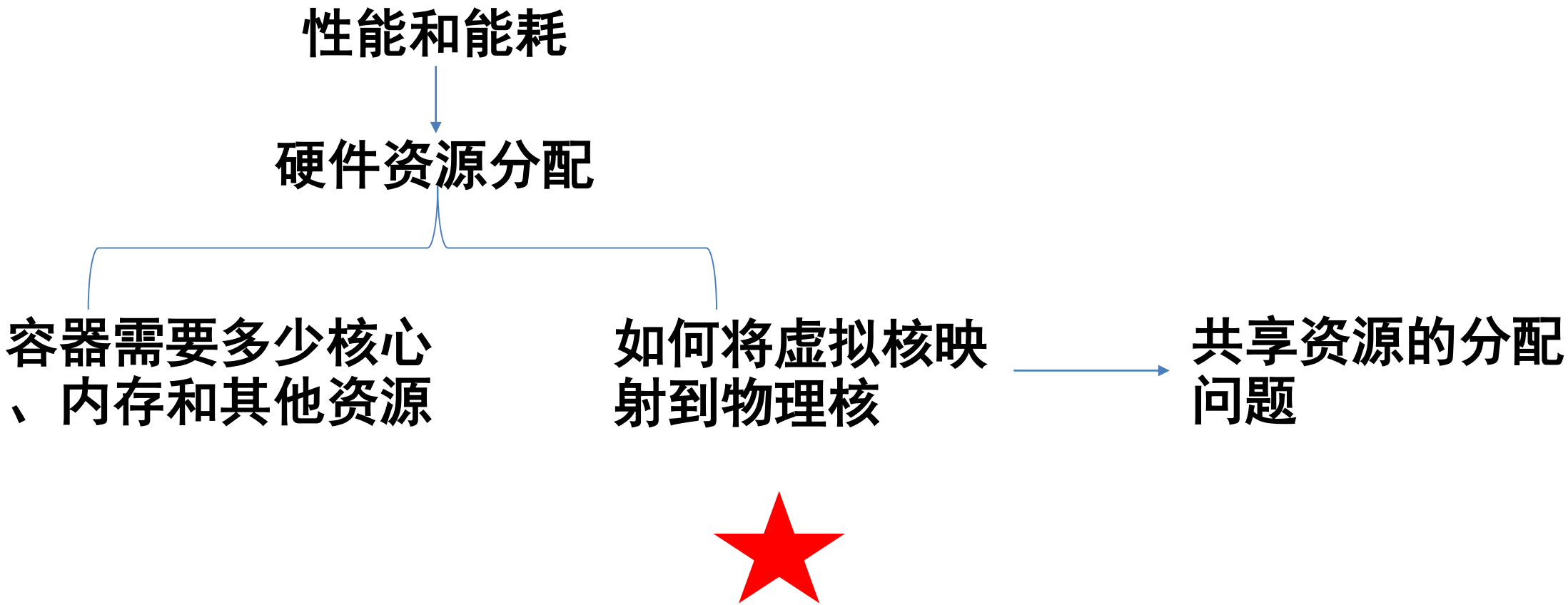
VS.



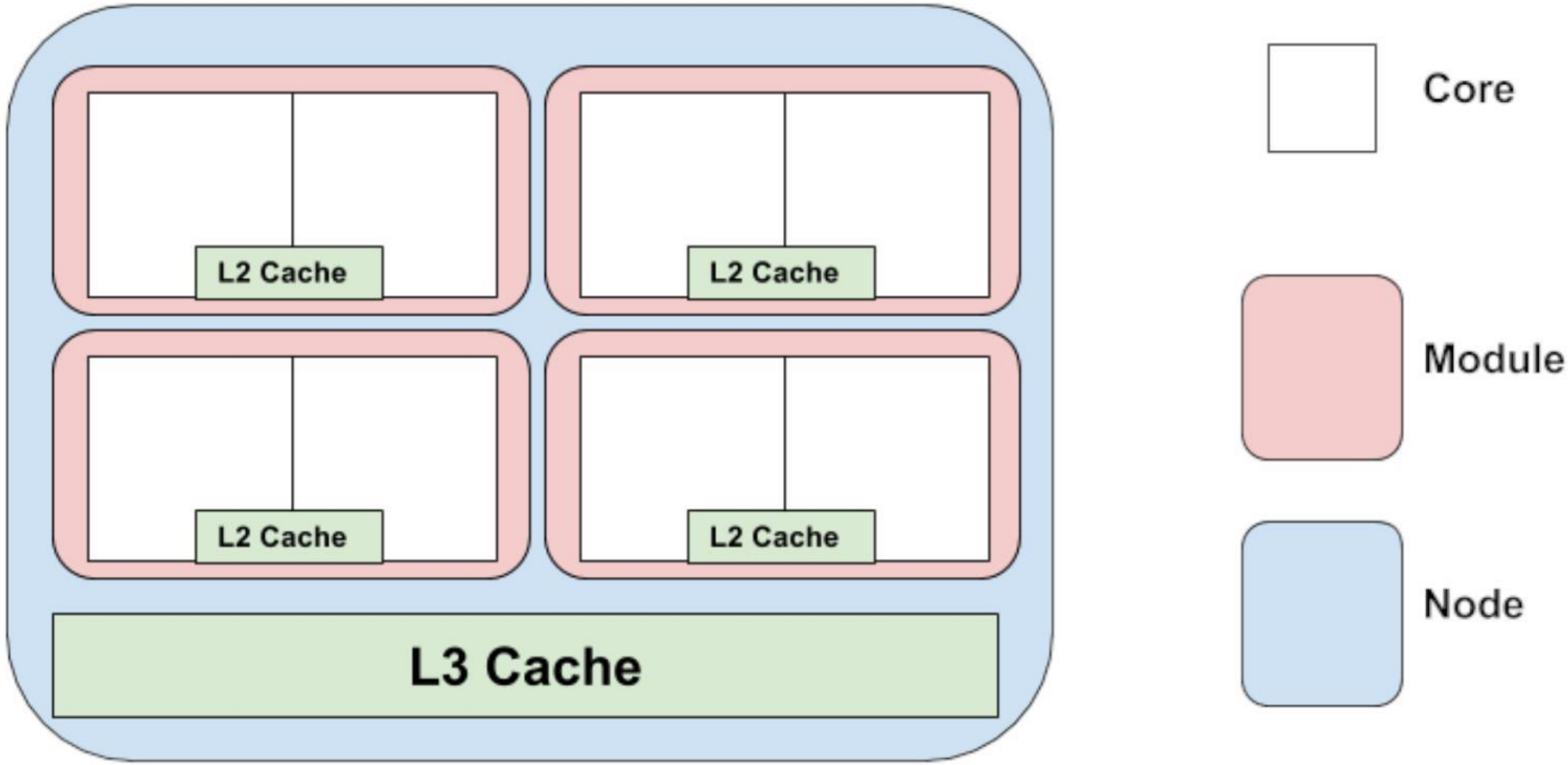


- **Half as many servers!**
- **Half as much energy!**
- **Half as much infrastructure!**
- **Performance?**

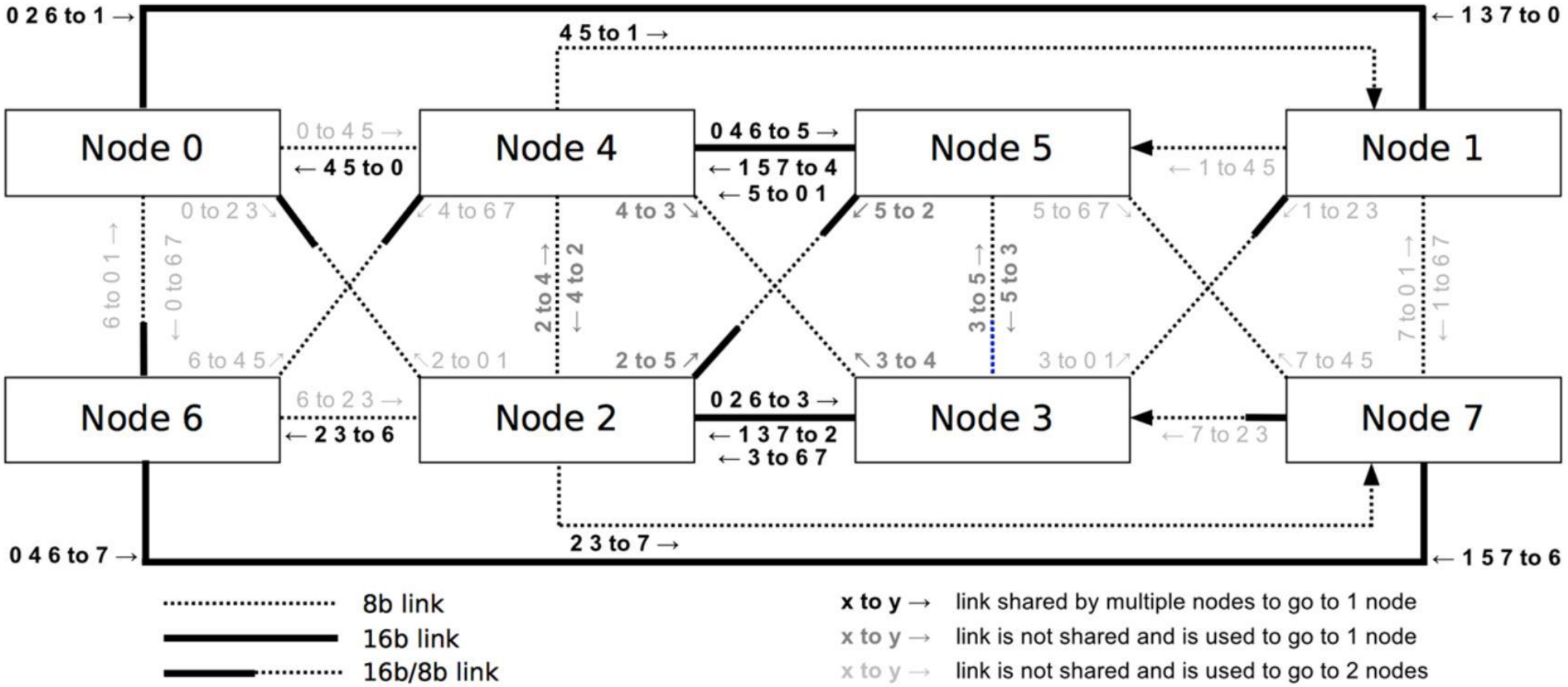
Motivation – Server Packing



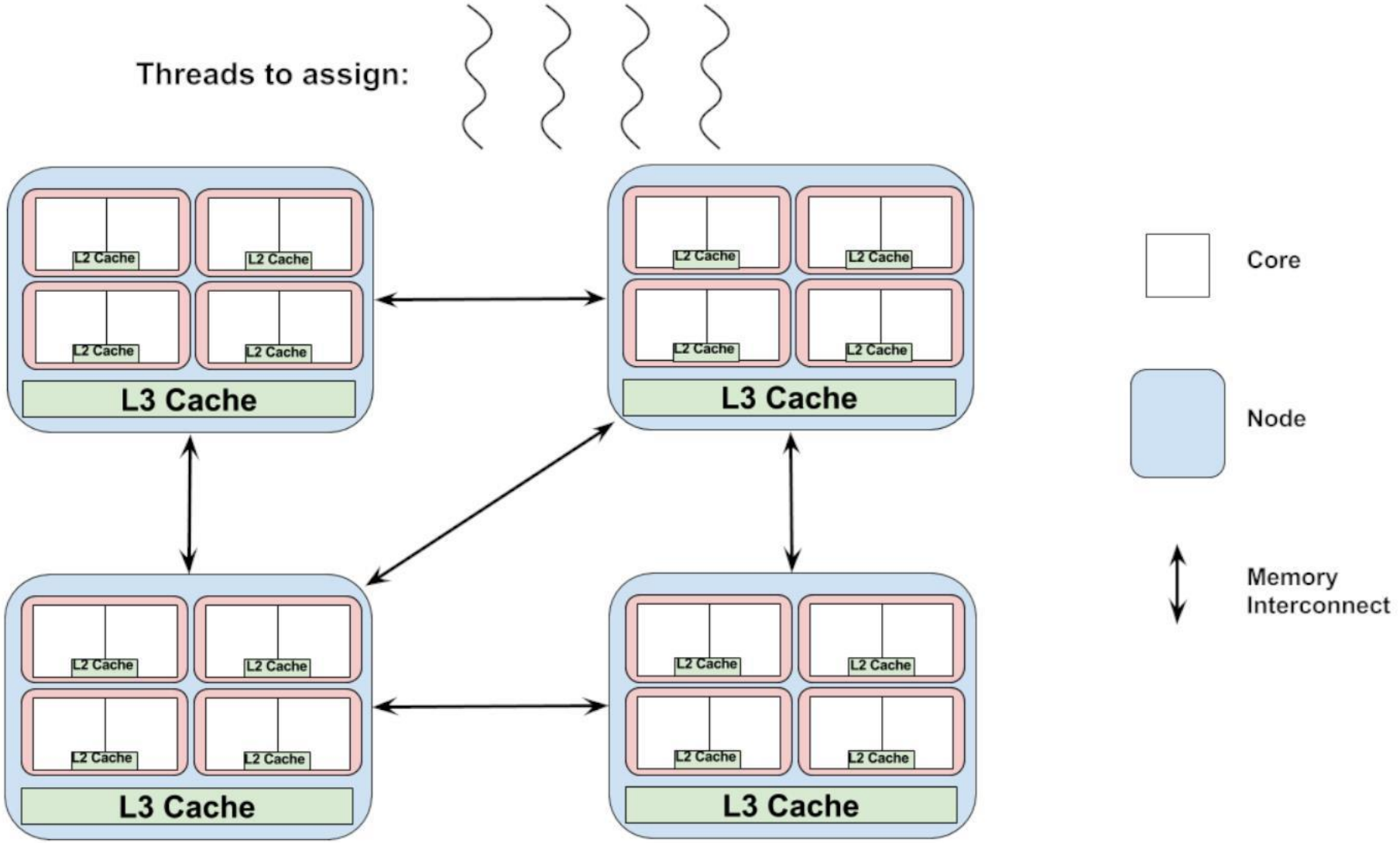
Motivation – NUMA Node



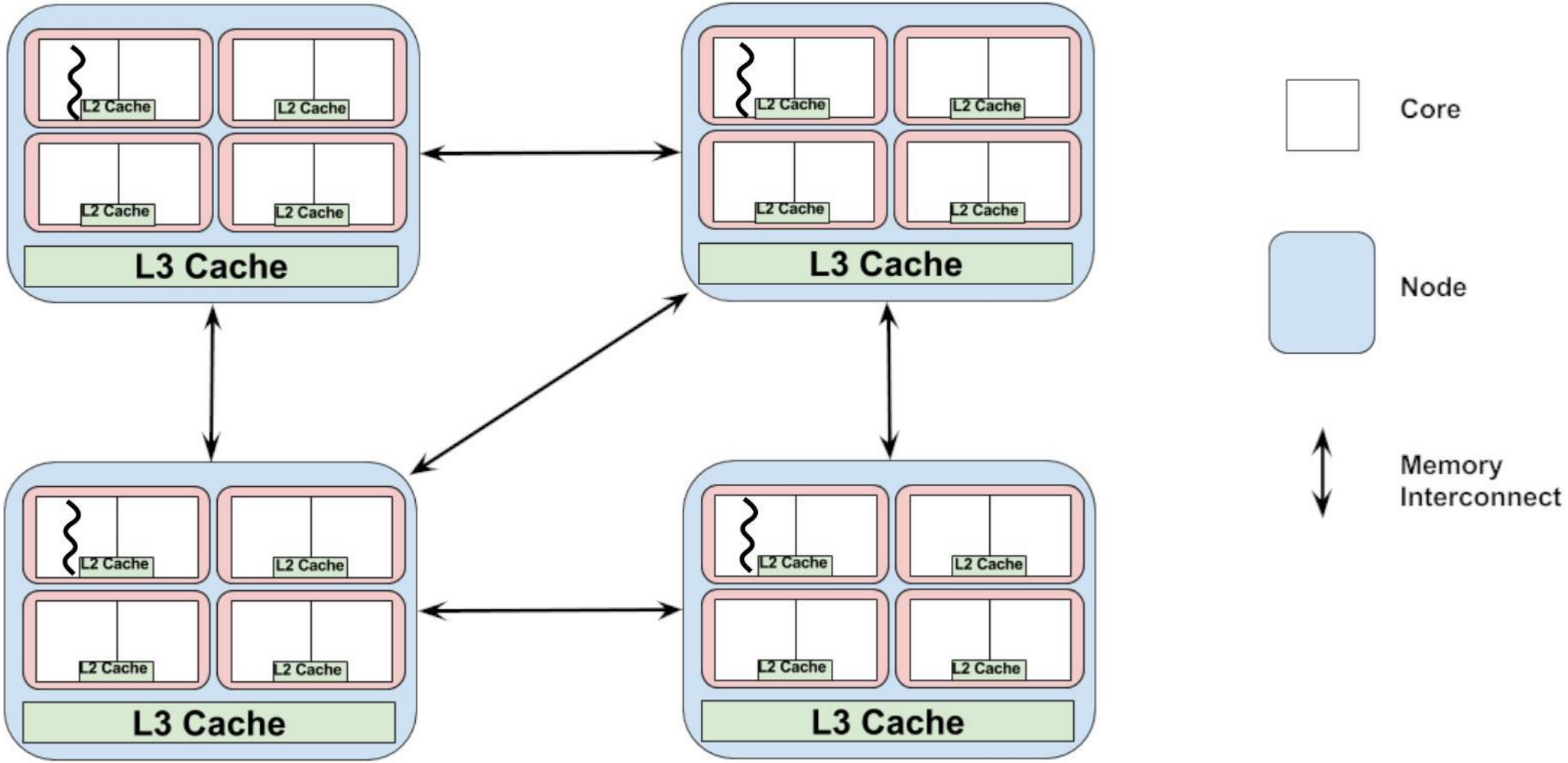
Motivation – Interconnect Topology



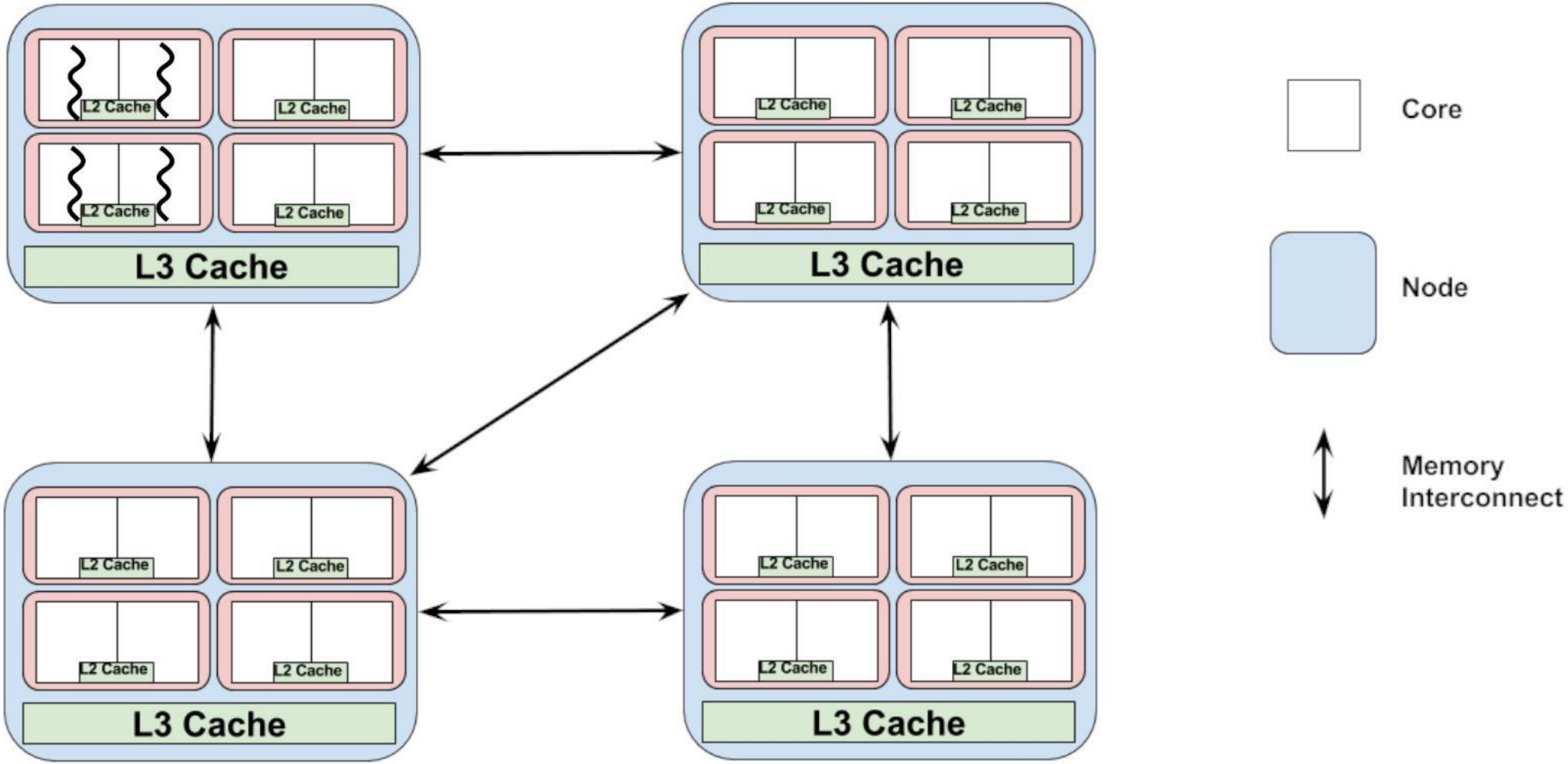
Example of placements



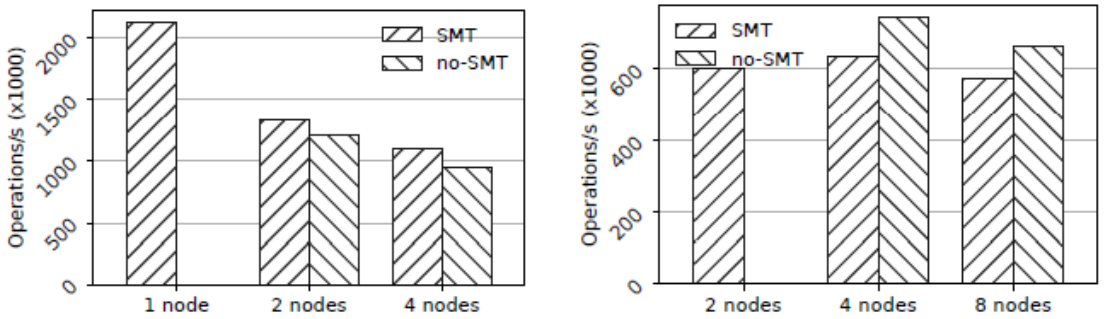
Example of placements



Example of placements



Motivation – Server Packing



(a) WiredTiger, Intel (b) WiredTiger, AMD

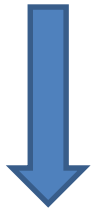
Figure 1: Throughput of the WiredTiger key-value store on two NUMA systems.

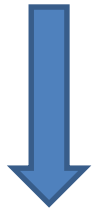
资源共享存在竞争（线程竞争缓存空间和硬件队列）和协作（线程为批次预取数据）

在Intel系统更快的通信和协作资源共享的好处超过了的资源争用成本

Solution steps

找出重要配置 → 预测配置性能





Abstract machine model

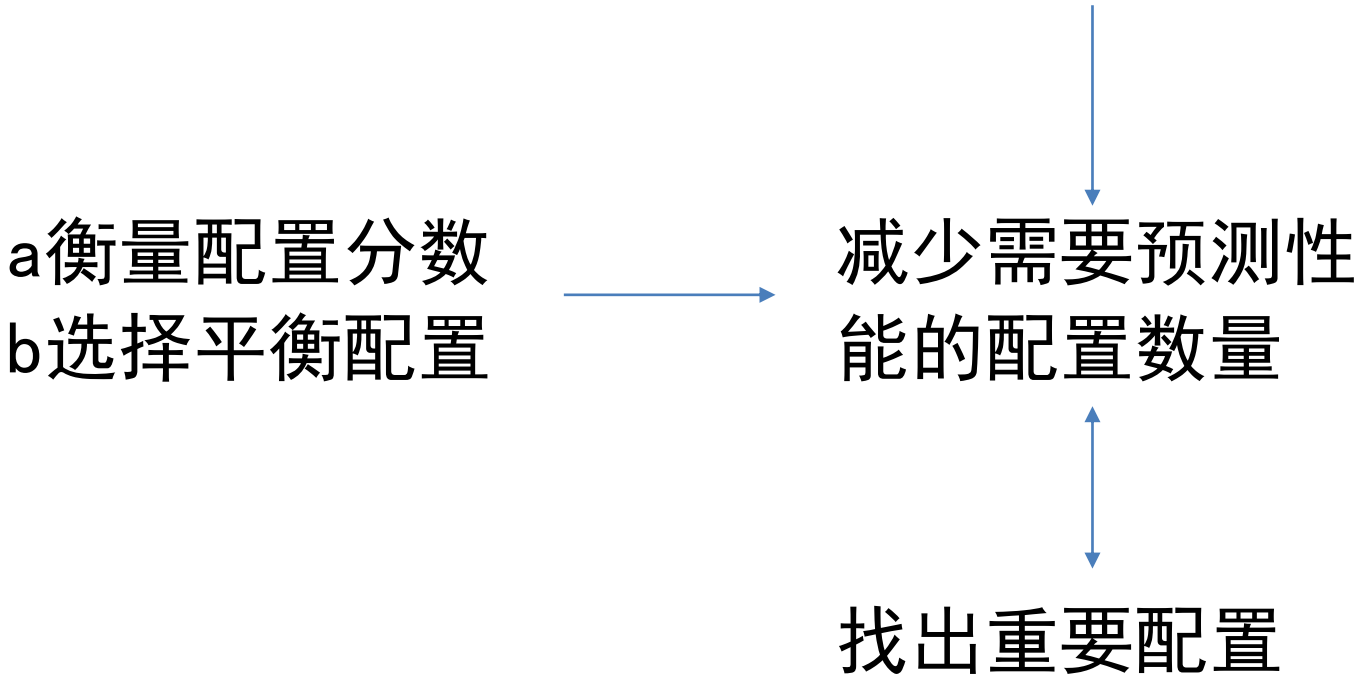
Performance prediction model

Assumptions

- 相同分数的配置会产生相同的效果
- 工作负载封装在虚拟容器中
- NUMA节点是资源分配的单位
- 只考虑平衡的配置

Abstract machine model-what to do

考虑在64核的系统上分配16个虚拟核给容器，将近有10^{14}种



Abstract machine model-meaning of “score” and “balance”

分数表示特定资源的静态利用率，分数向量表示多种共享资源的利用率

例：

对于L2缓存资源，如果在配置中，所有虚拟核共享单个L2高速缓存，则L2高速缓存的分数将等于1。

如果核心分布在两个L2缓存上，则分数将等于2

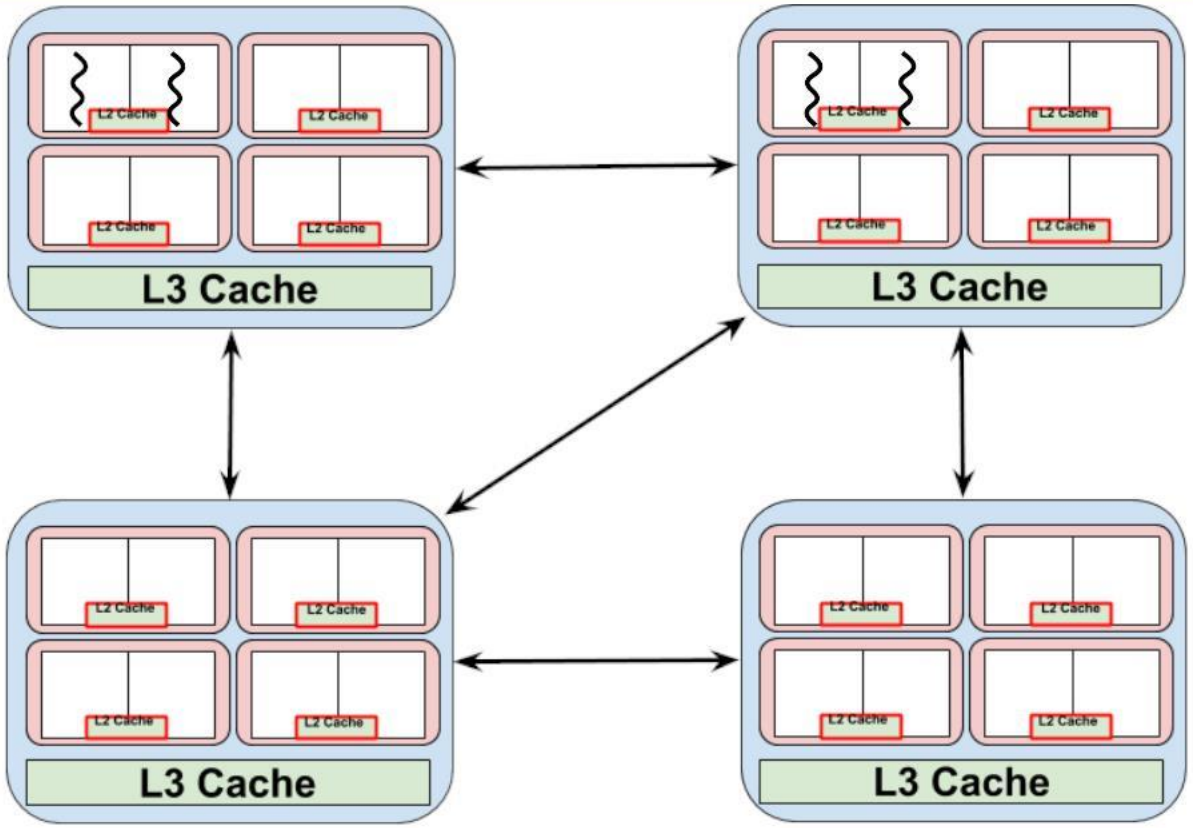
对于非对称资源，例如某些系统上的跨芯片互连，计算配置使用的所有链接的总可用带宽作为分数

平衡配置：vCPU的数量可被配置的共享资源单元均分

例：如果在系统上共享L3缓存，将只考虑每个L3缓存的vCPU数量相等的配置

Scheduling Concern Example – L2

Placement:



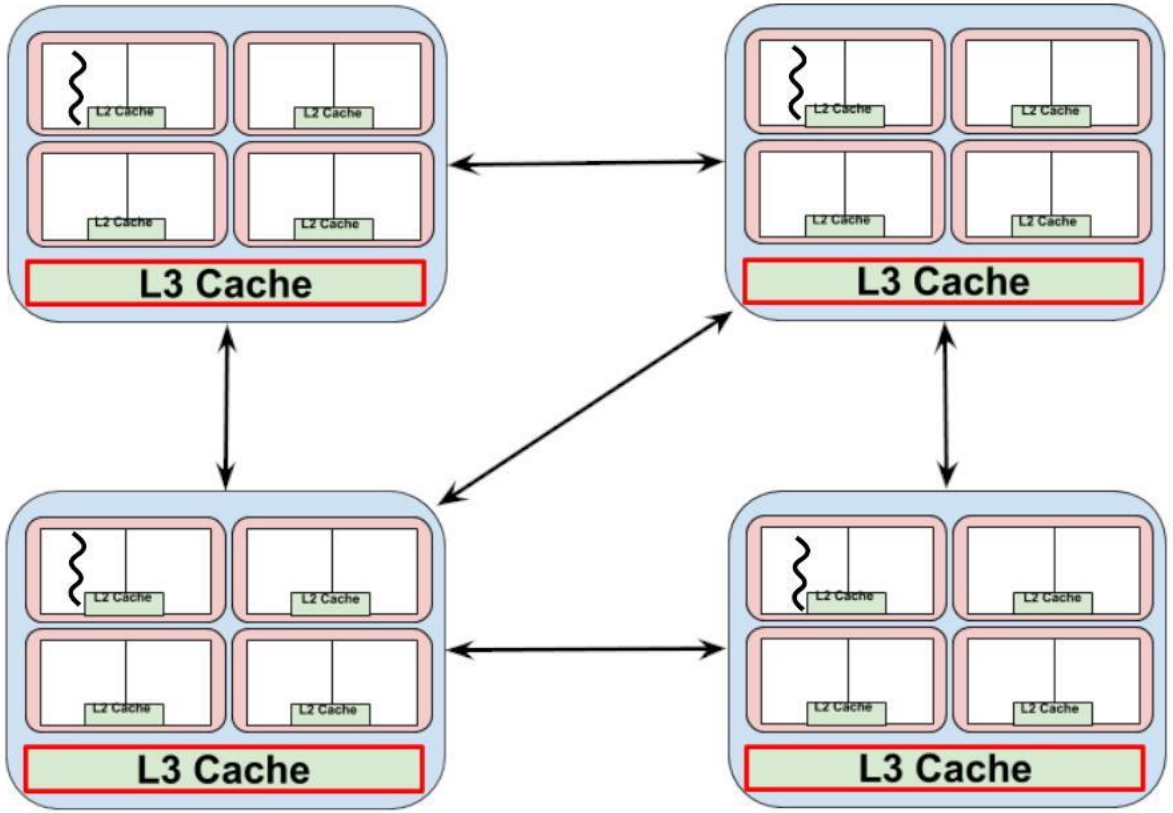
Scheduling Concern:

Score = # L2 Caches In-Use

2

Scheduling Concern Example – L3

Placement:



Scheduling Concern:

Score = # L3 Caches In-Use

4

Abstract machine model-how to do

衡量配置分数

具有相同分数向量的配置在资源共享方面性能相同，可以丢掉重复配置

丢弃严格较低分数的配置（限于网络带宽）

由于希望用户能够进行成本 - 性能权衡，因此分数较低但成本较低的配置仍然可选

对于某些工作负载（协作缓存共享），较小的分数（使用较少的L2缓存）实际上可以提高性能

Abstract machine model-algorithm

找出重要配置

满足条件

1符合平衡条件

$$v \bmod s = 0$$

2可行即不为一个硬件线程分配多个vCPU

$$v / s \leq capacity$$

3不能被更好的配置取代

算法步骤

1找出满足平衡可行条件的共享资源数量

2获取所有配置

3丢弃相同分数配置，根据带宽丢弃较低分数配置

Performance Prediction Model-purpose

- **outcome**

性能结果由性能向量表示

如果有三种配置，并且第二第三配置的性能比第一个基准配置的性能高20%和30%，则性能向量将为：[1.0,0.8,0.7]

- **Model-building methodology**

随机森林回归器（RF）

机器学习技术，能够在很少或没有调谐的情况下学习非线性函数

Performance Prediction Model – Features/Inputs

- **hardware performance events (HPEs)**

手动选择——硬件架构的复杂性及其对软件的影响

自动选择——数量多达几百，甚至超过一千

折中方案：先选择一些HPE，再通过SFS选择最优的

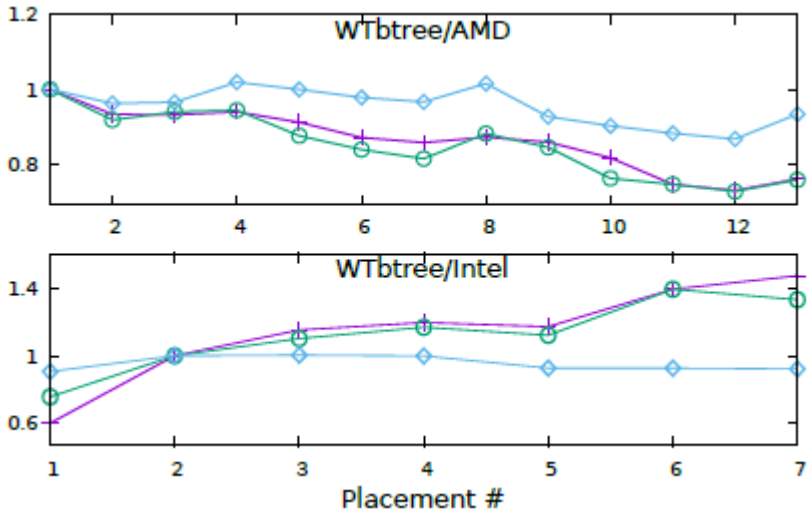
采用某个配置的HPE矢量输入，输出性能矢量

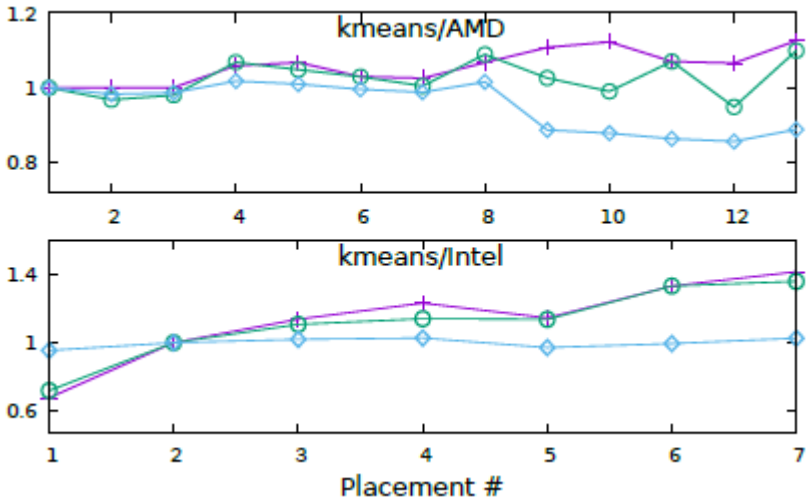
- **performance measurement**

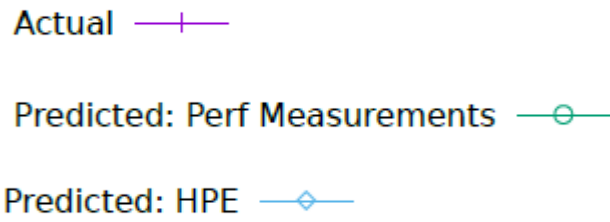
选择两种配置测量其性能

测量的两种配置性能输入，输出性能矢量

Performance Prediction Model -accuracy of predictions







结论：根据性能测量进行预测的准确性更高

原因：

使用HPE很难将延迟的敏感度与总体内存密集度分开

仅测量单个配置上的HPE，很难确定给定数量的L3缓存是否适合工作负载

Evolution

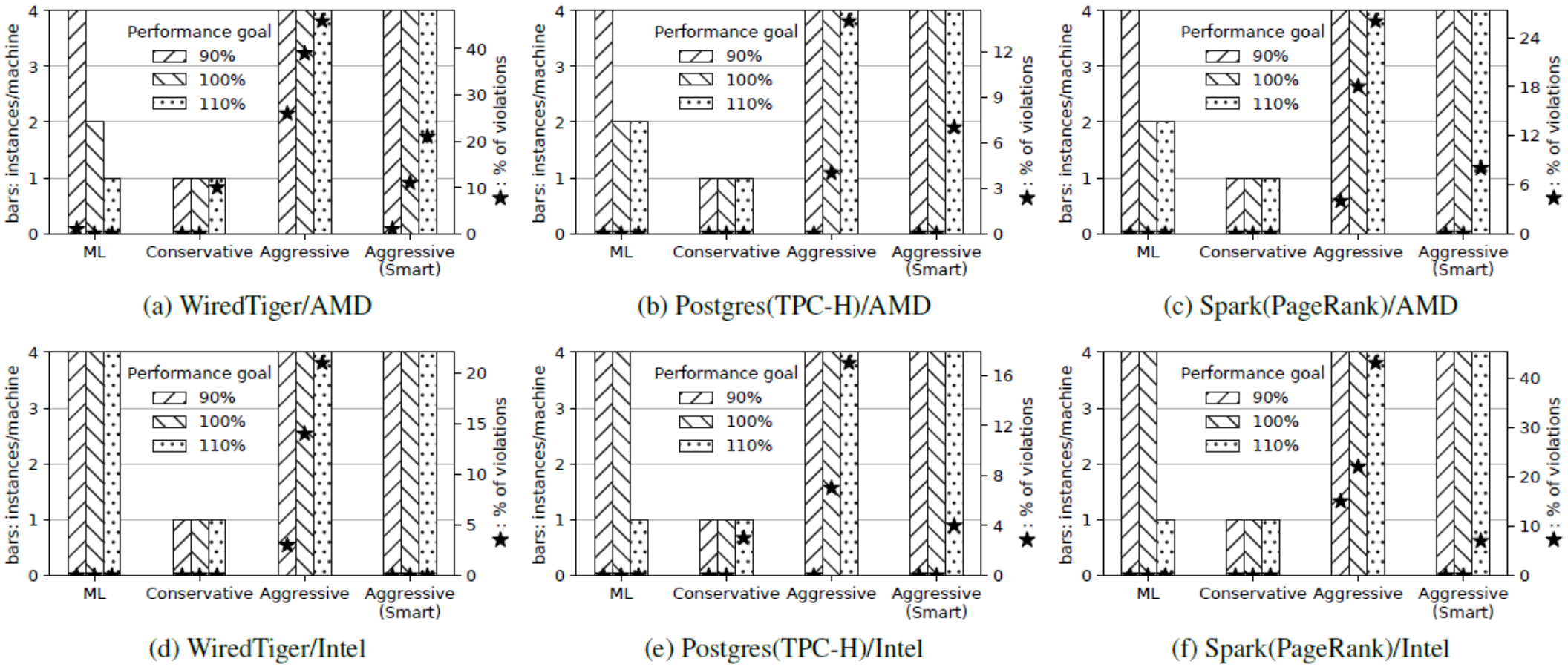


Figure 5: Instances per machine (left y-axis, higher is better) and % performance goal violation (right y-axis, lower is better).

Evolution

ML：本文介绍的策略
Conservative：将整个物理机分配给每个容器实例，每台机器只允许一个容器
Aggressive：尽可能多地填充容器，从而最大限度地提高机器利用率
Smart-Aggressive：类似于Aggressive，但每个实例都固定到最佳的最小节点集（定义为具有最高的互连带宽）

ML策略满足性能目标并且违规最小

Conservative策略违规原因：共享资源不均衡地分配给虚拟核，从而导致争用

Smart-Aggressive性能较低原因：没有考虑所有的配置

Evolution-container migration

memory migration

ML策略需要先获得两种配置的性能测量，再迁移至最佳性能配置，因此会有容器内存迁移

How to do

Freezing the container：迁移时间短，宕机时间长，适合非延迟敏感工作

throttle the bandwidth：迁移时间长，宕机时间段，适合延迟敏感工作

Workload Placement – Related Work

	Predicts Performance	Multiple Hardware Resources	Easily Adapted	Deployable Online
Our Solution	✓	✓	✓	✓
Pandia (EuroSys '17)	✓	✓	✗	✗
SMiTe (Micro '14)	✓	✓	✗	✓
Bubble-Flux (ISCA '13)	✓	✓	✗	✓
Asymsched (ATC '15)	✗	✗	✓	✓
DINO (ASPLOS '10)	✗	✗	✓	✓
Thread Clustering (EuroSys '07)	✗	✗	✓	✓

Smart-Aggressive

Conlusion

本文介绍的系统能够预测容器的虚拟核映射到物理核的性能，并给出最佳性能映射配置。

该系统拥有以下特点：
考虑了多种共享资源如缓存，存储，带宽；
先实际测量两种映射配置的性能作为输入再预测，相比于比直接利用硬件性能事件，不仅性能预测准确率增高，而且更能适用于不同的NUMA系统，另外不需要大量的人力，也能实际应用。

Thanks