Modeling Tail-Latencies

Max Helm, Florian Wiedner, Alexander Daichendt, Jonas Andre, Georg Carle

December 1, 2023

Chair of Network Architectures and Services
Department of Computer Engineering
Technical University of Munich
I. Extreme Value Theory Latency Models of Containers

II. Network Calculus as Latency Quantile Predictor Assistant
I. Extreme Value Theory Delay Models of Containers

What? Why?

- Containerized applications are important for sharing hardware resources and providing resources on-demand
- Applications with user interaction are latency-sensitive
- High impact of tail-latencies
- No available forwarding delay benchmark of containers
I. Extreme Value Theory Delay Models of Containers

What? Why?

- Containerized applications are important for sharing hardware resources and providing resources on-demand
- Applications with user interaction are latency-sensitive
- High impact of tail-latencies
- No available forwarding delay benchmark of containers

⇒ Can we predict tail-latency behavior of containers?
I. Extreme Value Theory Delay Models of Containers

Measurements
I. Extreme Value Theory Delay Models of Containers

Measurements

![Diagram showing LoadGen, DuT, and Timestamper with a graph representing Latency [μs] vs. Percentile [%]]
I. Extreme Value Theory Delay Models of Containers

Tail Latencies and Rare Events
I. Extreme Value Theory Delay Models of Containers

Tail Latencies and Rare Events

![Diagram showing density distribution of observable events, maximal observed delay, actual worst case, and upper bound for end-to-end delay.](image-url)
I. Extreme Value Theory Delay Models of Containers

Tail Latencies and Rare Events

Density

Observable events

Maximal observed delay

Actual worst case

Upper bound

End-to-end delay

Rare events

Over provisioning

Measurements, Simulation, Emulation
I. Extreme Value Theory Delay Models of Containers
Tail Latencies and Rare Events

Measurements, Simulation, Emulation

Queuing Theory

Maximal observed delay
Actual worst case
Upper bound

End-to-end delay

Observable events
Rare events
Over provisioning

Density
I. Extreme Value Theory Delay Models of Containers

Tail Latencies and Rare Events

- Queuing Theory
- Maximal observed delay
- Actual worst case
- Upper bound

End-to-end delay

Measurements, Simulation, Emulation

Formal methods, e.g., Network Calculus
I. Extreme Value Theory Delay Models of Containers
Tail Latencies and Rare Events

- Queuing Theory
- Maximal observed delay
- Actual worst case
- Upper bound
- Observables events
- Rare events
- Over provisioning
- End-to-end delay

Measurements, Simulation, Emulation

Formal methods, e.g., Network Calculus
I. Extreme Value Theory Delay Models of Containers

Modeling Approach: Extreme Value Theory

Extreme Value Theory (EVT):

- Predict future extreme events based on historical data
- Previously used for natural disaster prediction
- High latencies are a type of extreme event in networks

Modeling Approach:

- Select a threshold (what are tail latencies?)
- Fit a Generalized Pareto Distribution (GPD) to values above threshold using, e.g., a Maximum Likelihood Estimator (MLE)
- Obtained model can be used to extrapolate to future events, assess “expected worst-case behavior”
I. Extreme Value Theory Delay Models of Containers

EVT Model

Performance metrics:
- Return level: Expected worst-case latency
- Return period: Within this timeframe
- E.g., within 10 minutes the expected worst-case latency is 30µs, within 20 minutes it is 35µs

Model convergence:
- Expected worst-case latency converges or diverges based on sign of tail parameter
- Return period $\rightarrow \infty$
I. Extreme Value Theory Delay Models of Containers

EVT Model

Performance metrics:
- Return level: Expected worst-case latency
- Return period: Within this timeframe
- E.g., within 10 minutes the expected worst-case latency is 30µs, within 20 minutes it is 35µs

Model convergence:
- Expected worst-case latency converges or diverges based on sign of tail parameter
- Return period $\rightarrow \infty$
I. Extreme Value Theory Delay Models of Containers

EVT Model

Performance metrics:
- Return level: Expected worst-case latency
- Return period: Within this timeframe
- E.g., within 10 minutes the expected worst-case latency is 30µs, within 20 minutes it is 35µs

Model convergence:
- Expected worst-case latency converges or diverges based on sign of tail parameter
- Return period $\rightarrow \infty$
I. Extreme Value Theory Delay Models of Containers

Return level is the expected worst case latency for a given timespan.
I. Extreme Value Theory Delay Models of Containers

Return level is the expected worst case latency for a given timespan.
Return level is the latency that is expected to be exceeded exactly once during a given timespan.

Experiment:
• Divide container latency measurements into 20% training, 80% evaluation.
• Fit an EVT model to the 20%.
• Make predictions for the remaining 80%.

Platform exceedances of return level:
- Optimal Model: 1.00
- Container: 1.50
- Virtual Machine: 2.58

⇒ The predicted worst-case latency is exceeded 1.5 times instead of the expected one time on average.

This type of verification of an EVT model is typically not done in literature due to scarcity of evaluation data.
I. Extreme Value Theory Delay Models of Containers

Return level is the expected worst-case latency for a given timespan.
Return level is the latency that is expected to be exceeded exactly once during a given timespan.

Experiment:
- Divide container latency measurements into 20% training, 80% evaluation.
- Fit an EVT model to the 20%.
- Make predictions for the remaining 80%.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Exceedances of return level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Model</td>
<td>1.00</td>
</tr>
<tr>
<td>Container</td>
<td>1.50</td>
</tr>
</tbody>
</table>
I. Extreme Value Theory Delay Models of Containers

Return level is the expected worst-case latency for a given timespan

Return level is the latency that is expected to be exceeded exactly once during a given timespan

Experiment:

- Divide container latency measurements into 20% training, 80% evaluation
- Fit an EVT model to the 20%
- Make predictions for the remaining 80%

<table>
<thead>
<tr>
<th>Platform</th>
<th>Exceedances of return level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Model</td>
<td>1.00</td>
</tr>
<tr>
<td>Container</td>
<td>1.50</td>
</tr>
<tr>
<td>Virtual Machine</td>
<td>2.58</td>
</tr>
</tbody>
</table>

⇒ The predicted worst-case latency is exceeded 1.5 times instead of the expected one time on average
Return level is the expected worst-case latency for a given timespan.
Return level is the latency that is expected to be exceeded exactly once during a given timespan.

Experiment:

- Divide container latency measurements into 20% training, 80% evaluation.
- Fit an EVT model to the 20%.
- Make predictions for the remaining 80%.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Exceedances of return level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Model</td>
<td>1.00</td>
</tr>
<tr>
<td>Container</td>
<td>1.50</td>
</tr>
<tr>
<td>Virtual Machine</td>
<td>2.58</td>
</tr>
</tbody>
</table>

⇒ The predicted worst-case latency is exceeded 1.5 times instead of the expected one time on average.

(this type of verification of an EVT model is typically not done in literature due to scarcity of evaluation data.)
II. Network Calculus as Latency Quantile Predictor Assistant

What? Why?
II. Network Calculus as Latency Quantile Predictor Assistant

Network Calculus Basics

\[ \alpha(t) = r \cdot t + b \]

- **Service curve**
- **Arrival curve**
- **Latency Bound**
- **Backlog Bound**

Data

flow rate

flow burst

server proc. delay

server rate

Time
II. Network Calculus as Latency Quantile Predictor Assistant

Network Topologies

(a) Network I

(b) Network II

(c) Network III
II. Network Calculus as Latency Quantile Predictor Assistant

Latency Quantile Point Predictions

<table>
<thead>
<tr>
<th>Predicted Metric</th>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>$\eta_{90}$</th>
<th>$\eta_{99}$</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Error</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

NC Bounds
- False
- True
II. Network Calculus as Latency Quantile Predictor Assistant

Importance of Network Calculus Results

Analysis methods:

- Total Flow Analysis (TFA): Bounds on flow aggregates on per-hop basis
- Separate Flow Analysis (SFA): Bounds per flow using left-over service curves and service curve convolutions
- SFA bounds tighter or equally tight as TFA bounds
- (other analytical and linear programming-based approaches exist)
II. Network Calculus as Latency Quantile Predictor Assistant

Importance of Network Calculus Results

Analysis methods:

- Total Flow Analysis (**TFA**): Bounds on flow aggregates on per-hop basis
- Separate Flow Analysis (**SFA**): Bounds per flow using left-over service curves and service curve convolutions
- SFA bounds tighter or equally tight as TFA bounds
- (other analytical and linear programming-based approaches exist)
Conclusion

I. Extreme Value Theory Latency Models of Containers

- EVT suitable to model tail latencies


II. Network Calculus as Latency Quantile Predictor Assistant

- Network Calculus bounds helpful for other modeling approaches