SLOMO: Contention-Aware Performance Prediction for Virtualized Network Functions

Antonis Manousis, Rahul Anand Sharma, Vyas Sekar, Justine Sherry
Carnegie Mellon University
Meeting SLAs in NFV: An operator’s nightmare

Server A or server B?

SLA: >1Gbps
Throughput Solo: 1.5Gbps
Meeting SLAs in NFV: An operator’s nightmare

A = 1.3Gbps

B = 0.7Gbps

VS.
I need a tool predicting the firewall’s throughput as it contends with other NFs for server resources.
The performance prediction problem

What will the target’s throughput be when co-located with any competitors?
SLOMO in a nutshell

SLOMO is a contention-aware performance prediction framework that:

1. Models the multiple facets of resource contention in the memory subsystem
2. Produces composable prediction models
3. Achieves <8% prediction error, up to 60% improvement from prior work
Agenda

1. Analysis of Contention in Memory Subsystem
2. Contentiousness & Sensitivity: A blueprint for performance prediction
3. Modeling Contentiousness & Sensitivity for NFV
4. Evaluation
Prior work: Using CAR for slowdown modeling

Throughput $= \mathcal{F}(\sum_i \text{CAR}_{\text{Comp}_i})$

[Dobrescu et al. NSDI ’12]
Reality: CAR is insufficient at predicting performance

Large prediction error with modern architectures and NFs
Predictions can only be as good as our understanding of contention!
The Life of packet in NFV: Understanding contention chokepoints

C1. Contention during packet IO
C2. Contention in Last Level Cache
C3. Contention for memory bandwidth

* DDIO: Data Direct IO
C1. Contention at Packet IO: Evictions from DDIO

Setup: Stateless FW vs. L2-Forwarding competitor.

As packet rates increase, DDIO contention evicts more target packets to DRAM.
C2. Contention in LLC: NF Data Evictions to DRAM

Setup: IP Router vs. synthetic competitor.

Competing LLC occupancy and access rate shape LLC contention.
C3. Contention in memory: Variable latency

Setup: IP Router vs. synthetic competitor & LLC isolation

As memory bandwidth utilization increases, memory latency is no longer fixed.
Contention manifests **simultaneously** and **independently** at at least three different chokepoints across the memory subsystem.
Agenda

1. Analysis of Contention in Memory Subsystem

2. Contentiousness & Sensitivity: A blueprint for performance prediction

3. Modeling Contentiousness & Sensitivity in NFV

4. Evaluation
Contentiousness & Sensitivity: Blueprint for Prediction

**Sensitivity:** How susceptible target NF is to contentiousness

**Contentiousness:** Usage of shared resources by competing NFs

[Mars et. al, MICRO’11]
Contentiousness: “Pressure” on shared resources

Contentiousness $\geq$ Contentiousness
Sensitivity: Target’s susceptibility to contention
Agenda

1. Analysis of Contention in Memory Subsystem

2. Contentiousness & Sensitivity: A blueprint for performance prediction

3. Modeling Contentiousness & Sensitivity in NFV

4. Evaluation
SLOMO: System Overview

Available NFs

Contentiousness Characterization
Sensitivity Modeling

Offline

NF Profiles

Contentiousness Composition
Prediction

Inputs to online predictor

Competitors

Target

Online

Throughput
SLOMO: Practical Challenges

1. How does SLOMO quantify contentiousness?

2. How does SLOMO model target’s sensitivity?
SLOMO: Practical Challenges

1. How does SLOMO quantify contentiousness?
2. How does SLOMO model target’s sensitivity?
3. How does SLOMO compose aggregate contentiousness of any set of competitors?
SLOMO: Practical Challenges

1. How does SLOMO quantify contentiousness?
2. How does SLOMO model target’s sensitivity?
3. How does SLOMO compose the aggregate contentiousness of any set of competitors?
1. How does SLOMO quantify contentiousness?

2. How does SLOMO model target’s sensitivity?

3. How does SLOMO compose the aggregate contentiousness of any set of competitors?
Contentiousness is a multivariate quantity

PCM counters
- LLC Miss Rate
- LLC Access Rate
- Memory Reads
- Memory Writes
- LLC Occupancy
- ...

What is a practical set of PCM metrics for SLOMO?
Different counters for different chokepoints

**Setup**: Measure correlation between counters and target performance

### Contention in DDIO

<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRITE</td>
<td>Memory Write traffic</td>
<td>0.90</td>
</tr>
<tr>
<td>READ</td>
<td>Memory Read traffic</td>
<td>0.88</td>
</tr>
<tr>
<td>IPC</td>
<td>Instructions/Cycle</td>
<td>0.88</td>
</tr>
<tr>
<td>L3OCC</td>
<td>LLC Occupancy</td>
<td>0.80</td>
</tr>
<tr>
<td>L2MPI</td>
<td>LLC Accesses/Instruction</td>
<td>0.76</td>
</tr>
<tr>
<td>L3HIT</td>
<td>LLC Hit Rate</td>
<td>0.40</td>
</tr>
<tr>
<td>LMB</td>
<td>Local NUMA bandwidth</td>
<td>0.20</td>
</tr>
</tbody>
</table>

### Contention in LLC

<table>
<thead>
<tr>
<th>Metric</th>
<th>Definition</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3MISS</td>
<td>LLC Miss Rate</td>
<td>0.98</td>
</tr>
<tr>
<td>L3OCC</td>
<td>LLC Occupancy</td>
<td>0.87</td>
</tr>
<tr>
<td>L3HIT</td>
<td>LLC Hit Rate</td>
<td>0.79</td>
</tr>
<tr>
<td>LMB</td>
<td>Local NUMA bandwidth</td>
<td>0.76</td>
</tr>
<tr>
<td>L2MISS</td>
<td>CAR</td>
<td>0.76</td>
</tr>
<tr>
<td>L2HIT</td>
<td>L2 Hit Rate</td>
<td>0.36</td>
</tr>
<tr>
<td>RMB</td>
<td>Remote NUMA bandwidth</td>
<td>0.13</td>
</tr>
</tbody>
</table>

≠
SLOMO: Practical Challenges

1. How does SLOMO quantify contentiousness?

2. How does SLOMO model target’s sensitivity?

3. How does SLOMO compose the aggregate contentiousness of any set of competitors?
Sensitivity: A complex function of contentiousness

**Observation:** Simple models cannot accurately capture sensitivity!
Insight: Sensitivity exhibits phase transitions

How does SLOMO model this complex behavior?
Approach: Model sensitivity as a piecewise function of input

Gradient Boosting Regression to model sensitivity as an ensemble of localized predictors
SLOMO: Practical Challenges

1. How does SLOMO quantify contentiousness?

2. How does SLOMO model target’s sensitivity?

3. How does SLOMO compose the aggregate contentiousness of any set of competitors?

Detailed analysis in our paper 😊
Agenda

1. Analysis of Contention in Memory Subsystem

2. Contentiousness & Sensitivity: A blueprint for performance prediction

3. Porting Contentiousness & Sensitivity to NFV

4. Evaluation
Experimental Setup: Hardware

Configuration 1:
CPU: Intel Xeon E5-2620 v4 (20MB inclusive LLC)
NIC: Intel XL710 - 40GbE (SR-IOV)

Configuration 2:
CPU: Intel Xeon Silver 4110 (11MB non inclusive LLC)
NIC: Mellanox MT27700 Family - 100GbE (SR-IOV)
Experimental Setup: NFs

We evaluated both research and industrial grade NFs

<table>
<thead>
<tr>
<th>Application</th>
<th>Description</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stateless Firewall</td>
<td>Click-based - Stateless</td>
<td>1K sequential rules</td>
</tr>
<tr>
<td>IP Router</td>
<td>Click-based - RadixIPLookup</td>
<td>130k rules</td>
</tr>
<tr>
<td>FlowStats</td>
<td>Click-based - AggregateIPFlows</td>
<td>1h Flow Timeout</td>
</tr>
<tr>
<td>VPN</td>
<td>Click-based – IPSec enabled</td>
<td>Encryption &amp; Authentication</td>
</tr>
<tr>
<td>Maglev LB</td>
<td>NetBricks Load Balancer</td>
<td>Default</td>
</tr>
<tr>
<td>Snort</td>
<td>IPS – Intrusion Detection Mode</td>
<td>Snort3.0 – Community Rules</td>
</tr>
<tr>
<td>Suricata</td>
<td>IPS – Intrusion Detection Mode</td>
<td>Suricata-5.0.1 Ruleset</td>
</tr>
<tr>
<td>PFSense</td>
<td>Open Source Stateful Firewall</td>
<td>1K rules</td>
</tr>
</tbody>
</table>
Experimental Setup: Comparisons

For prediction accuracy, we evaluate SLOMO against:

1. CAR-based performance predictor for NFV
   [Dobrescu et al. NDSI’ 12]

2. BubbleUP: Performance prediction for General Purpose applications
   [Mars et al. Micro’ 11]
SLOMO’s predictions are accurate

SLOMO achieves <8% avg error and improves prior work by up to 60%!
SLOMO improves server utilization of NFV cluster

<table>
<thead>
<tr>
<th>System</th>
<th>Resource Overhead (%)</th>
<th>SLA violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLOMO</td>
<td>1.5%</td>
<td>0</td>
</tr>
<tr>
<td>ResQ [NSDI ’18]</td>
<td>6%</td>
<td>3%</td>
</tr>
<tr>
<td>CAR-based [NSDI ’12]</td>
<td>14%</td>
<td>0</td>
</tr>
</tbody>
</table>
Takeaways

• Contention-aware performance prediction is critical for NFV orchestration.

• Prior predictors are inaccurate: Insufficient in modeling contention

• Our observation: Contention which happens at multiple chokepoints in memory.

• SLOMO Design Contributions
  • Microarchitectural blueprint for modeling slowdown through Contentiousness & Sensitivity
  • Practical ML models for Contentiousness, Sensitivity, Composition

• SLOMO is 60% more accurate; ensures up to 15% better resource usage.

SLOMO open source at github.com/cmu-snap-lab/SLOMO
Antonis Manousis ~ antonis@cmu.edu