SMiTe: Precise QoS Prediction on Real-System SMT Processors to Improve Utilization in Warehouse Scale Computers

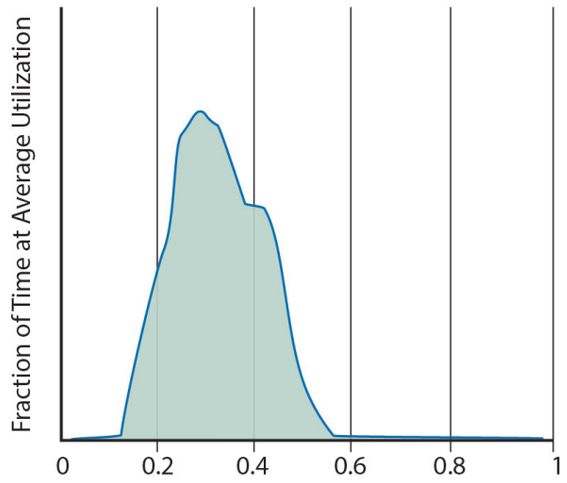
Yunqi Zhang, Michael A. Laurenzano, Jason Mars, Lingjia Tang

Clarity-Lab
Electrical Engineering and Computer Science
University of Michigan

Houston, we have a problem

- Warehouse scale computers are expensive
- Host large-scale Internet services

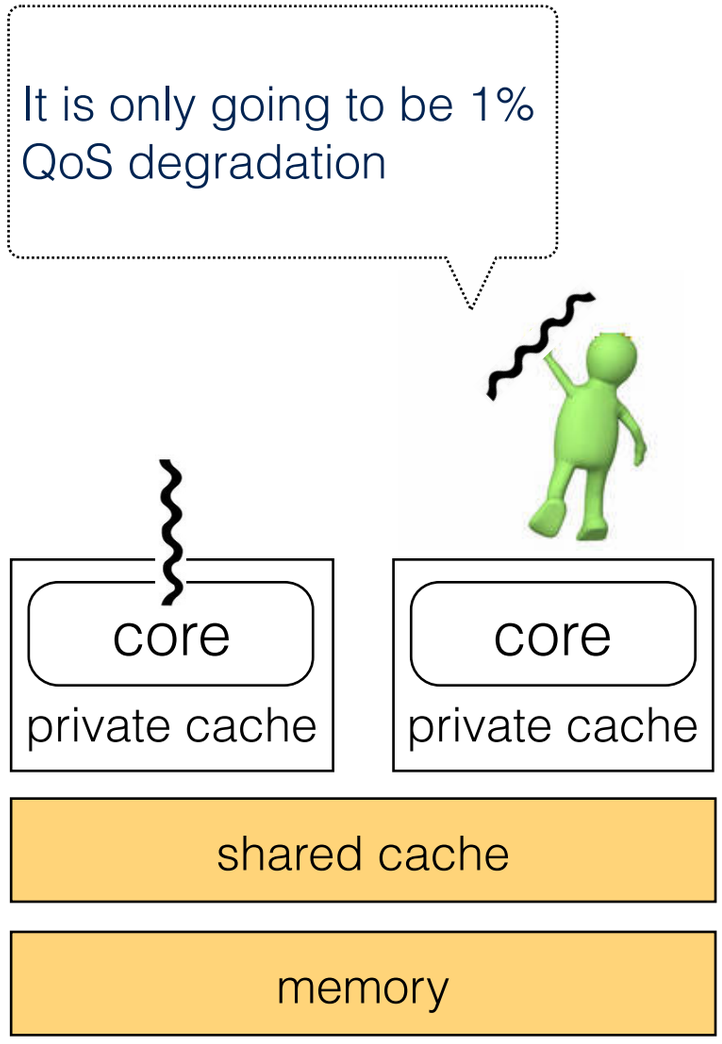
- Inefficiency due to low utilization
- Co-location can solve the problem



Server utilization distribution of a Google cluster.
(Borroso et al, “The datacenter as a computer: An Introduction to the Design of Warehouse-Scale Machines, Second edition”, Synthesis Lectures on Computer Architecture ’2013)

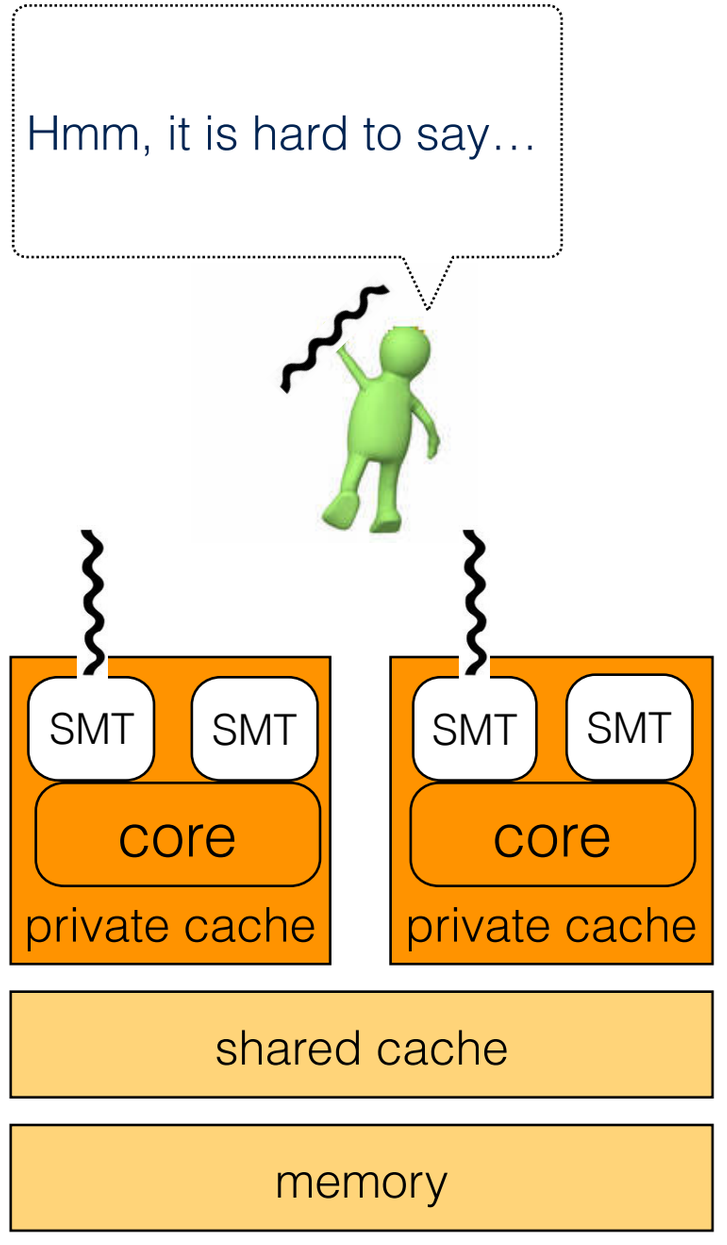
Keep calm and make predictions

- CMP co-location
  - Interference caused by contention on shared cache and memory bandwidth
- **Precise** QoS prediction for co-location [Bubble-Up 'MICRO2011, Bubble-flux 'ISCA2013, Whare-Map 'ISCA2013, Paragon 'ASPLOS2013, Quasar 'ASPLOS2014]
  - Identify "safe" co-locations
  - Improve server utilization

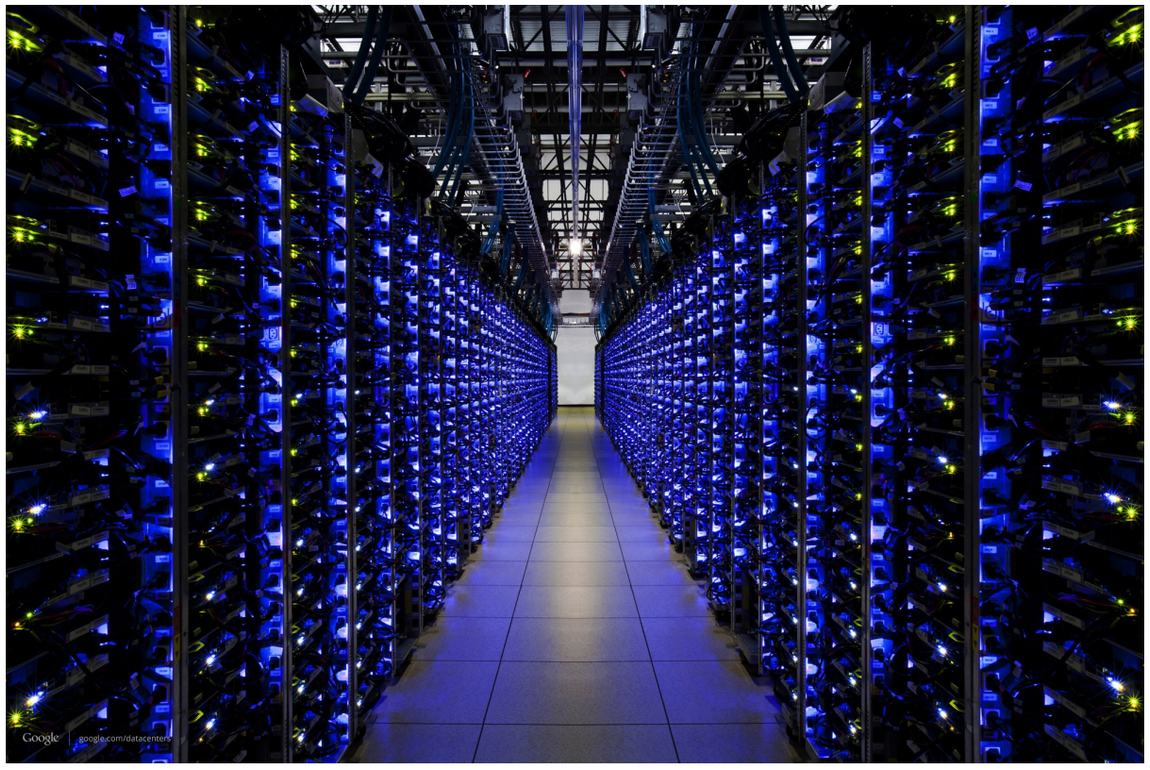


What about SMT

- **No** prior works on SMT co-locations
- Significantly more challenging than CMP co-location
  - Fine-grained resource sharing
  - Many more shared resources
- SMT is ubiquitous in modern WSCs



“For the Horde”

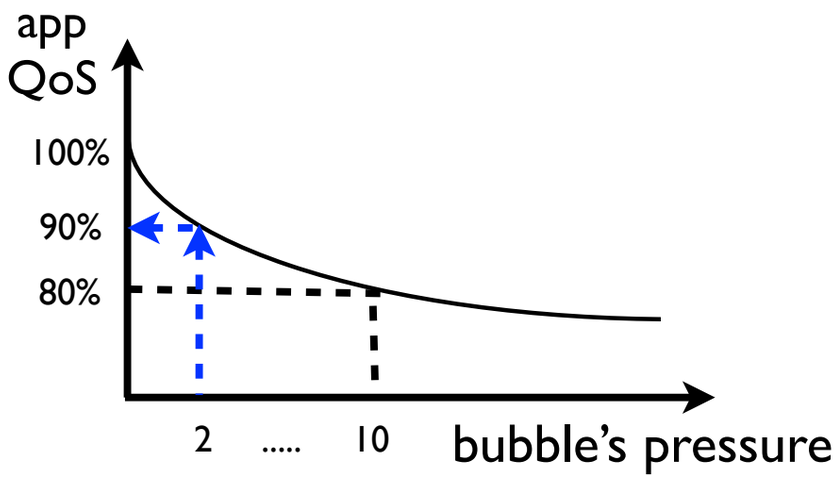


- Precise QoS interference prediction on **real-system** SMT processors
- Identify “safe” co-locations to improve server utilization

Is SMT co-location really different from CMP co-location?

Prior work for CMP co-location

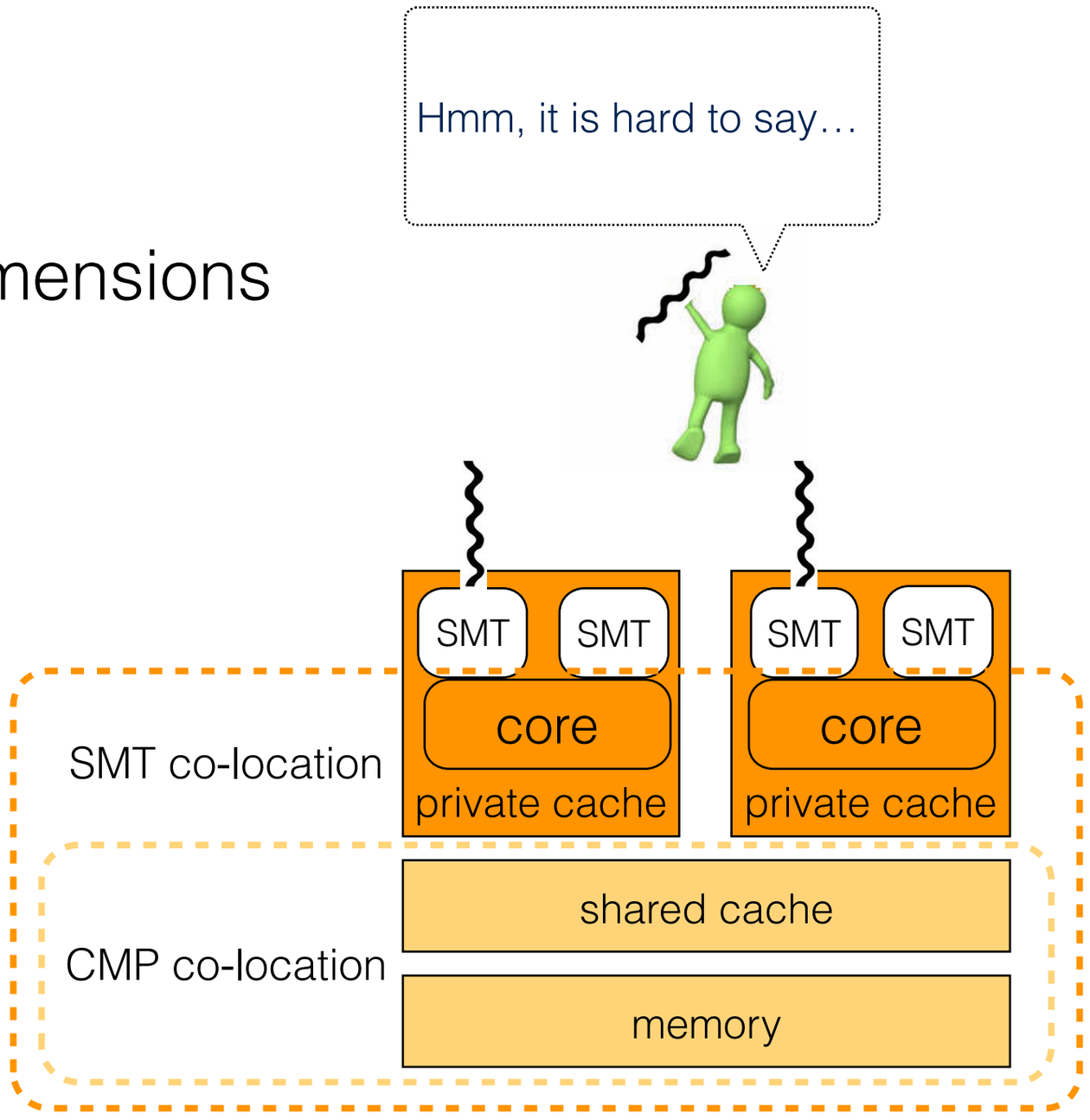
- One pressure score to quantify the contention
  - unified approach
  - limited # shared resources
- Can we still use the same approach for SMT co-location?



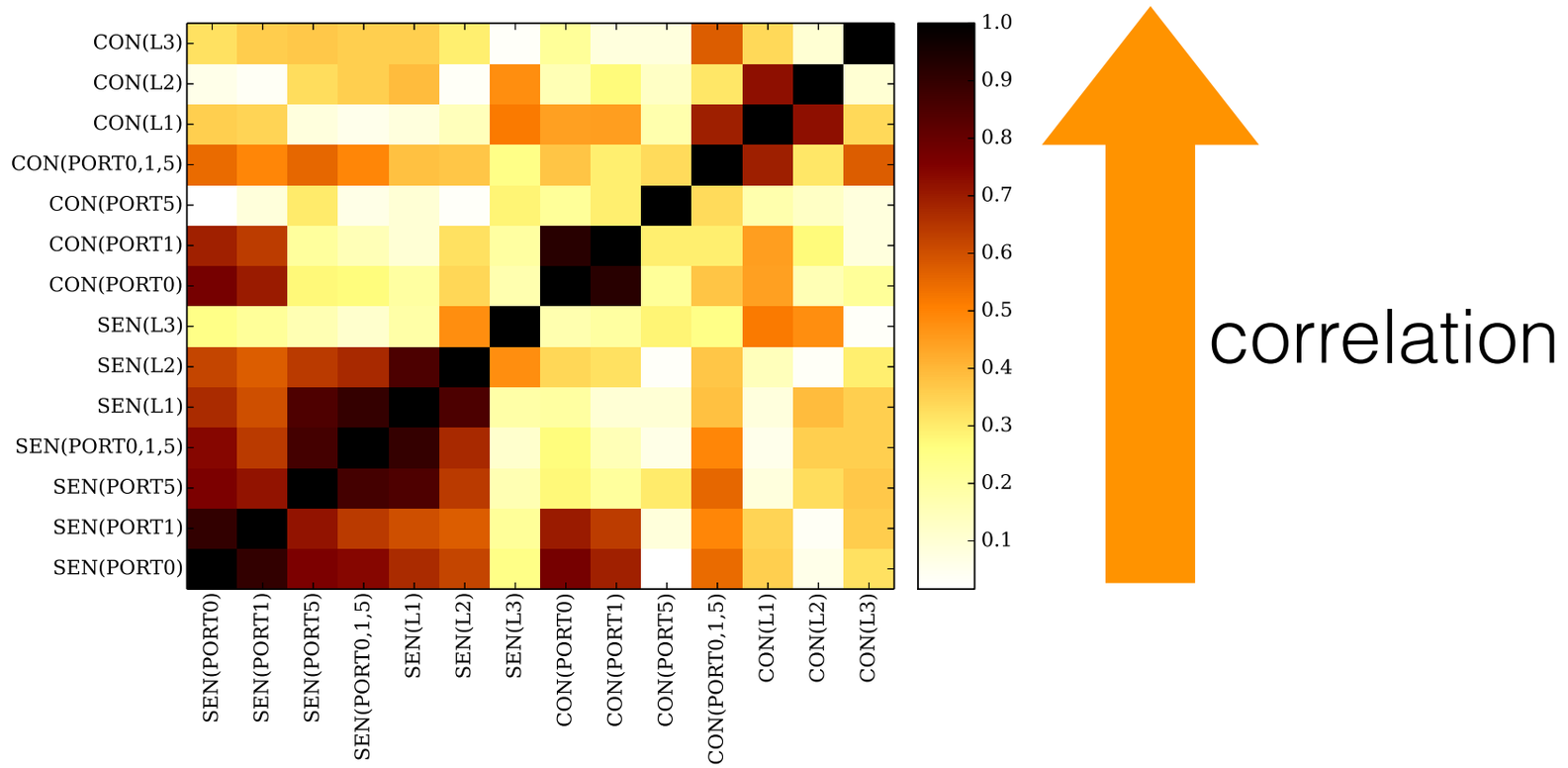
Unified metric to quantify the contention
[Bubble-Up 'MICRO2011]

Is it really different

- More resource sharing dimensions
  - private cache(s)
  - memory ports
  - functional units



What if they correlate



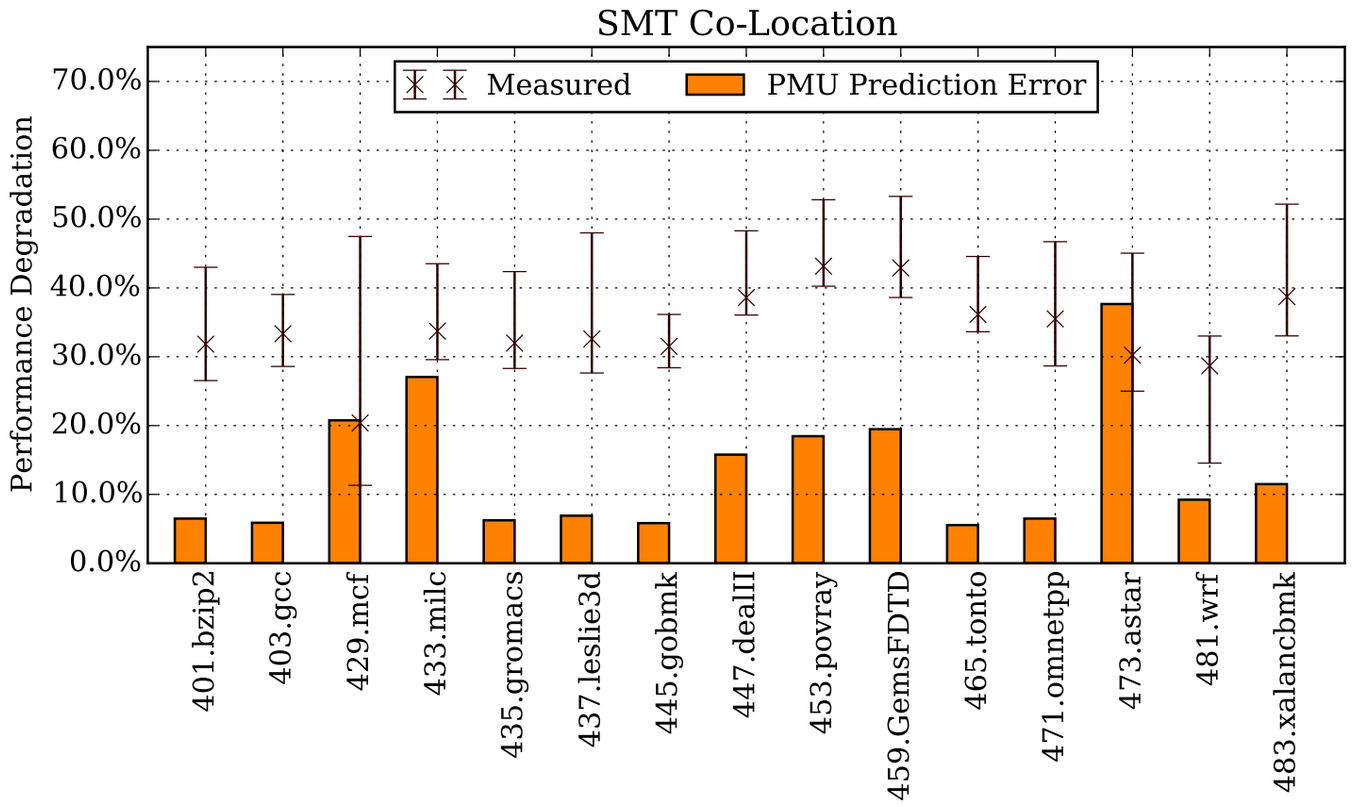
Absolute Pearson correlation coefficient. **97%** of the pairs < 0.8.

- No, different resources do not correlate
- A Unified approach cannot capture

A **decoupled** approach is required to quantify the contention for SMT co-location

Throw some PMUs
and a little regression to the problem

PMUs and regression models



- Regression model based on PMU measurements
- 14% prediction error on average

A **direct** measurement of contenting behavior is desirable for SMT co-location

Ruler-based Approach

A **decoupled** approach is required to quantify the contention for SMT co-location

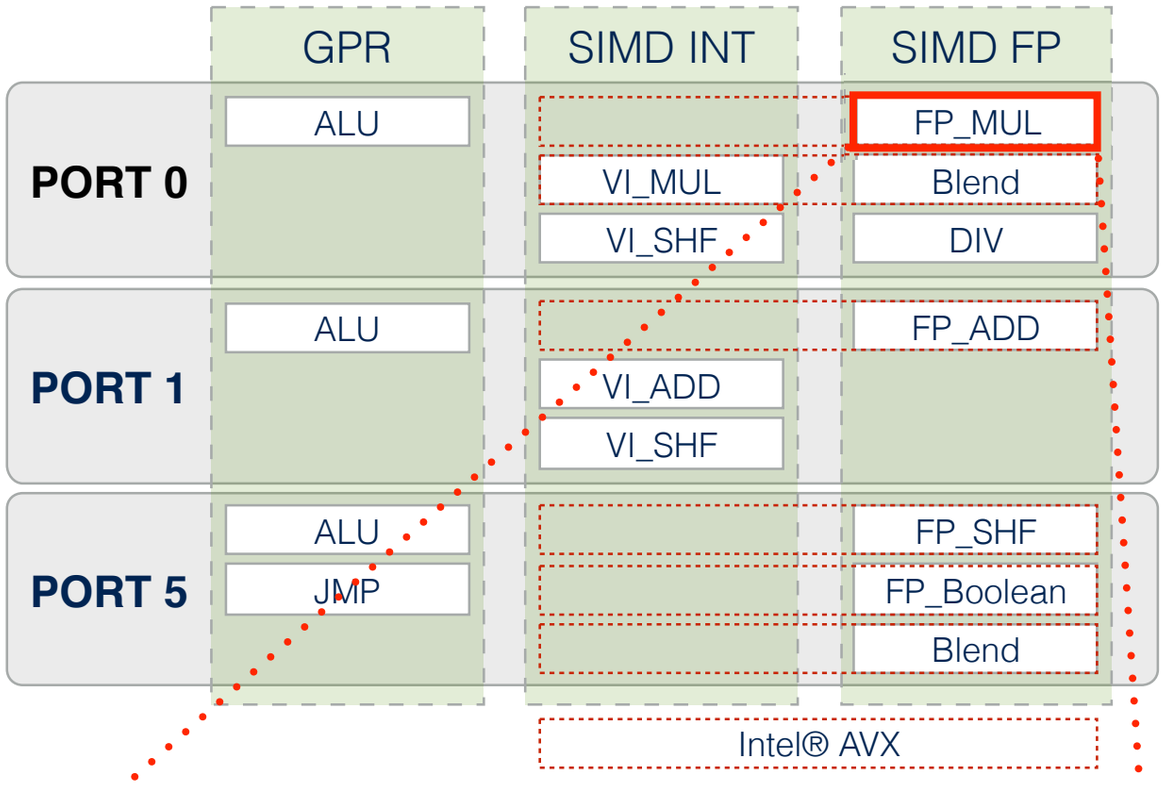
A **direct** measurement of contenting behavior is desirable for SMT co-location

Ruler-based approach

- Carefully designed set of micro-benchmarks
- Decouples contending behavior into each individual dimension in isolation
- Each one is extremely contentious in one specific resource sharing dimension

Ruler for functional units

- Port-specific instructions in commodity server designs

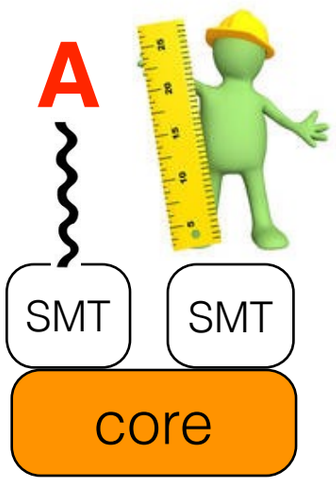


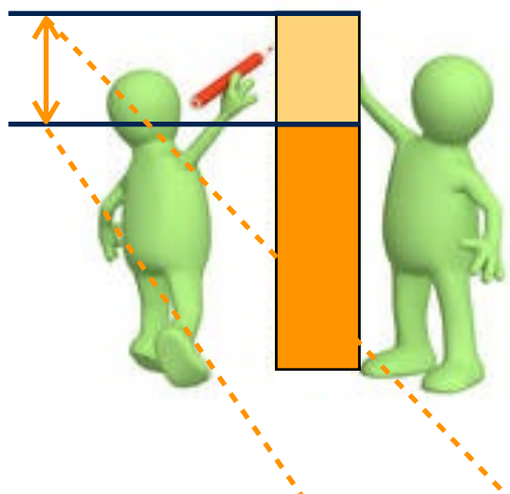
- Stream of independent instructions
- Achieve max utilization on specific port

```
loop:
{   mulps    %xmm0, %xmm0
    ……
    mulps    %xmm7, %xmm7
    ……
    jmp loop
```

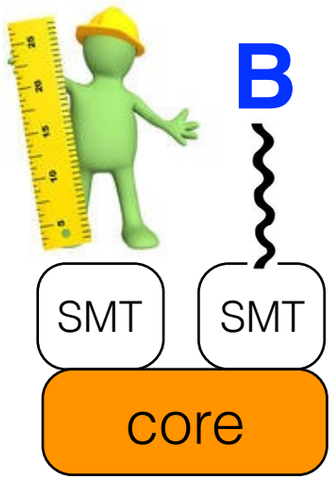
(a) FP_MUL (PORT0)

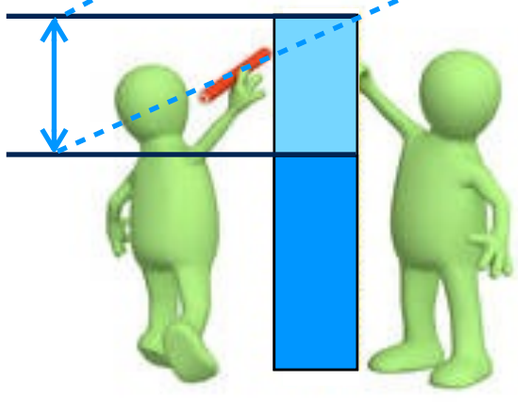
Use of Rulers

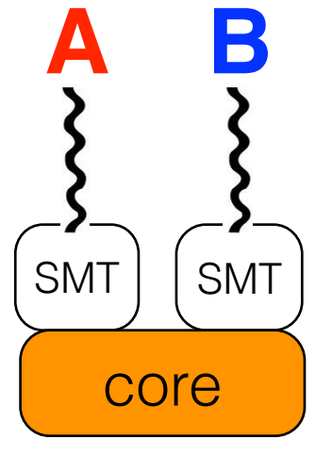




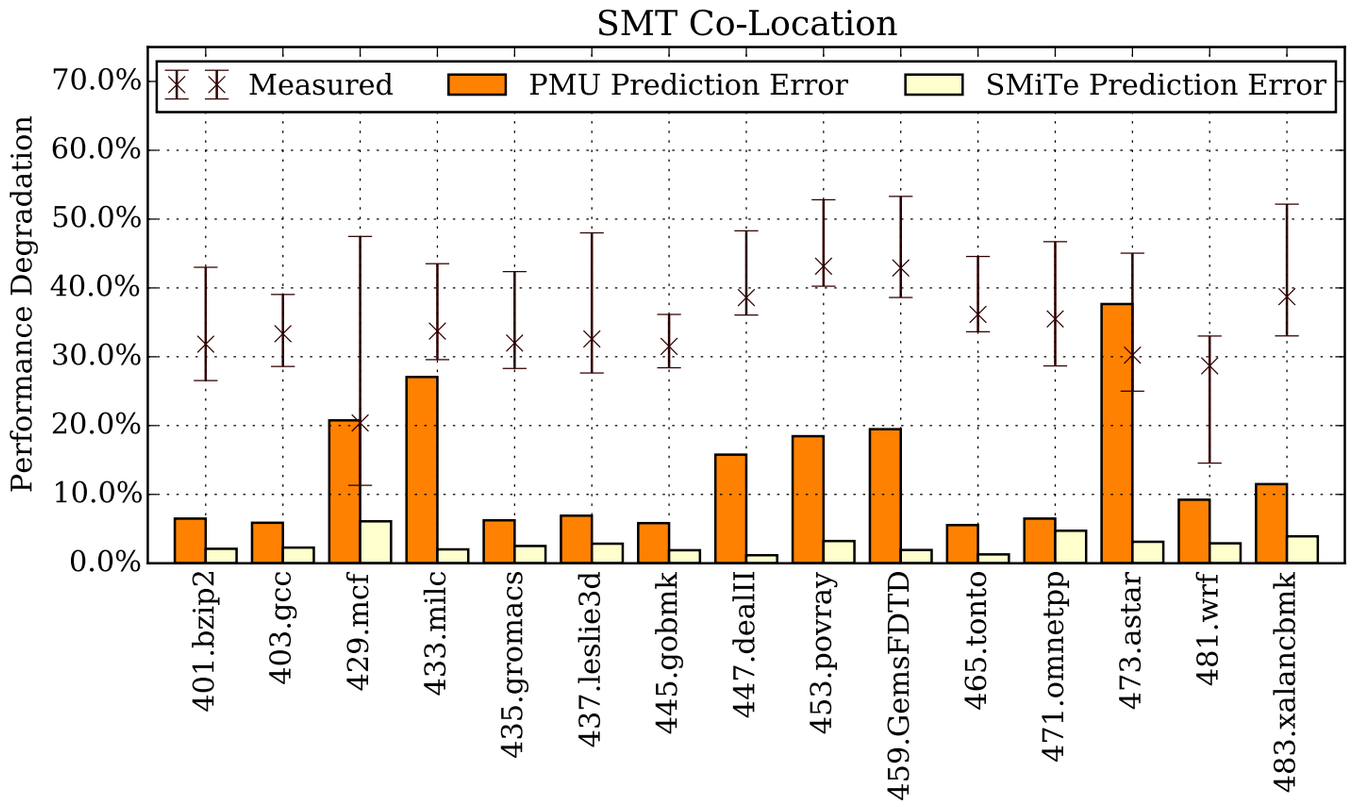
$$Deg^{A}_{co-locate/B} = \sum_{i}^{N}(c_i \times Sen_i^{A} \times Con_i^{B}) + c_0$$





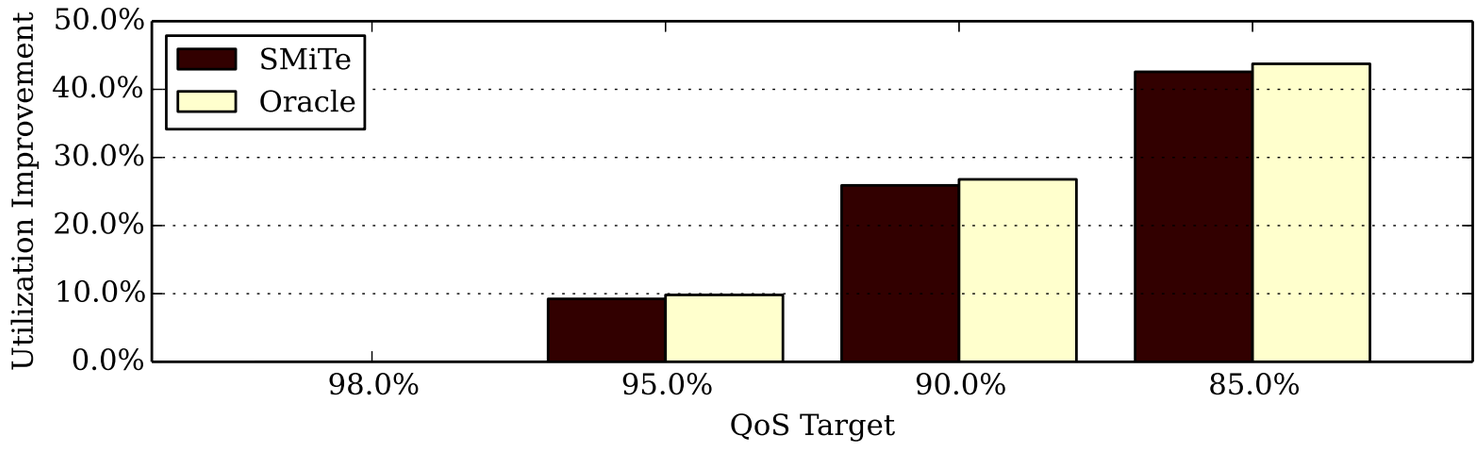


SMiTe prediction

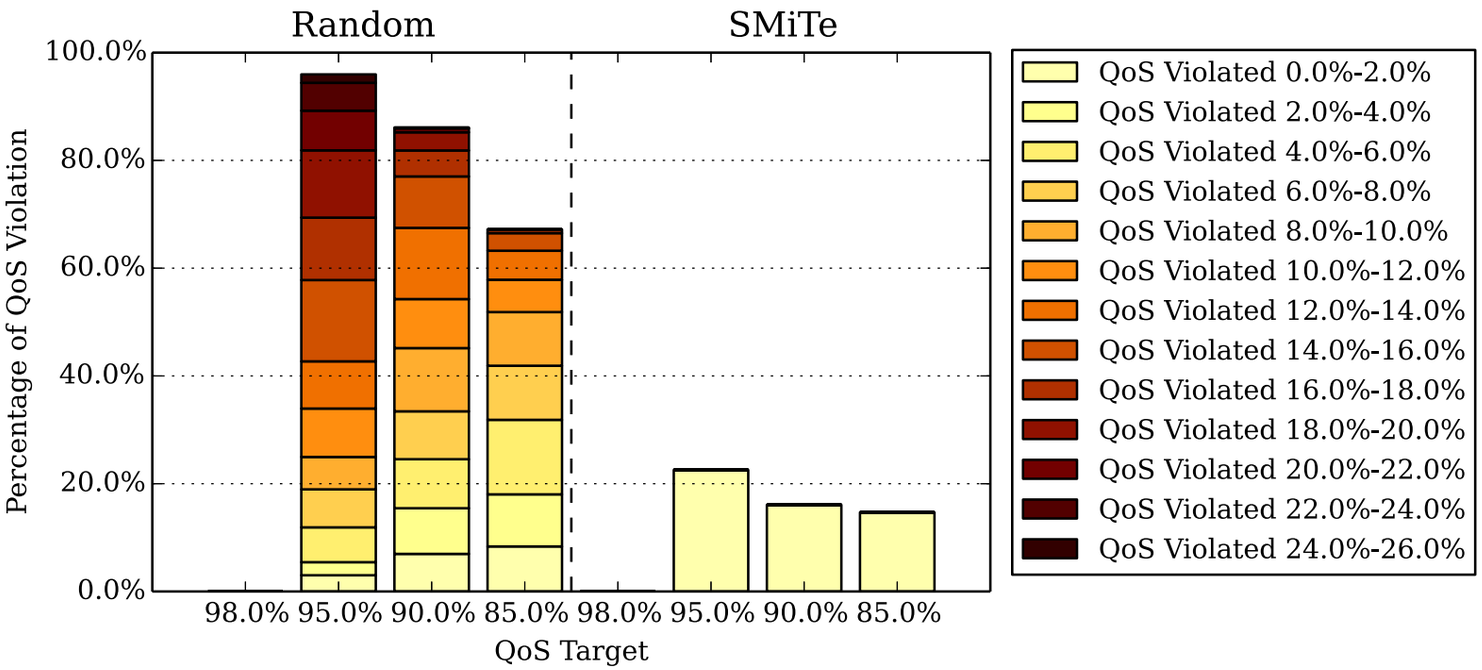


- Regression model based on Ruler characterization
- Evaluated on real-system SMT processors
- **2%** prediction error on average (14% PMU-based)

Putting in all together



- Close to Oracle
- **42%** Improvement



- **< 2%** Violation
- QoS Awareness

Conclusion

- A decoupled methodology to quantify contention is required for precise interference prediction
  - more shared resources in SMT co-location
  - contending behaviors in different dimensions do not correlate
- Ruler-based approach provides precision on real systems
  - 2% prediction error
- Improve warehouse scale computer utilization
  - 42% server utilization improvement

Questions