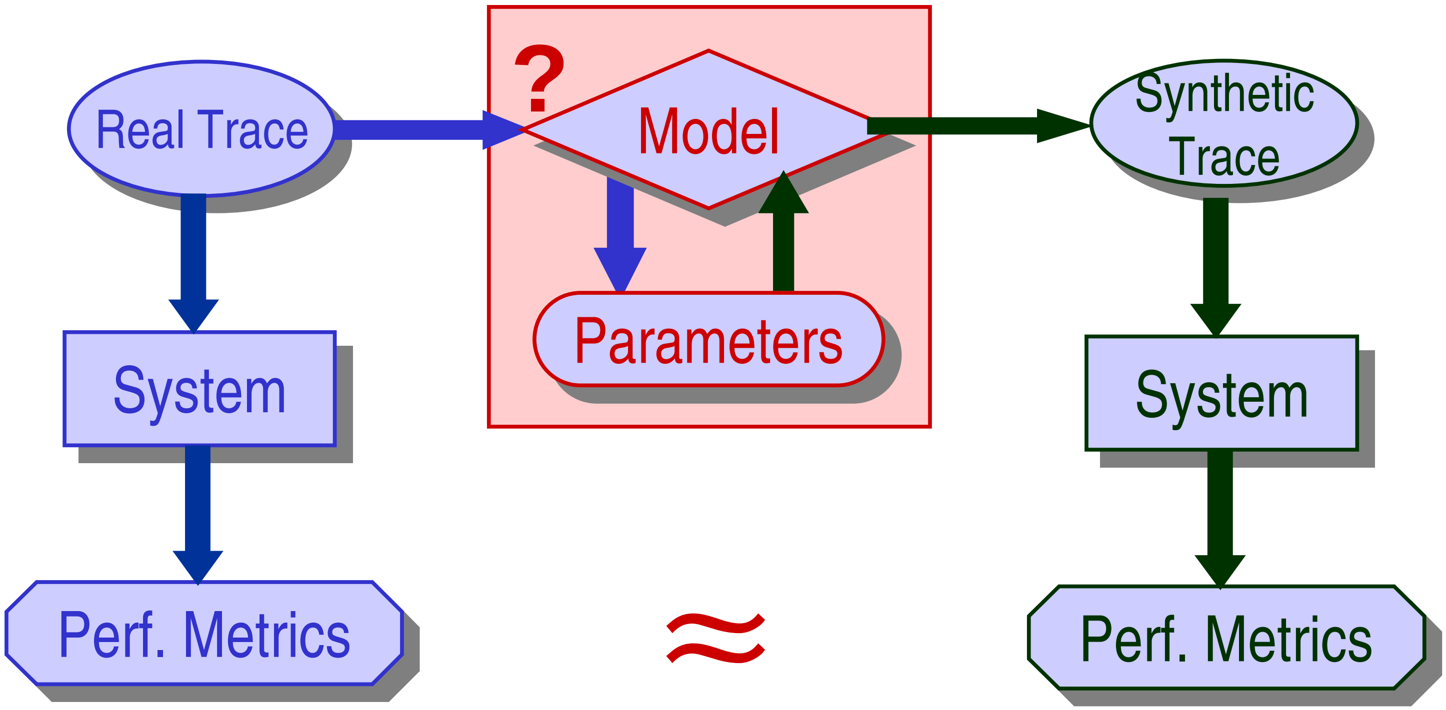
Capturing the Spatio-Temporal Behavior of Real Traffic Data

Mengzhi Wang, Anastassia Ailamaki, and Christos Faloutsos

PARALLEL DATA LABORATORY

Carnegie Mellon University

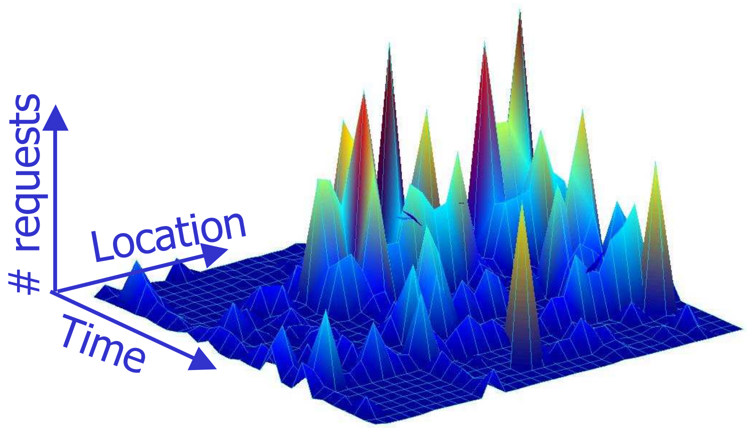
Traffic Modeling

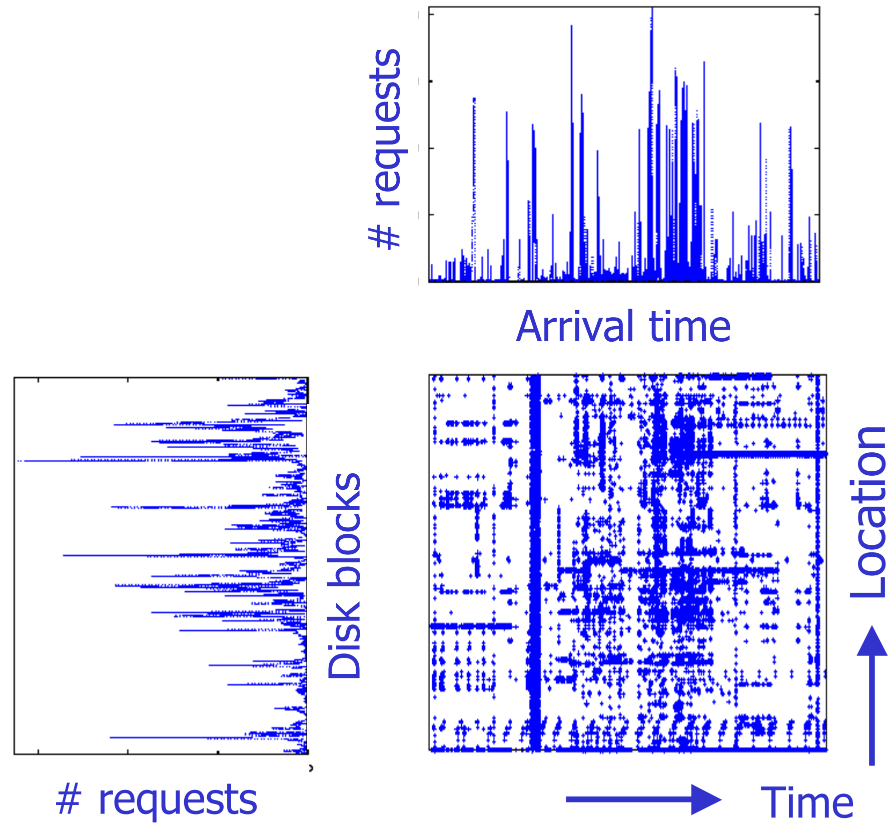


- Type of traffic: Disk/Memory
  - (arrival_time, location, size, R/W, …)

Challenges: Burstiness

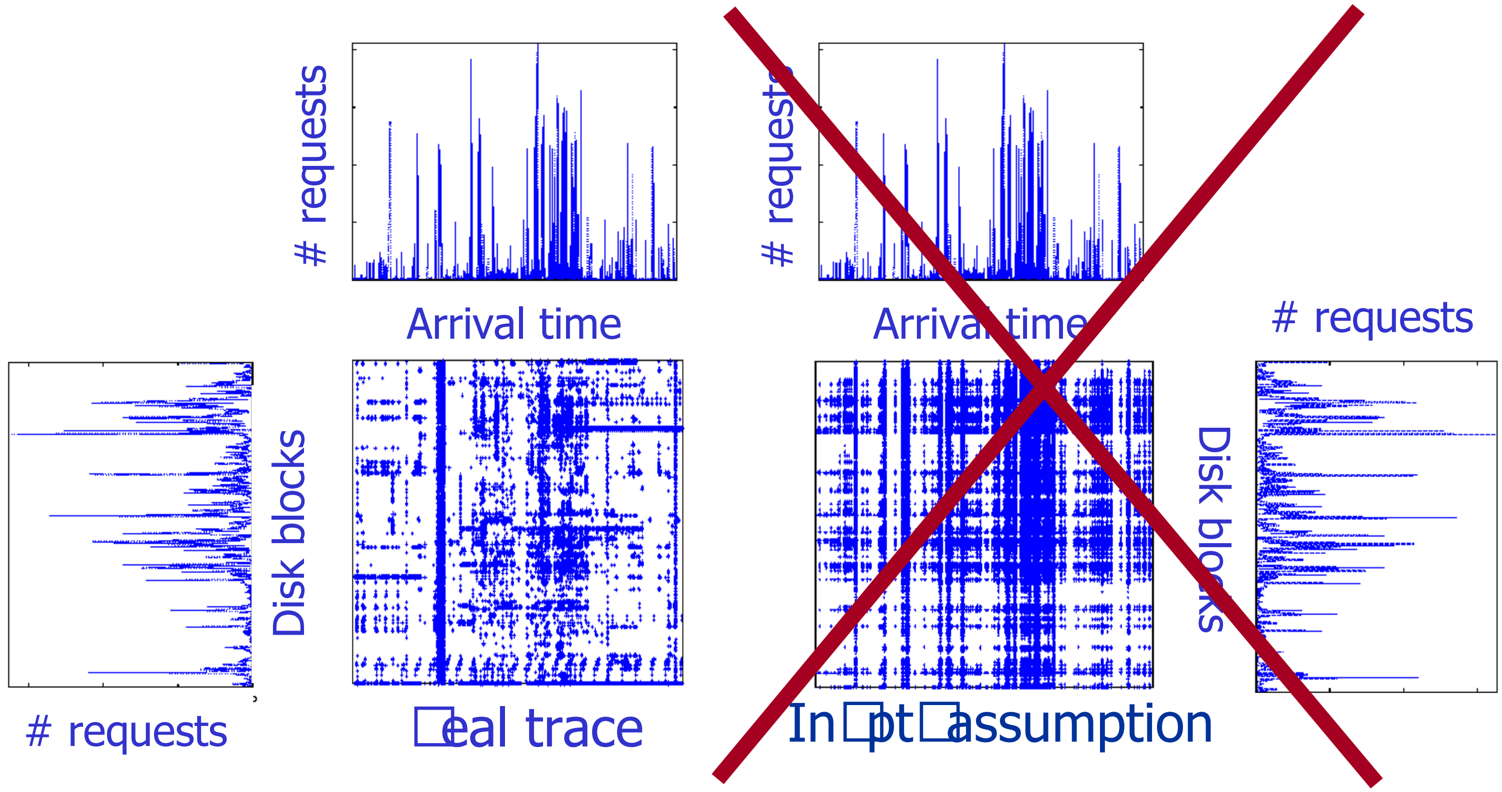
- Real traffic:
  - –Is bursty along time
  - –Is bursty along space





Cello Disk Trace

Challenges: Correlation



- **Real traffic has strong spatio-temporal correlation**

Traffic Modeling: Goal

- A statistical model which
  - Can generate realistic traffic
    - Captures burstiness over time and space
    - Captures spatio-temporal correlation
  - Is practical
    - Accurate
    - Concise
    - Fast to compute
    - Fast in parameter fitting

Outline

- Motivation
- Quantifying burstiness and correlation
- Our approach: PQRS model
- Experiments
- Conclusions

Question

- Question:
  - How to measure the burstiness and correlation?

- Answer: Entropy and Mutual Information!

Quantifying

- Existing measures:
  - **Entropy:** uniformity of a discrete probability function
  - **Mutual Information:** correlation between two random variables

- Applied to our domain:
  - Entropy for **burstiness**
  - Mutual Information for **correlation**

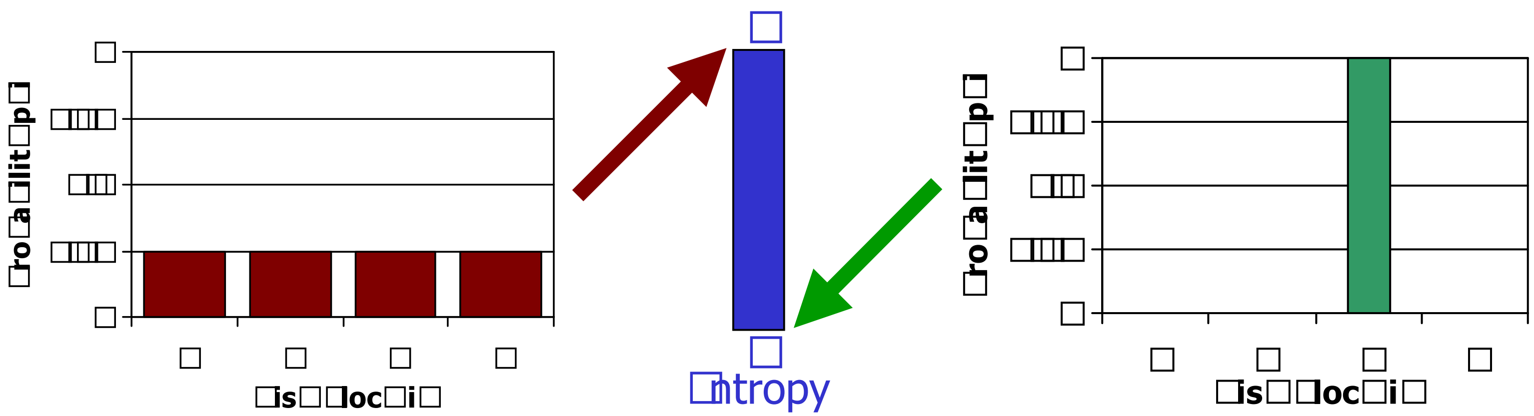
Reminder: Entropy

- **Entropy:**

  $H(E) = - p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n,$

  for a random variable E. (e.g., disk block id)

  - Maximum value for uniform distributions
  - Small value for high-biased distributions



Reminder: Mutual Information

- **Mutual information:**

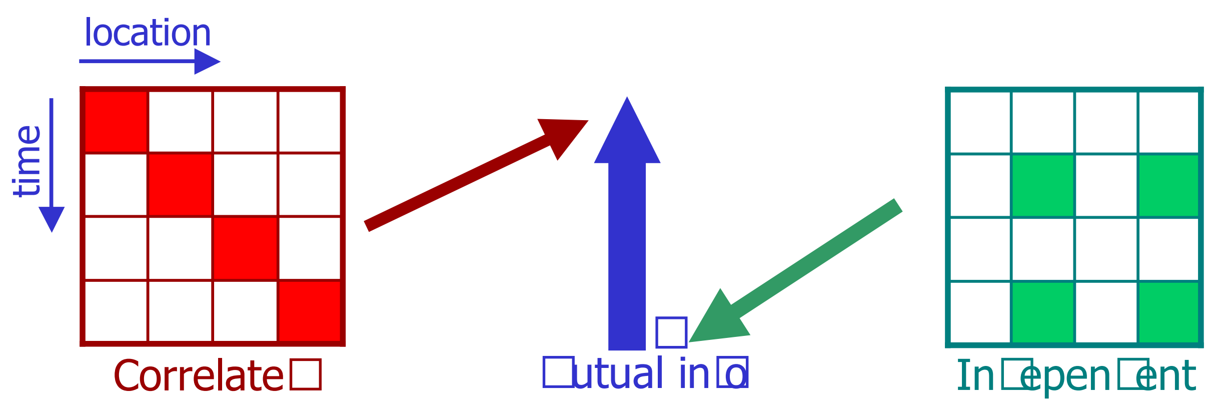
  $I(E;F) = H(E) + H(F) - H(E, F),$

  for two random variables E and F. (e.g., disk block id and arrival time)

  where $H(E, F)$ is the **joint entropy** definded as $-\Sigma_{i,j}\, p_{i,j} \log_2 p_{i,j}.$

⇒ Positive value for correlated variables

⇒ Zero for independent variables



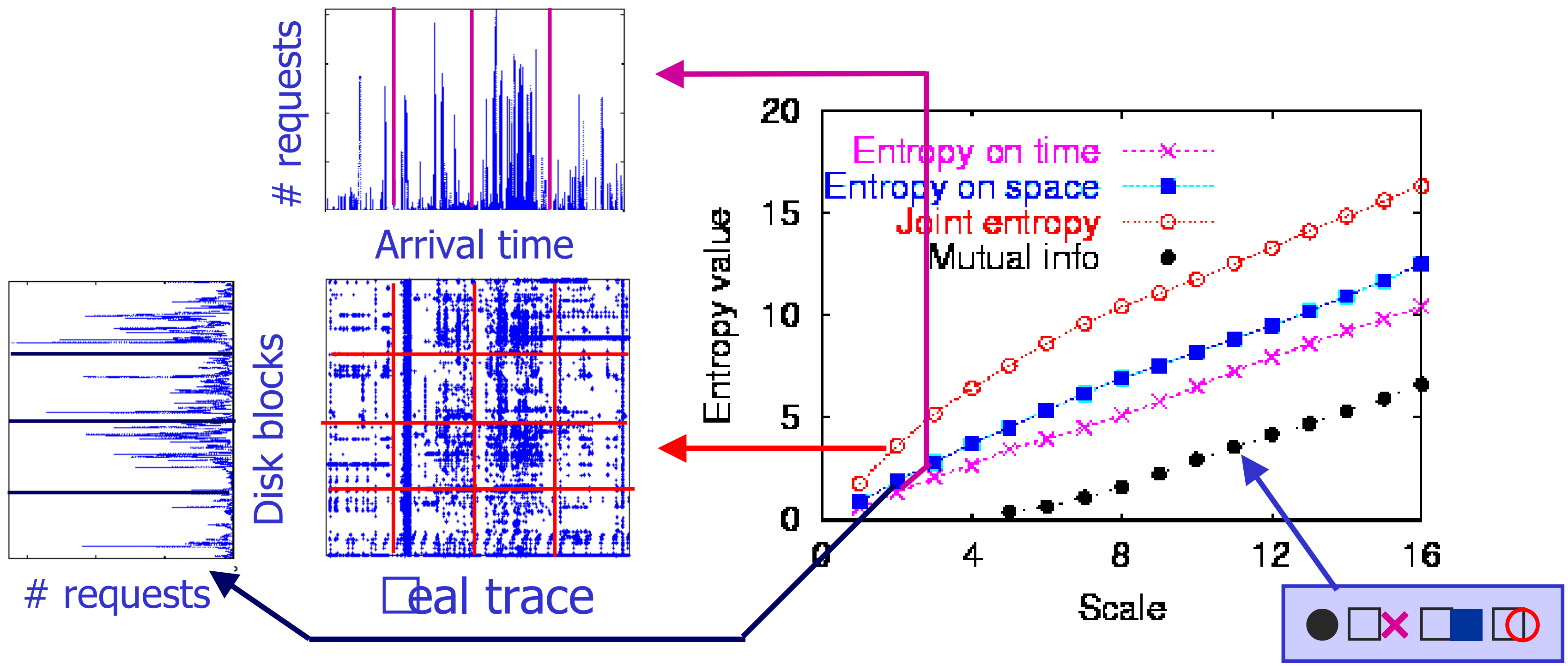
Quantifying

- Question:
  - How to apply entropy on real traffic data?

- Answer: Entropy plots!

Our Approach: Entropy Plot

- **Entropy plot**: Entropy values against scale

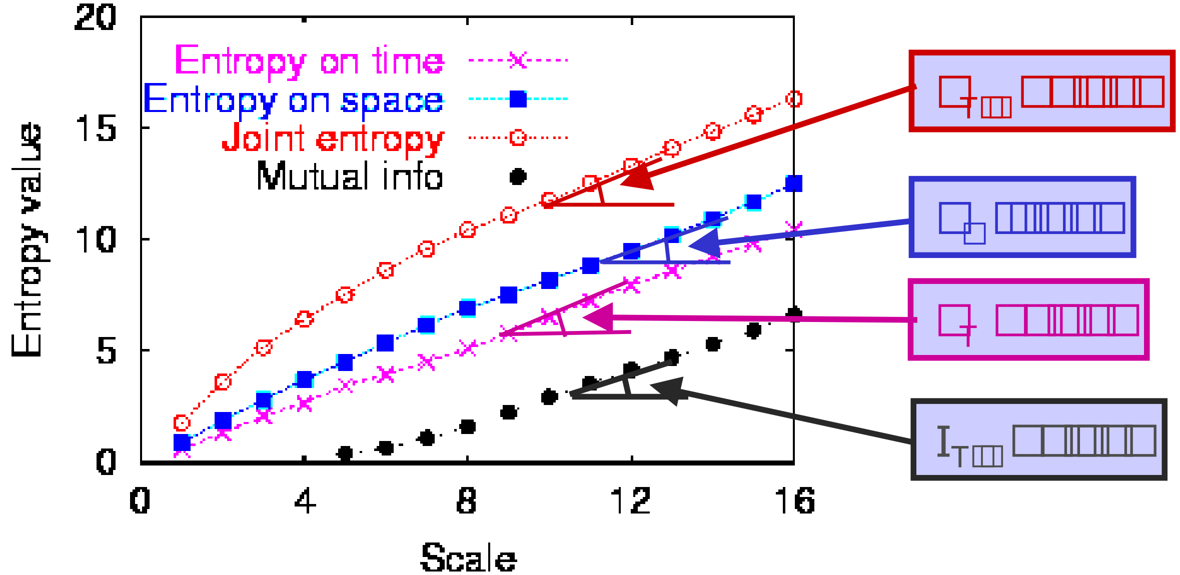


□ □ntrop□plot for real traffic □ata is linear

Carnegie Mellon
Parallel Data Laboratory

Reading Entropy Plots

- □Real traffic □ata sho□s □□rstiness an□strong correlation
  - □Linear temporal entropy plot □□□uni□orm□□□e□tremely bursty□
  - □Linear spatial entropy plot □□□uni□orm□□□e□tremely bursty□
  - □Linear mutual in□ormation entropy plot □□□in□epen□ent□



Real Traffic Data Summary

- Real traffic data shows
  - Burstiness along time and space
  - Strong spatio-temporal correlation
  - Linear entropy plots

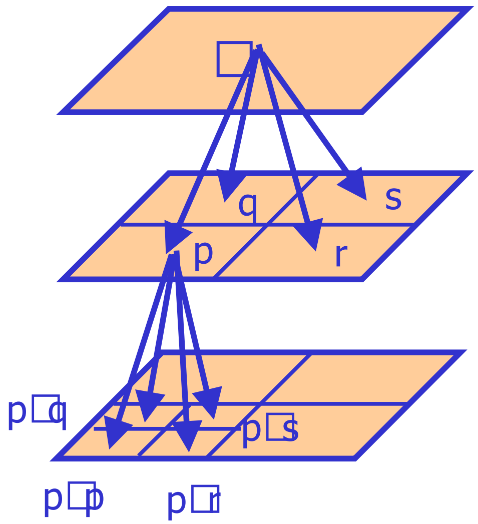
Question

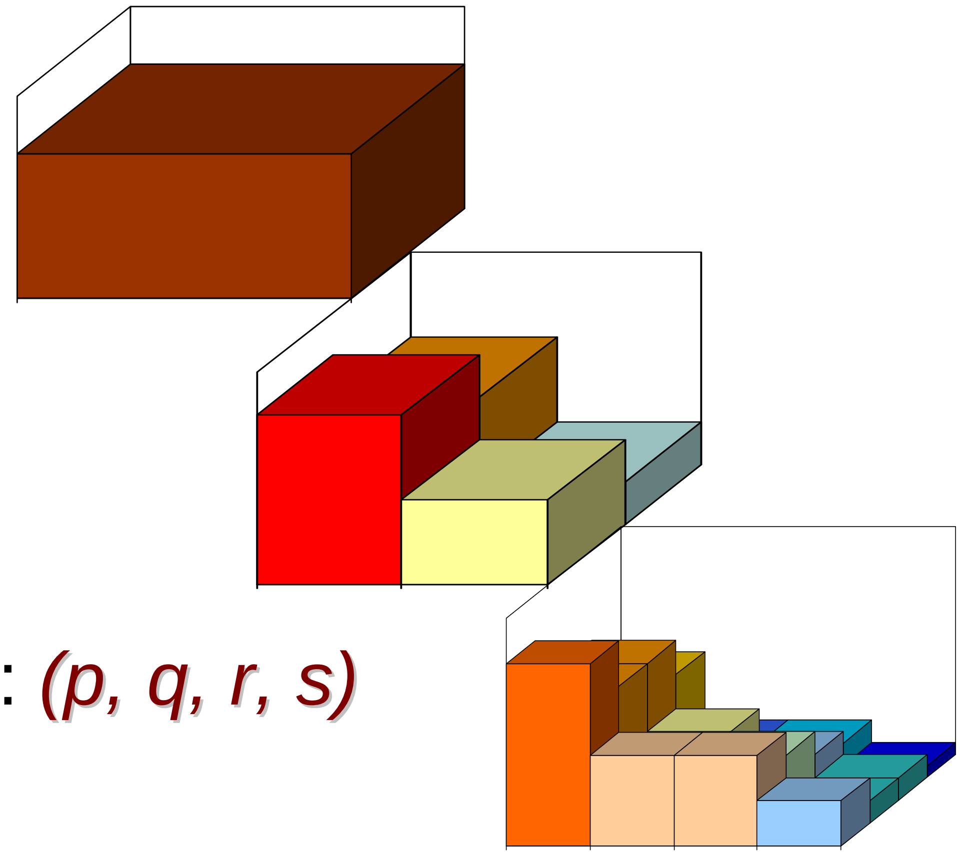
- Question:
  - How to mimic such traffic?
- Answer: PQRS Model!

Outline

- Motivation
- Quantifying burstiness and correlation
- Our approach: PQRS model
- Experiments
- Conclusions

Our Approach: PQRS Model

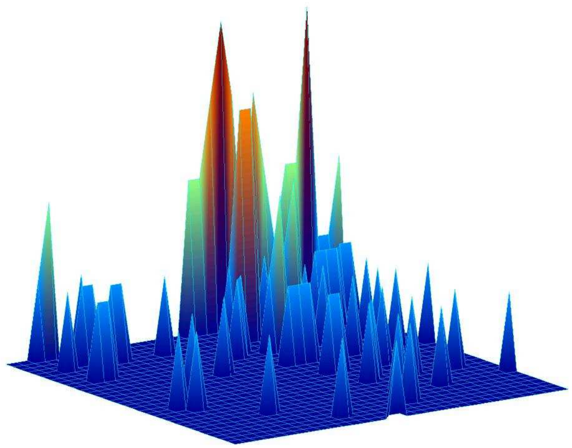




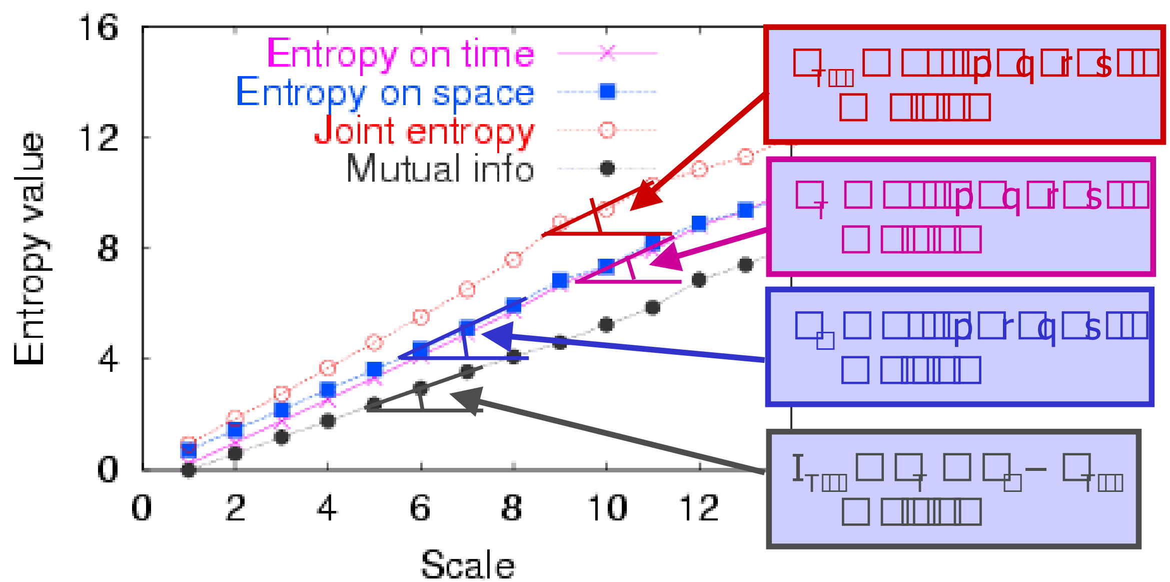
- Four parameters: *(p, q, r, s)*

Entropy Plots for PQRS Model

- Lemma 5: PQRS traces have linear entropy plots.



(p,q,r,s) = **(0.200, 0.027, 0.001, 0.772)**



PQRS Model Properties

- Lemma 7: PQRS model includes Poisson model on arrival time and location when

  $p=q=r=s = 0.25$

- Lemma 8: PQRS model includes independence model for arrival time and location when

  $p/q = r/s$

Efficiency of PQRS Model

- Lemma 9: Generating a PQRS trace takes O(M*N) time: linear on # of requests.

- Lemma 10: Computing entropy plot takes O(M*N) time: linear on # of requests.

  - M: # of requests in the trace
  - N: # of steps in generation / # of points in entropy plot

Traffic Modeling: Goal

- A statistical model which
  - Can generate realistic traffic
    - ü Captures burstiness over time and space
    - ü Captures spatio-temporal correlation
  - Is practical
    - Accurate
    - ü Concise
    - ü Fast to compute
    - ü Fast in parameter fitting

Outline

- Motivation
- Quantifying burstiness and correlation
- Our approach: PQRS model
- Experiments
- Conclusions

Evaluation Methodology

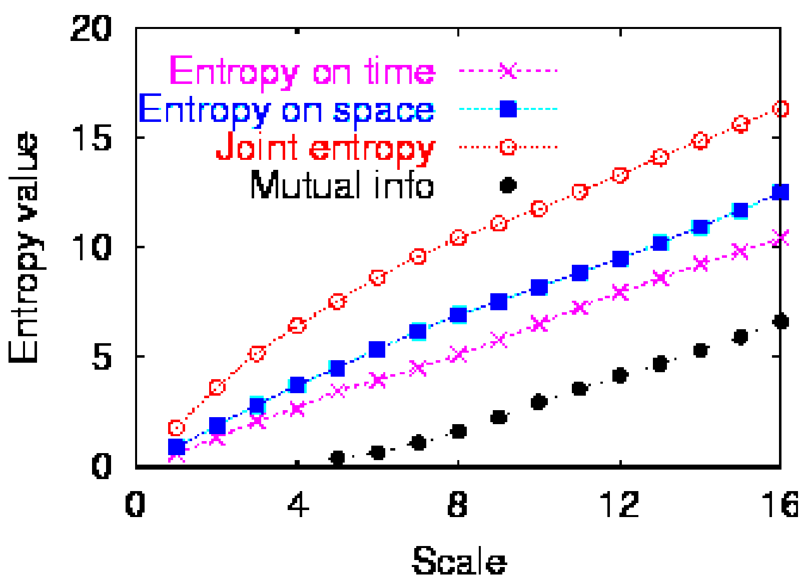
- Datasets:
  - Cello disk traces from HP
    - Unix file servers, one day long, accurate to microseconds
  - TPC-C memory reference traces
    - Online transaction system, one run for each type of transactions

- Tools
  - Response time / queue length distributions for disk traces
    - Generated by DiskSim 2.0
  - Cache miss rates for memory traces
    - Generated by a realistic cache simulator (associativity, LRU)

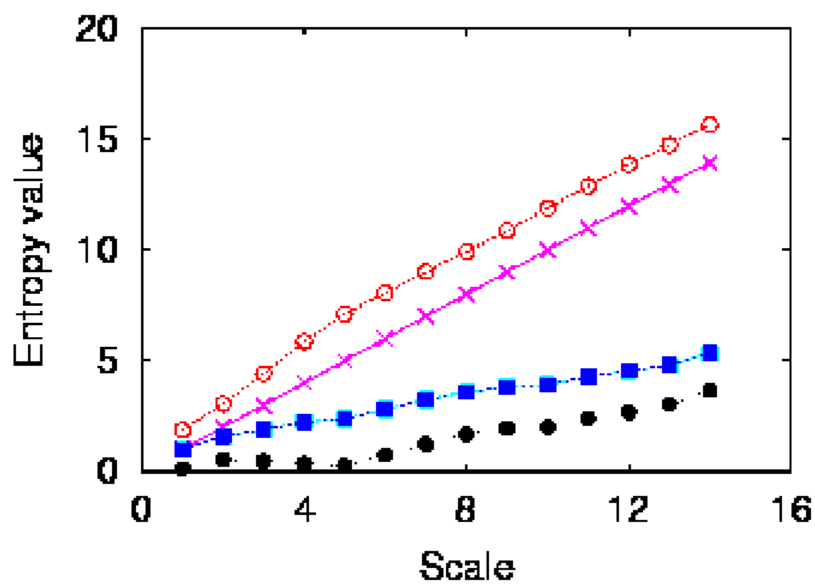
Evaluation Methodology

- Questions:
  - Q1: Are entropy plots of real traffic data linear?
  - Q2: Does correlation matter in terms of performance?
  - Q3: How does the PQRS model perform?

Q1: Linear Entropy Plots!



Entropy plot for a disk trace

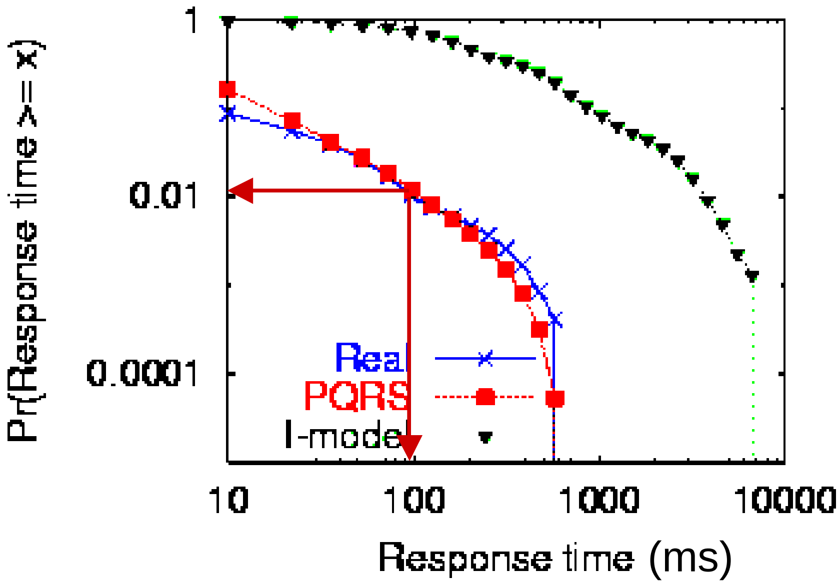


Entropy plot for a memory trace

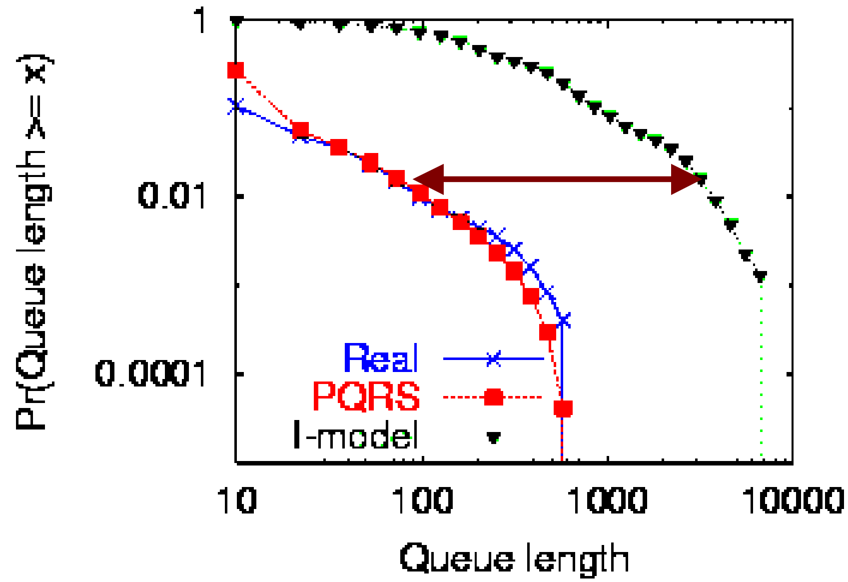
- **PQRS model is applicable!**

Q2: Correlation Matters!

Disk trace performance comparison



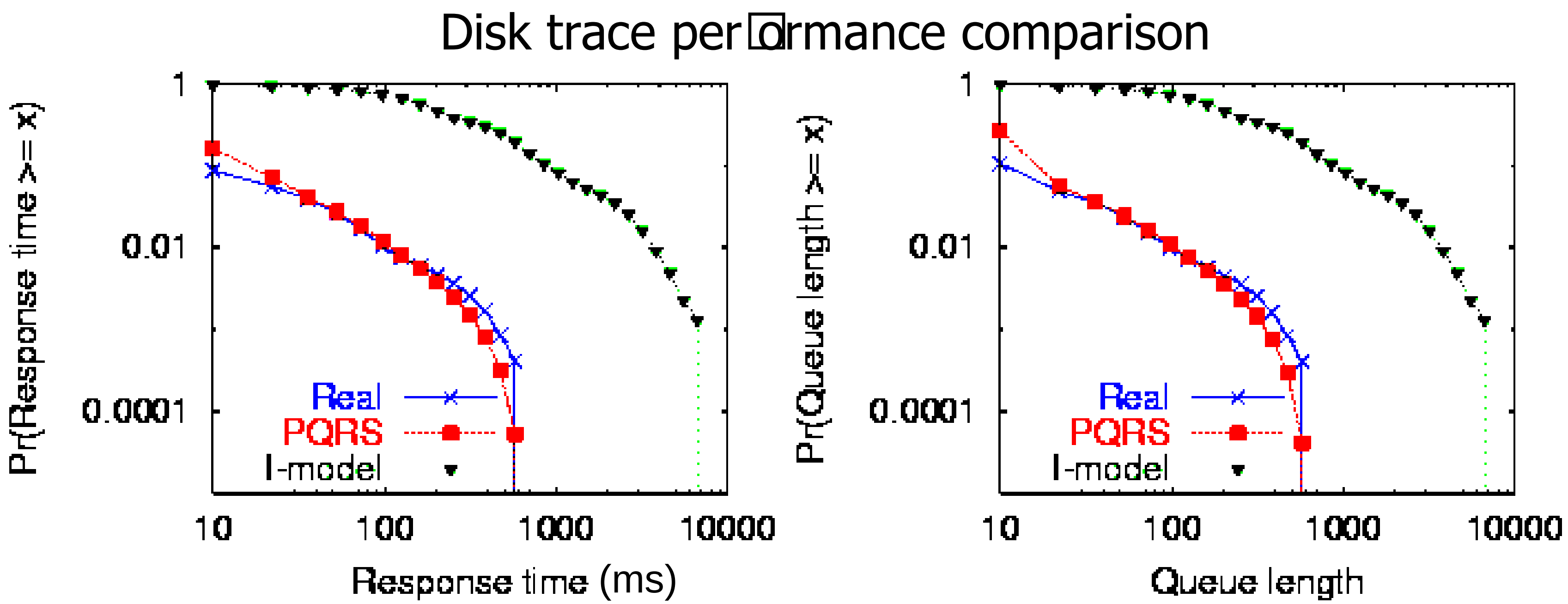
Response time distribution



Queue length distribution

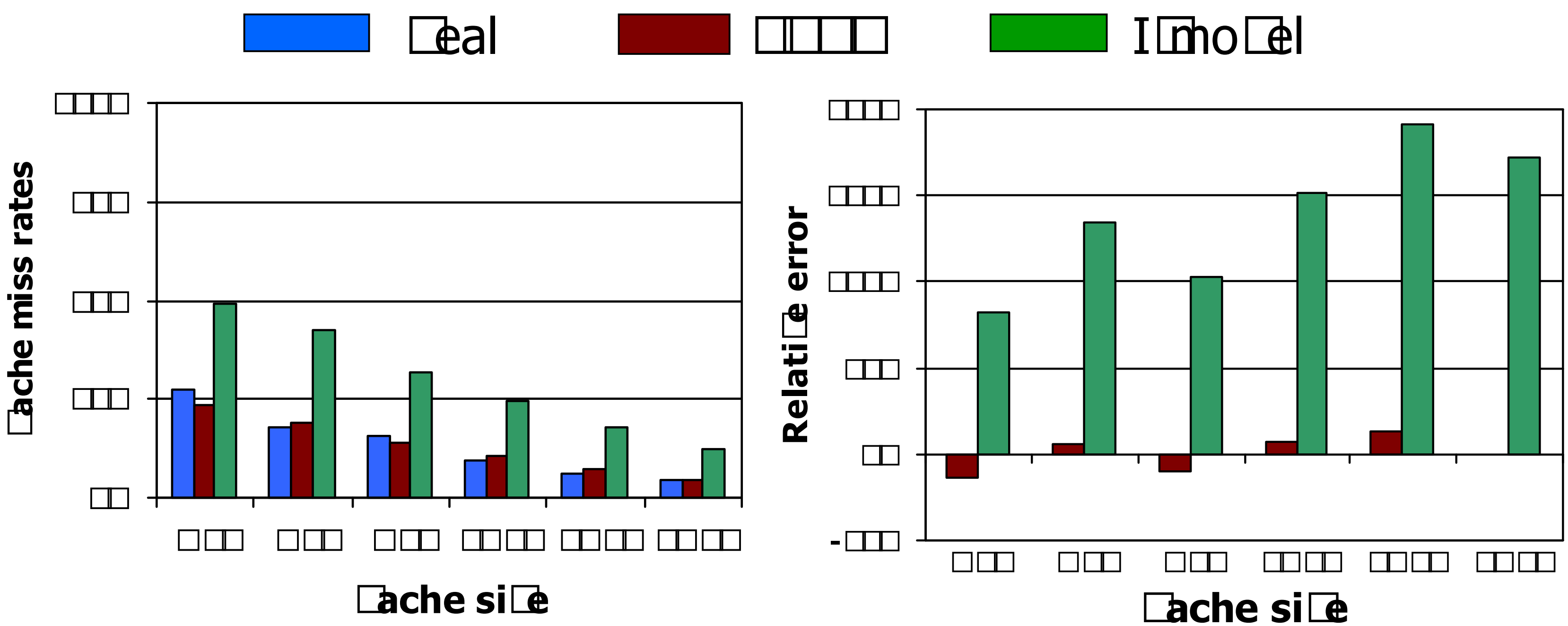
Independence leads to [illegible] worse performance

Q3: PQRS Wins! (Disk)



Response time distribution **Queue length distribution**

Q3: PQRS Wins! (Memory)



Traffic Modeling: Goal

- A statistical model which
  - Can generate realistic traffic
    - ü Captures burstiness over time and space
    - ü Captures spatio-temporal correlation
  - Is practical
    - ü Accurate
    - ü Concise
    - ü Fast to compute
    - ü Fast in parameter fitting

Outline

- Motivation
- Quantifying burstiness and correlation
- Our approach: PQRS model
- Experiments
- Conclusions

Conclusions

- Quantifying burstiness and correlation
  - Entropy plots as measurement
  - Existence of strong spatio-temporal correlation
- PQRS model:
  - Captures the burstiness and correlation
  - Has good scalability
  - Matches the performance properties of real traces
    - Independence assumption leads to 10x worse performance
    - PQRS provides low relative error

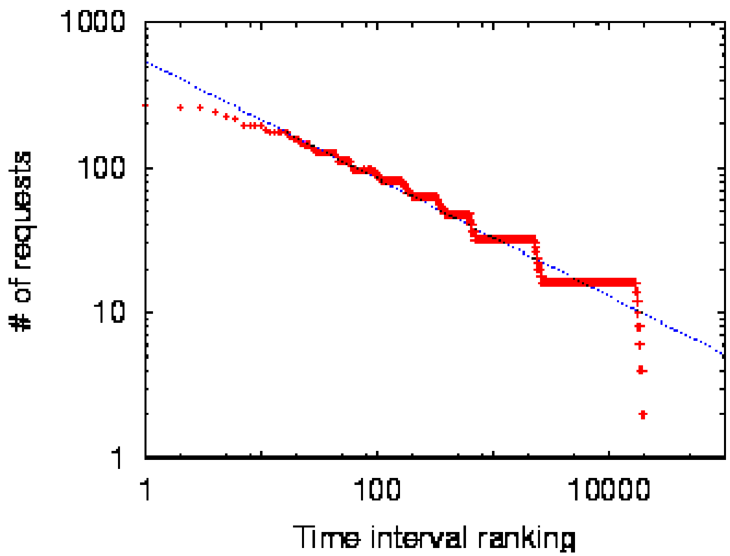
Related Work

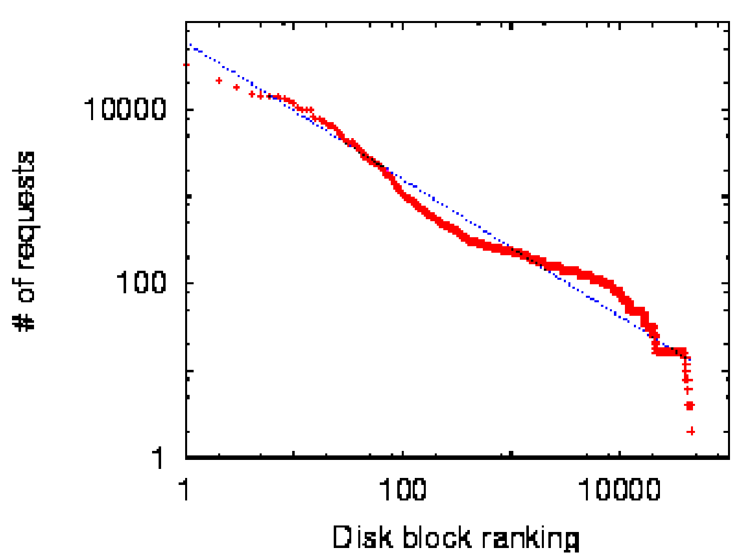
- Temporal modeling
  - fBm[Leland94], fARIMA[Garrett94], ON/OFF[Barford98], MWM[Riedi99], b-model[Wang02], …
  - No location information
- Web cache trace generating
  - [Almeida96], [Breslau99], [Busari01], …
  - Categorical space
- Disk traffic modeling
  - [Ganger95], [Shriver98], [Gomez00], …

Ł **No measure for spatio-temporal correlation**

Zipf or Not?

- Not exactly Zipf-like distributions

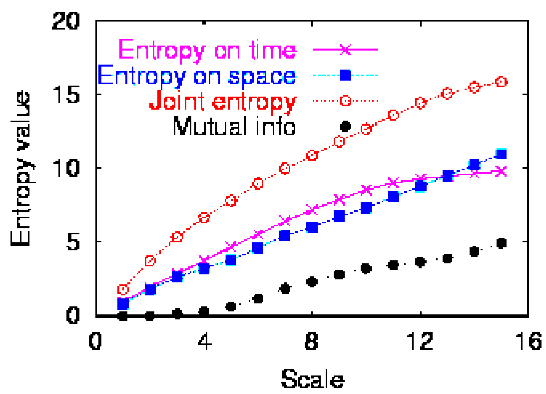




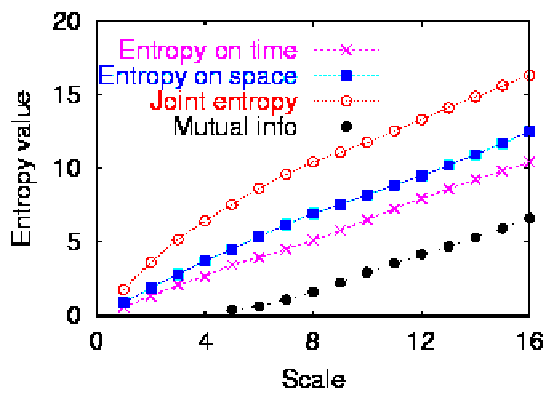
Rank-frequency plots for time ticks and disk blocks

Entropy plots for disk traces

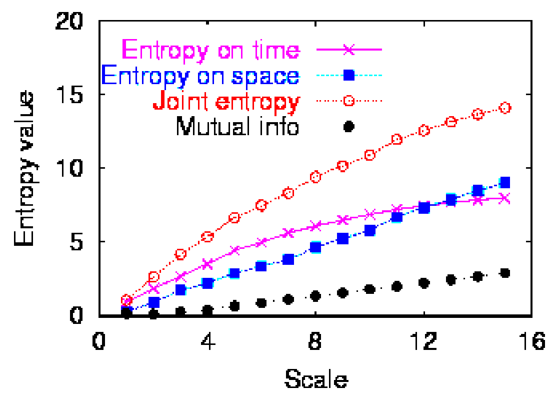
- All the disk traces have linear entropy plots
  - Bursty along time and space
  - Of strong spatio-temporal correlation



Disk



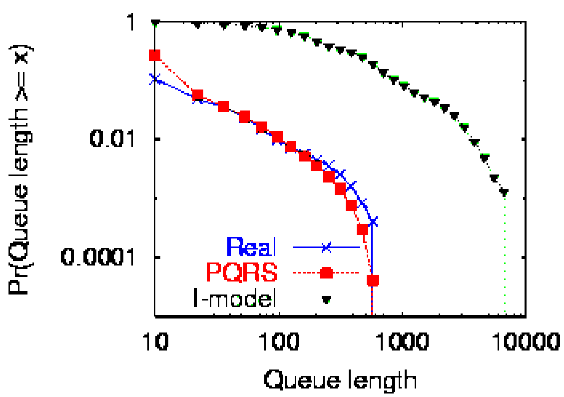
Disk



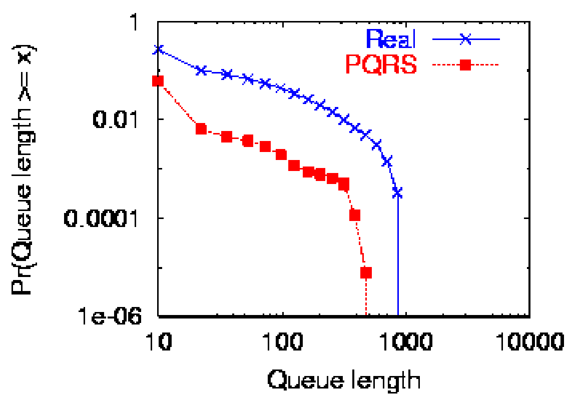
Disk

Disk Trace Performance (1/2)

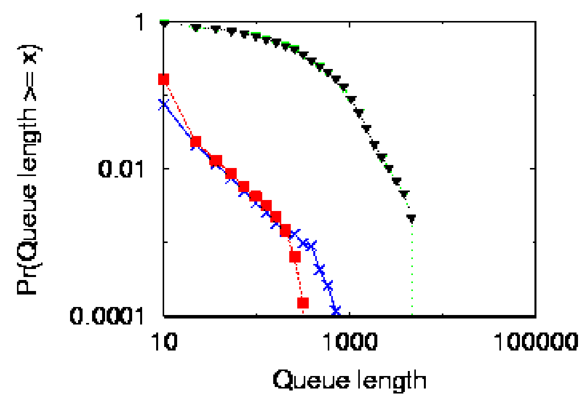
- Correlation matters
  - Strong correlation leads to shorter queues



Disk□□



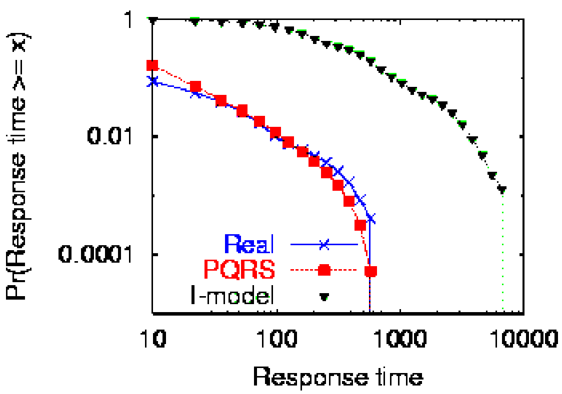
Disk□□



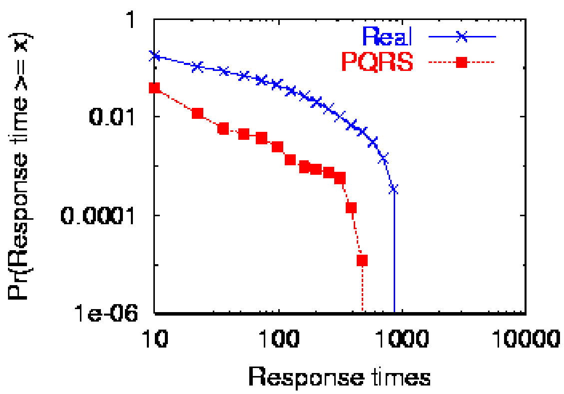
Disk□□

Disk Trace Performance (2/2)

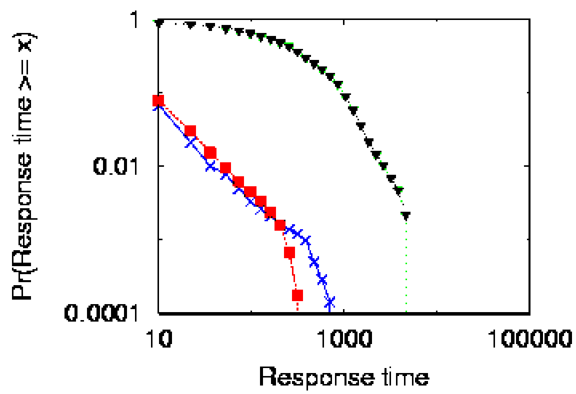
- PQRS model works very well
  - Capturing the burstiness and correlation accurately



Disk



Disk

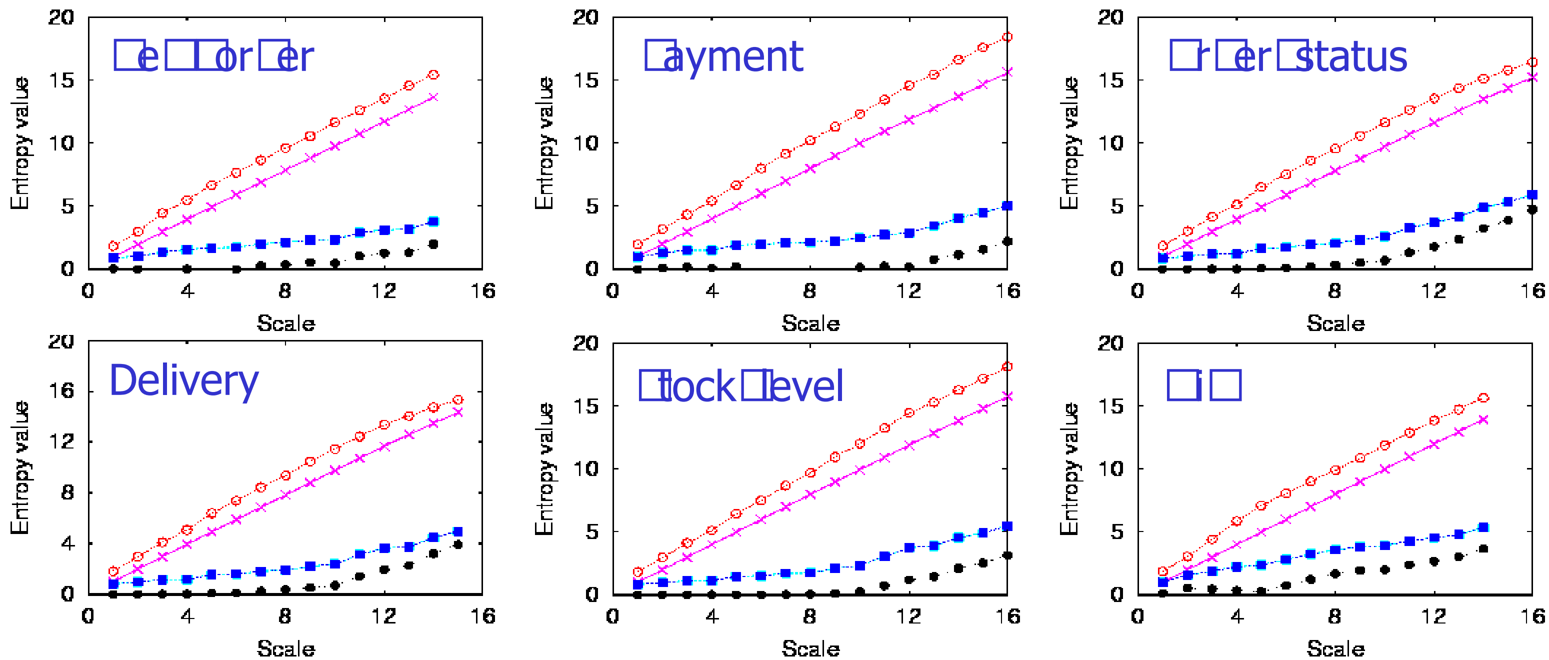


Disk

Carnegie Mellon
Parallel Data Laboratory

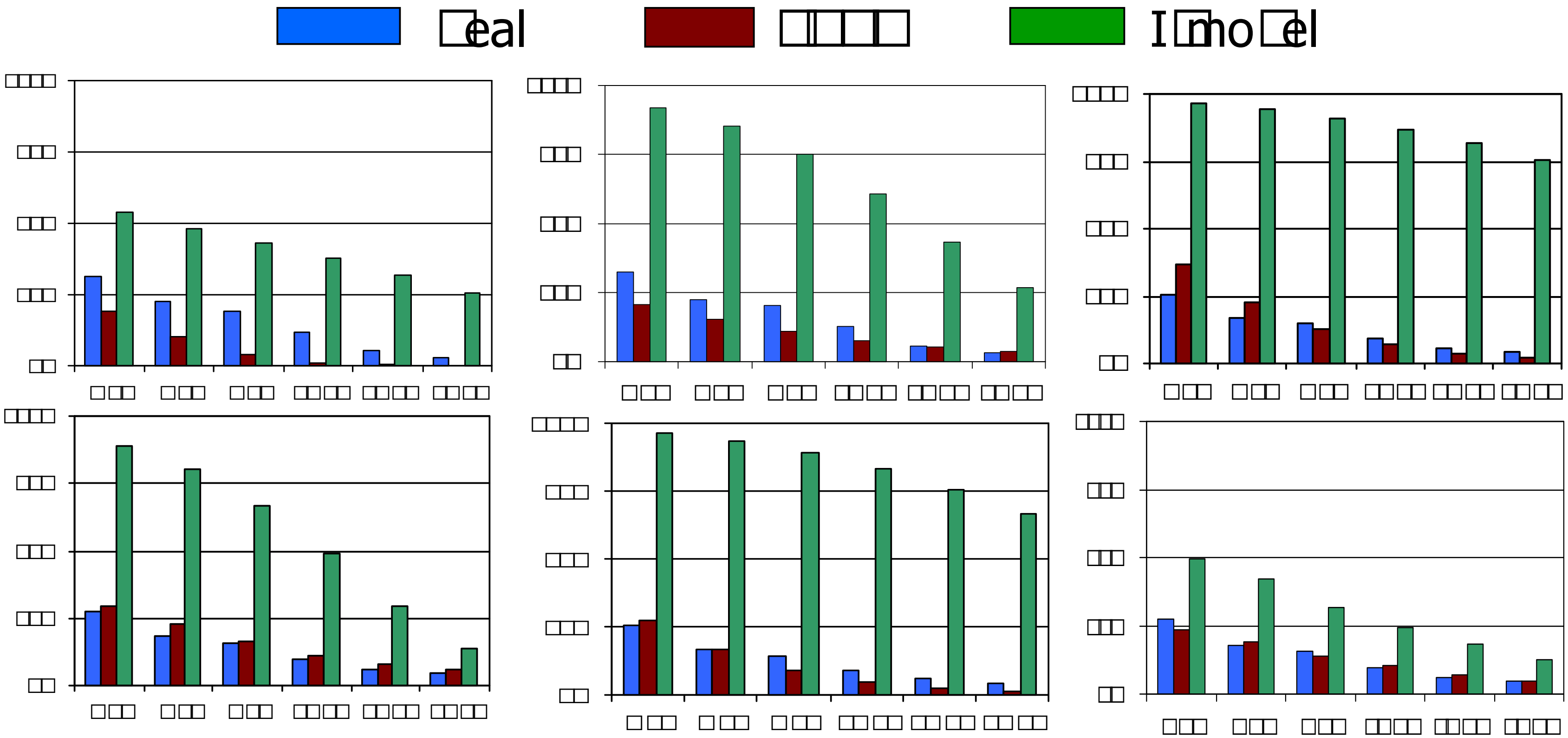
Entropy Plots for Memory Traces

- All memory traces have linear entropy plots
  - Bursty along space but smooth along time
  - Of moderate spatio-temporal correlation



Performance of Memory Traces

- PQRS model works well !



Parameter Fitting (Disk)

- Disk traces are
  - Bursty along time and space
  - Of strong spatio-temporal correlation

Trace	# requests	R_T	R_S	$R_{T,S}$	$I_{T,S}$	(p,q,r,s)
Disk-a	4,575,798	**0.641**	**0.819**	1.058	**0.402**	(0.862, 0.001, 0.257, 0.741)
Disk-r	1,822,781	**0.847**	**0.833**	0.984	**0.696**	(0.016, 0.258, 0.720, 0.006)
Disk-w	3,300,628	**0.641**	**0.728**	0.992	**0.377**	(0.150, 0.013, 0.053, 0.784)
Disk-0	1,101,416	**0.814**	**0.690**	0.941	**0.563**	(0.043, 0.184, 0.722, 0.001)
Disk-2	1,396,649	**0.790**	**0.723**	0.904	**0.609**	(0.200, 0.027, 0.001, 0.772)
Disk-7	371,320	**0.722**	**0.573**	0.881	**0.414**	(0.056, 0.135, 0.808, 0.001)

Parameter Fitting (Memory)

- Memory traces are
  - Smooth along time
  - Highly bursty along space
  - Strong spatio-temporal correlation

Traces	Length	# requests	R_T	R_S	$R_{T,S}$	$I_{T,S}$	(p,q,r,s)
New order	14,990,636	4,000,000	**0.962**	**0.200**	0.996	**0.166**	(0.030, 0.255, 0.001, 0.714)
Payment	17,242,172	4,573,044	**0.963**	**0.281**	1.042	**0.202**	(0.239, 0.047, 0.713, 0.001)
Order status	1,355,168	268,943	**0.950**	**0.456**	0.417	**0.417**	(0.095, 0.185, 0.001, 0.722)
Delivery	525,100	129,388	**0.957**	**0.439**	0.409	**0.409**	(0.090, 0.192, 0.001, 0.717)
Stock level	14,453,440	3,613,360	**0.974**	**0.349**	0.271	**0.271**	(0.231, 0.064, 0.704, 0.001)
Mix	12,268,876	4,000,000	**0.983**	**0.309**	0.302	**0.302**	(0.248, 0.054, 0.697, 0.001)