

Exact Analysis of TTL Cache Networks

Daniel S. Berger, Philipp Gland,
Sahil Singla, and Florin Ciucu

IFIP WG 7.3 Performance
October 7, 2014

Classes of Caching

Classes of Caches

capacity-driven eviction

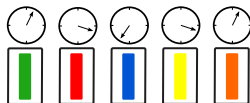


Least-Recently-Used (LRU)

First-In-First-Out (FIFO)

Random Eviction (RND)

timer-based eviction



Time-To-Live models

$\text{TTL-}\mathcal{R}$

$\text{TTL-}\Sigma$

$\text{TTL-min}(\Sigma, \mathcal{R})$

Notation for a TTL Caching Node



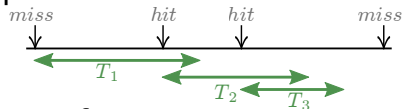
inter-arrival – TTL – cache model

cache model $\in \{ \mathcal{R}, \Sigma, \min(\Sigma, \mathcal{R}) \}$

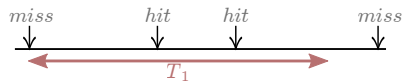
examples: $G-G-\mathcal{R}$ or $M-D-\Sigma$

Some TTL Policies

Policy $\mathcal{R}$: Reset **TTL** with every request.
Evict object upon TTL expiry.



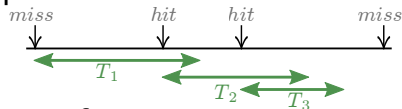
Policy Σ : Reset **TTL** only at the times of
unsuccessful requests (misses).
Evict object upon TTL expiry.



Some TTL Policies

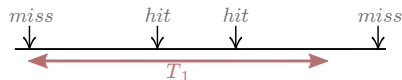
Policy $\mathcal{R}$: Reset **TTL** with every request.

Evict object upon TTL expiry.



Policy Σ : Reset **TTL** only at the times of unsuccessful requests (misses).

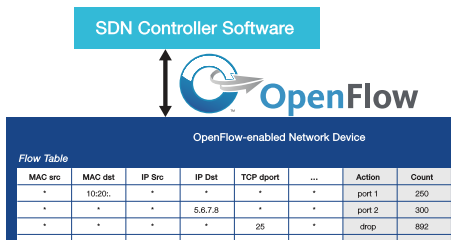
Evict object upon TTL expiry.



Policy $\min(\Sigma, \mathcal{R})$: Two parallel TTLs, with $\mathcal{R}$ and Σ reset policies, respectively. Evict object upon expiry of *either* TTL.

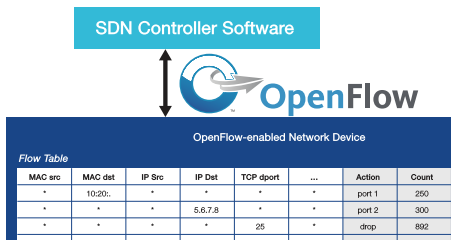
Some TTL Policies

Policy $\min(\Sigma, \mathcal{R})$: Two parallel TTLs, with $\mathcal{R}$ and Σ reset policies, respectively. Evict object upon expiry of *either* TTL.



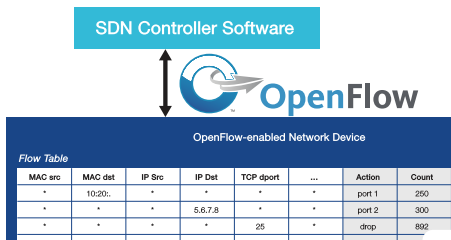
Some TTL Policies

Policy $\min(\Sigma, \mathcal{R})$: Two parallel TTLs,
with $\mathcal{R}$ and Σ reset policies, respectively.
Evict object upon expiry of *either* TTL.



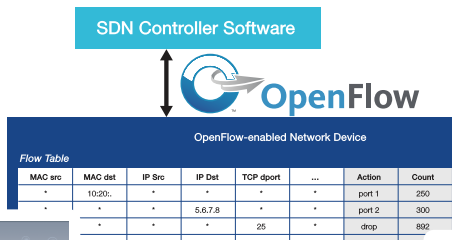
Some TTL Policies

Policy $\min(\Sigma, \mathcal{R})$: Two parallel TTLs, with $\mathcal{R}$ and Σ reset policies, respectively. Evict object upon expiry of *either* TTL.



Some TTL Policies

Policy $\min(\Sigma, \mathcal{R})$: Two parallel TTLs,
with $\mathcal{R}$ and Σ reset policies, respectively.
Evict object upon expiry of *either* TTL.



LRU, FIFO, RND $\in$ TTL

plots omitted

copyright [Martina et al., A unified approach to the performance analysis of caching systems, IEEE INFOCOM 2014]

$$\text{LRU} \approx M-D-\mathcal{R} \quad \text{FIFO} \approx M-D-\Sigma \quad \text{RND} \approx M-M-\Sigma$$

plots © [Martina et al, IEEE Infocom 2014]

Contribution of this work

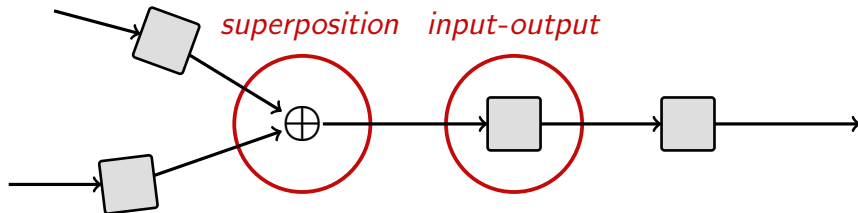
I) unifies the analysis of $\mathcal{R}$ and Σ



[Choungmo Fofack et al., VALUETOOLS 2012 and 2013]

II) first analysis of $\min(\Sigma, \mathcal{R})$

III) first exact analysis of feedforward networks

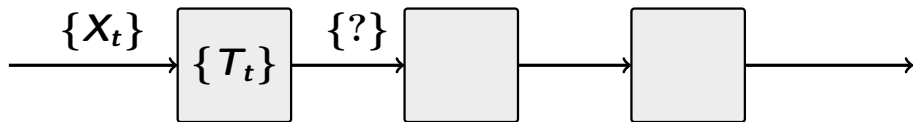


Outline

- 1 Introduction and Contribution
- 2 Analysis
 - Line Networks (Renewal Request Model)
 - Feedforward Networks (MAP Request Model)
- 3 Accuracy of the Poisson Approximation
- 4 Model Comparison and Conclusions

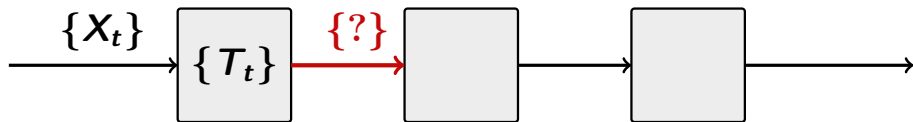
Line Networks

$\{X_t\}_{t \geq 1}$ and $\{T_t\}_{t \geq 1}$: independent renewal processes



Line Networks

$\{X_t\}_{t \geq 1}$ and $\{T_t\}_{t \geq 1}$: independent renewal processes

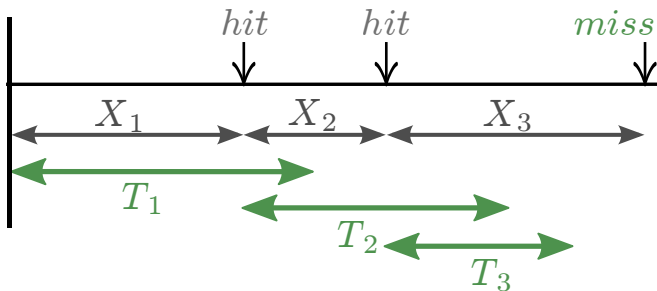


Stopping Time Formalization

$\{X_t\}_{t \geq 1}$ and $\{T_t\}_{t \geq 1}$: independent renewal processes

the miss process: $S_\tau := X_1 + \dots + X_\tau$

$$\mathcal{R} : \tau := \min\{t : X_t > T_t\}$$



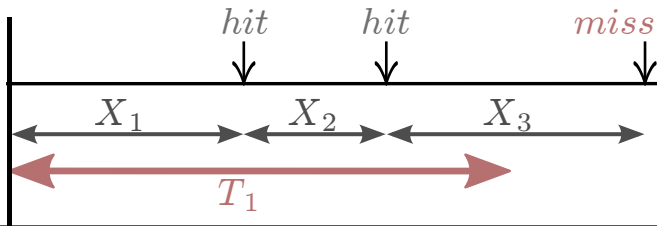
Stopping Time Formalization

$\{X_t\}_{t \geq 1}$ and $\{T_t\}_{t \geq 1}$: independent renewal processes

the miss process: $S_T := X_1 + \dots + X_T$

$$\mathcal{R} : \tau := \min\{t : X_t > T_t\}$$

$$\Sigma : \tau := \min\{t : \sum_{s=1}^t X_s > T_1\}$$



Stopping Time Formalization

$\{X_t\}_{t \geq 1}$ and $\{T_t\}_{t \geq 1}$: independent renewal processes

the miss process: $S_{\mathcal{T}} := X_1 + \dots + X_{\mathcal{T}}$

$$\mathcal{R} : \tau := \min\{t : X_t > T_t\}$$

$$\Sigma : \tau := \min\left\{t : \sum_{s=1}^t X_s > T_1\right\}$$

$$\min(\Sigma, \mathcal{R}) : \tau := \min \left\{ \begin{array}{l} \min\left\{t : \sum_{s=1}^t X_s > T_1^{\Sigma}\right\}, \\ \min\left\{t : X_t > T_t^{\mathcal{R}}\right\} \end{array} \right\}$$

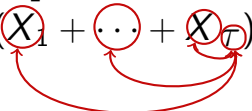
Laplace Transform of Miss Process

Find

$$\begin{aligned}(\mathcal{L}S_\tau)(\omega) &= \mathbb{E} \left[e^{-\omega S_\tau} \right] \\ &= \mathbb{E} \left[e^{-\omega(X_1 + \cdots + X_\tau)} \right]\end{aligned}$$

Laplace Transform of Miss Process

Find

$$\begin{aligned}(\mathcal{L}S_\tau)(\omega) &= \mathbb{E} \left[e^{-\omega S_\tau} \right] \\ &= \mathbb{E} \left[e^{-\omega (X_1 + \dots + X_\tau)} \right]\end{aligned}$$


Challenge

dependencies $\tau \leftrightarrow X_i$

(Fractional) Change of Measure $\mathbb{P} \rightarrow \tilde{\mathbb{P}}$

$$\forall t \quad \tilde{\mathbb{P}}_t(F) := E \left[\frac{e^{-\omega S_t}}{\mathcal{L}(X)^t} 1_F \right]$$
$$F \in \mathcal{F}_t = \sigma((X_1, T_1), \dots, (X_t, T_t))$$

Theorem

For an a.s. finite stopping time τ :

$$(\mathcal{L}S_\tau) = \tilde{\mathbb{E}} \left[(\mathcal{L}X)^\tau \right]$$

(Fractional) Change of Measure $\mathbb{P} \rightarrow \tilde{\mathbb{P}}$

$$\forall t \quad \tilde{\mathbb{P}}_t(F) := E \left[\frac{e^{-\omega S_t}}{\mathcal{L}(X)^t} 1_F \right]$$
$$F \in \mathcal{F}_t = \sigma((X_1, T_1), \dots, (X_t, T_t))$$

Theorem

For an a.s. finite stopping time τ :

$$(\mathcal{L}S_\tau) = \tilde{\mathbb{E}} \left[(\mathcal{L}X)^\tau \right]$$

constant

(Fractional) Change of Measure $\mathbb{P} \rightarrow \tilde{\mathbb{P}}$

$$\forall t \quad \tilde{\mathbb{P}}_t(F) := E \left[\frac{e^{-\omega S_t}}{\mathcal{L}(X)^t} 1_F \right]$$
$$F \in \mathcal{F}_t = \sigma((X_1, T_1), \dots, (X_t, T_t))$$

Theorem

For an a.s. finite stopping time τ :

$$(\mathcal{L}S_\tau) = \tilde{\mathbb{E}} \left[(\mathcal{L}X)^\tau \right]$$

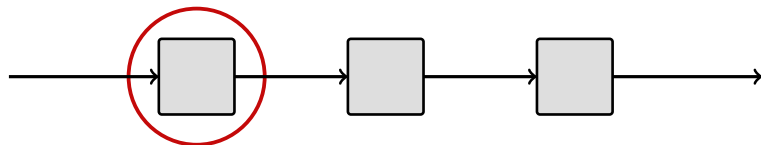
$\tilde{\mathbb{P}}(\tau = k)$

constant

From Renewals to MAPs

input-output
renewals ✓

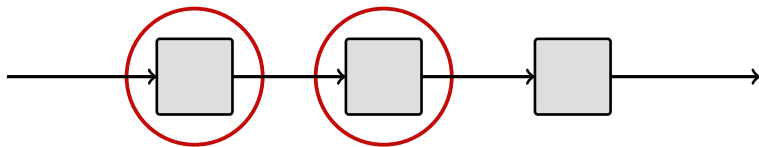
input-output



From Renewals to MAPs

input-output
renewals ✓

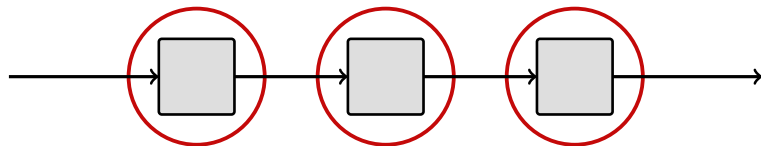
input-output



From Renewals to MAPs

input-output
renewals ✓

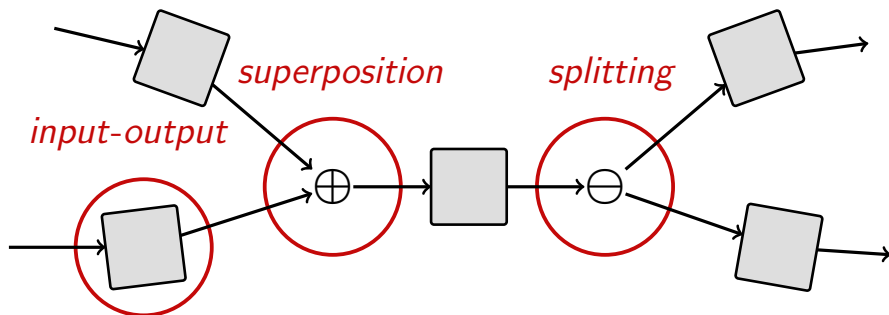
input-output



From Renewals to MAPs

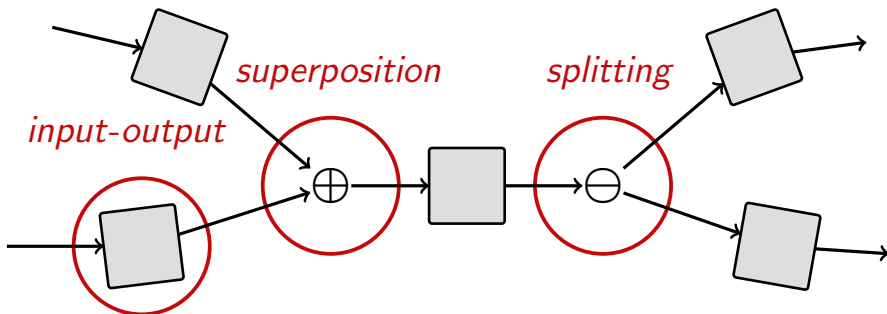
renewals input-output superposition splitting

✓



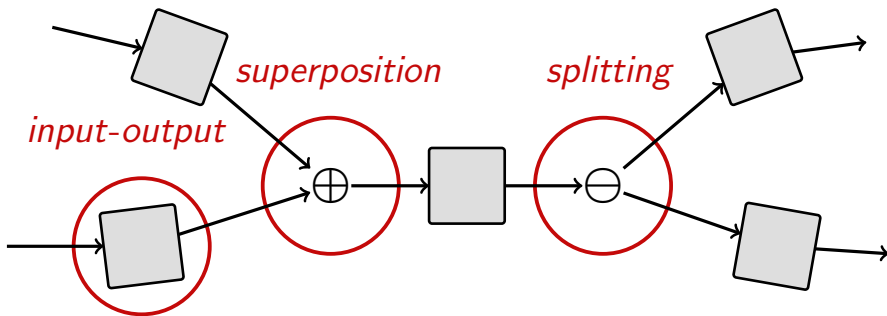
From Renewals to MAPs

	input-output	superposition	splitting
renewals	✓	✗	✗



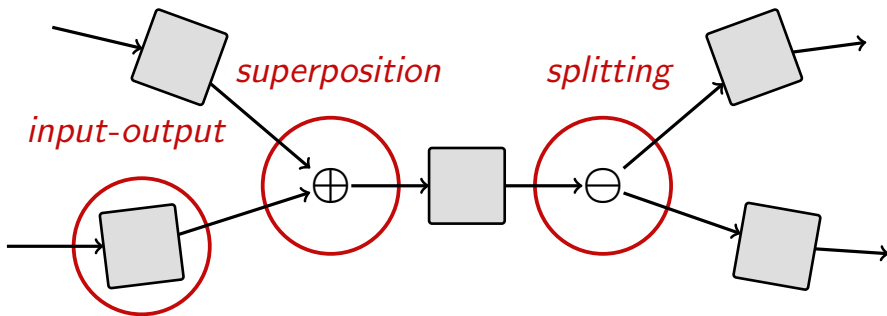
From Renewals to MAPs

	input-output	superposition	splitting
renewals	✓	✗	✗
MAPs			



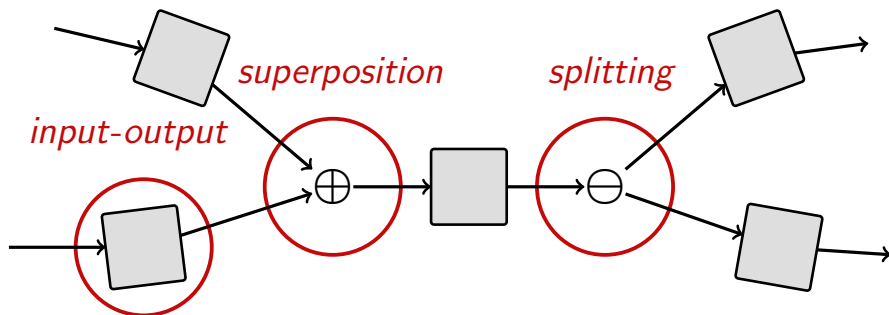
From Renewals to MAPs

	input-output	superposition	splitting
renewals	✓	✗	✗
MAPs		✓	✓



From Renewals to MAPs

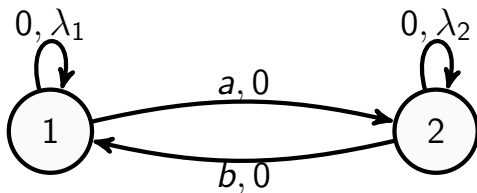
	input-output	superposition	splitting
renewals	✓	✗	✗
MAPs	?	✓	✓



Markov Arrival Processes (MAPs)

example

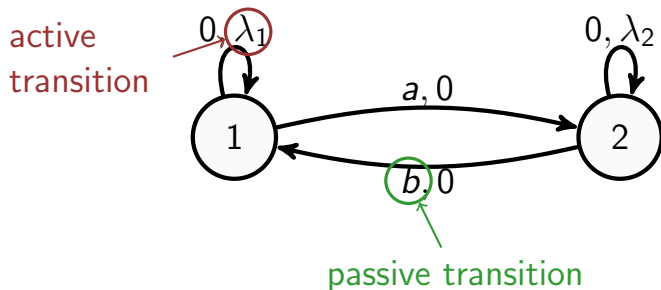
MMPP in MAP representation



Markov Arrival Processes (MAPs)

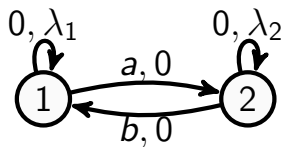
example

MMPP in MAP representation



MMPP-M- Σ : Cache-Miss MAP

arrival MAP:

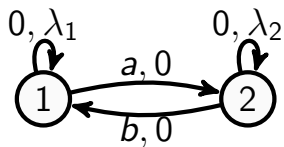


phase-type TTL:



MMPP-M- Σ : Cache-Miss MAP

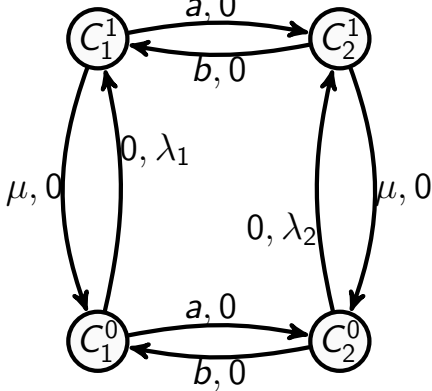
arrival MAP:



phase-type TTL:

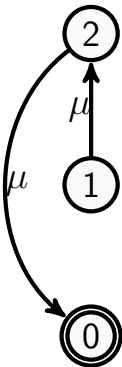


cache-miss MAP:

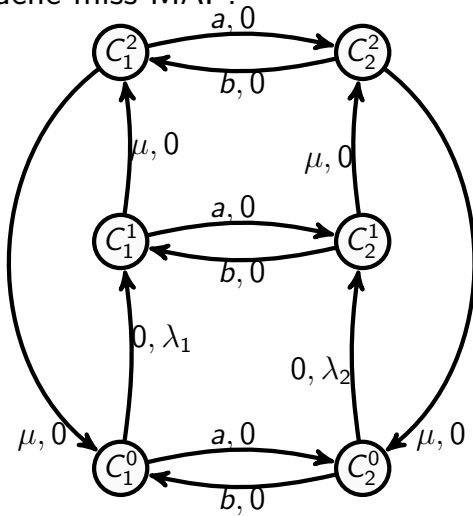


MMPP- E_2 - Σ : Cache-Miss MAP

phase-type TTL:



cache-miss MAP:



General Theorem for *MAP-PH- Σ* Cache

arrival MAP ($\mathbf{D}_0, \mathbf{D}_1$)

phase-type TTL ($\mathbf{S}, \pi$) with generator $\mathbf{P}$

the miss process is a MAP ($\mathbf{D}'_0, \mathbf{D}'_1$) given by

$$\mathbf{D}'_0 = (\mathbf{P} \oplus \mathbf{D}_0) + \mathbf{1}' \otimes D_1$$

$$\mathbf{D}'_1 = \begin{pmatrix} \mathbf{0}_{n \times n} & \pi \mathbf{D}_1 \\ \mathbf{0}_{m \times n} & \mathbf{0}_{n \times m} \end{pmatrix}$$

similar for *MAP-PH- $\mathcal{R}$* and *MAP-PH-min*($\Sigma, \mathcal{R}$)

Accuracy of the Poisson Approximation

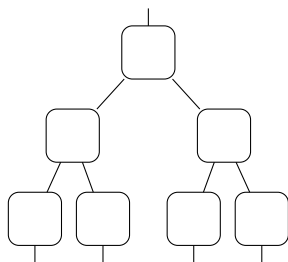
Accuracy of the Poisson Approximation

a cache's miss process $\neq$ Poisson

relative error in a binary tree

$M-M-\Sigma$ at lowest level

1000 objects & Zipf(0.85)



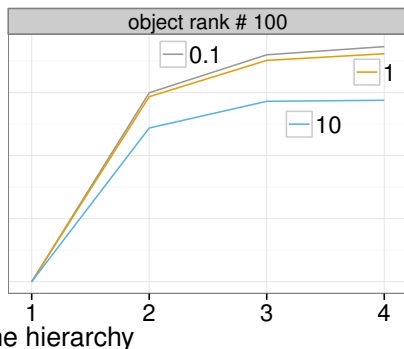
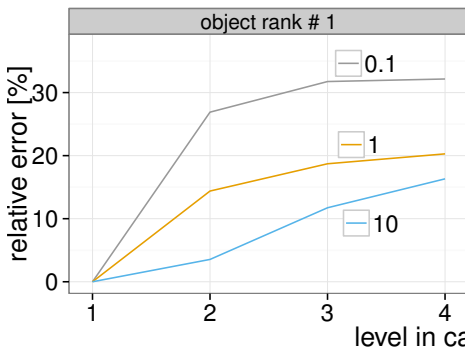
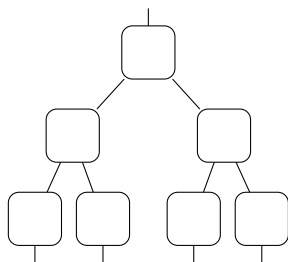
Accuracy of the Poisson Approximation

a cache's miss process $\neq$ Poisson

relative error in a binary tree

$M-M-\Sigma$ at lowest level

1000 objects & Zipf(0.85)



Summary and Conclusion

Conclusions

three practical cache models $\mathcal{R}$, Σ , $\min(\Sigma, \mathcal{R})$

renewal request model

- computationally efficient

- limited to line networks

MAP request model

- broadly applicable

- exponential space complexity $\rightarrow$ numerical methods

future work: performance **bounds** and **tool** support

Exact Analysis of TTL Cache Networks

Daniel S. Berger, Philipp Gland,
Sahil Singla, and Florin Ciucu

IFIP WG 7.3 Performance
October 7, 2014



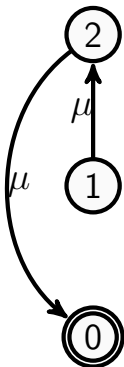
Backup Slides

Model Comparison and Summary

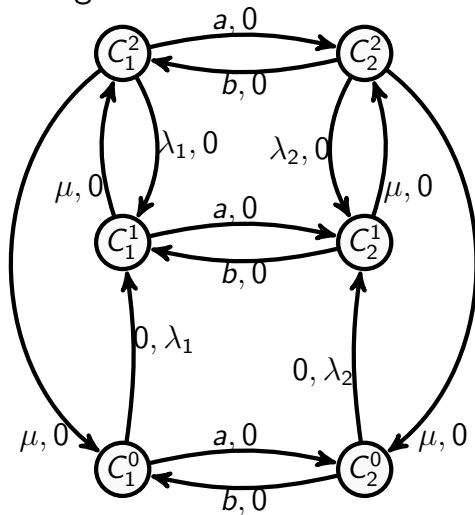
	Renewals	MAPs
request model	IID assumption	dense in point processes
timer assumptions	renewals & deterministic	phase-type ($\approx$ deterministic)
superposition	approximations	exact analysis
splitting	p -thinning	p -thinning & Markovian
type of solution	explicit/ numerical Laplace inversion	explicit (<100 states) numerical ($<10e5$ states)

Caching MAP: $MMPP-E_2-\mathcal{R}$

phase-type TTL:



caching MAP:



1 Introduction and Contribution

2 Analysis

- Line Networks (Renewal Request Model)
- Feedforward Networks (MAP Request Model)

3 Accuracy of the Poisson Approximation

4 Model Comparison and Conclusions